Phase II Trials Powered to Detect Tumor Subtypes

John D. Roberts\textsuperscript{1,2} and Viswanathan Ramakrishnan\textsuperscript{3}

Abstract

Classical phase II trial designs, including "adaptive" designs, require the prospective characterization of tumors. We propose a 2-stage phase II design that allows for characterizations of tumors and selection of a subtype of interest at the conclusion of stage 1. The stage 2 objective is a classical estimate of the response rate for either the tumor or a subtype, or a formal test of the hypothesis that the response rate for a subtype is greater than the overall response rate. Considering likely scenarios, stage 1 sample sizes approximately range from 20 to 100 with a usual size of 50. This compares with typical classical stage 1 sample sizes of 12 to 30. Total sample sizes range from sizes identical to classical designs (tens to scores) to large sizes typical of phase III trials in metastatic disease (hundreds). Our design is more efficient than previous adaptive designs because it allows for the selection of a tumor subtype of interest on the basis of results from stage 1. It complements classical phase II and phase III designs in which investigators compare different treatments in similar patients and tumors by positioning a treatment as fixed (control) and using tumor subtype as the variable of interest. Clin Cancer Res; 17(17); 5538–45. ©2011 AACR.

Introduction

Classical oncology phase II trial designs provide the minimum sample sizes necessary to estimate the activity of a treatment against a tumor type with a certain precision \((1, 2)\). These designs typically involve 2 stages, the first of which is an early stopping rule for futility. Whereas these designs consider a specified tumor type as homogeneous with regard to the probability of response, for most tumors there is likely to be an incompletely understood, predictable heterogeneity in the probability of response. Investigators have proposed "adaptive" modifications of classical phase II trial designs to address this issue \((3–7)\). To date, however, all of these designs require prospective tumor characterization, which is cumbersome and limits adaptation. We propose a phase II design that does not require prospective tumor characterization in stage 1, and therefore is less cumbersome in stage 1 and more adaptable in stage 2.

Materials and Methods

Definitions

Consider "tumor type" to refer to a set of tumors that conforms to the tumor-related eligibility criteria of a phase II trial. Consider tumor "subtype" to refer to a subset of those tumors that is identifiable by further characteriza-

\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
\textbf{Definition} & \textbf{Description} \\
\hline
Tumor type & A set of tumors that conforms to the tumor-related eligibility criteria of a phase II trial. \\
Subtype & A subset of tumors that is identifiable by further characterization. \\
Response rate & The probability that a tumor responds to a treatment. \\
Prevalence & The proportion of tumors in a population that have a specific characteristic. \\
Conditional probability & The probability that a tumor responds to a treatment given a specific characteristic. \\
\hline
\end{tabular}
\end{table}

Stage 1 sample size calculation

In classical phase II trial designs, the parameter of interest is the response rate, or probability of response, \(p_R\) (Table 1). To formulate the early stopping rule, one must specify a minimum overall response rate of interest. Once one presumes the existence of a subtype, to calculate the stage 1 sample size, one must specify something about the subtype in terms of prevalence and response rate. One approach is to specify a minimum subtype prevalence of interest as defined by the presence of a biomarker, \(p_{M/B}\), and a minimum subtype response rate of interest, \(p_{R|M/B}\), which is the conditional probability of response within the subset of biomarker-positive tumors. Another approach is to specify minimum subtype prevalence within responding tumors of interest, \(p_{M|R/B}\), which is the conditional probability of the subtype within the subset of responding tumors, and a minimum response rate of interest, \(p_R\). We consider the former approach to be more intuitive. In either case, the product of prevalence and conditional probability yields the probability that a tumor is both a responder and biomarker positive, \(p_{M/R/B}\), a joint probability:

\[ p_{M/R} = p_{M/R/B} = p_{M/B} p_{R|M/B} = p_{R/B} p_{R|M/B} \] (1)
Of course, a third approach is to specify directly the minimum prevalence of tumors that are both responders and positive for the biomarker of interest.

To calculate the stage 1 sample size, let \( N \) denote the number of subjects with tumors evaluable for response. Each subject has a tumor that is either responding or nonresponding. Let \( R \) denote the number of responding tumors. Each tumor is either biomarker positive or biomarker negative; however, we intend to characterize only responding tumors. Let \( J \) be the number of biomarker-positive, responding tumors. Now, the entire set of subjects can be allocated among 3 mutually exclusive categories: subjects with biomarker-positive, responding tumors; subjects with biomarker-negative, responding tumors; and subjects with nonresponding tumors. The number of subjects in each of these 3 categories is \( J \), \( R - J \), and \( N - R \), respectively, and corresponds to a multinomial distribution with cell probabilities \( p_{J}, p_{R} - p_{J}, \) and \((1 - p_{J})\). The proposed design is a straightforward extension of Gehan’s (1) design using this multinomial model.

As in classical designs, stage 1 is an early stopping rule for futility. Define \( n_{J} \) as the minimum number of biomarker-positive, responding tumors required to proceed to stage 2, and \( n_{R} \) as the minimum number of responses required to proceed to stage 2. For a specified false-negative rate, \( \beta \), the stage 1 sample size will be determined by requiring the joint probability of the event \( J < n_{J} \) and the event \( R < n_{R} \) to be smaller than \( \beta \). Under the multinomial model, this translates to numerically solving the equation

\[
P(J < n_{J}, R < n_{R}|N) = \sum_{J=0}^{n_{J}} \sum_{R=0}^{n_{R}} \frac{N!}{(J-r)!((N-r)!)} \left( p_{J} \right)^{J} \left( p_{R} - p_{J} \right)^{r} (1 - p_{J})^{N-J} < \beta
\]

for \( N, n_{R}, \) and \( n_{J} \). For a fixed \( n_{R} \), several combinations of \( n_{R} \) and \( N \) may satisfy the above inequality. The recommended combination is the case that leads to the smallest \( N \) such that \( n_{R} \) is the minimum needed to observe at least \( n_{J} \) where \( n_{J} \geq 1 \).

**List of candidate biomarkers**

A list of candidate biomarkers that are predictive of response is created from any available knowledge. In general, a candidate biomarker can be any feature that is differentially expressed within a tumor type. Such a biomarker may or may not have known prognostic or predictive significance. For example, for an agent that is targeted to a specific mutation, the presence of that mutation would be an obvious candidate biomarker.

**Analysis of stage 1**

Count the number of responding tumors \((R)\). Test the responding tumors for as many candidate biomarkers as is feasible or desired. Count the number of responding tumors \((J)\) for each candidate biomarker. Compare the numbers of responding \((R)\) and biomarker-positive, responding tumors \((J)\) with the minimum numbers needed to proceed to stage 2 \((n_{R}, n_{J})\) with 4 possible outcomes (Fig. 1): (1) the overall response rate is not potentially of interest, and the prevalence of biomarker-positive, responding tumors for any candidate biomarker is not potentially of interest \((R < n_{R} \) and \( J < n_{J})\); (2) the overall response rate is not promising, but there is a promising biomarker \((R < n_{R} \) and \( J > n_{J})\); (3) the overall response rate is promising, but there is no promising biomarker \((R \geq n_{R} \) and \( J < n_{J})\); and (4) the overall response rate is promising, and there is a promising biomarker \((R \geq n_{R} \) and \( J \geq n_{J})\). In scenarios 1 and 3, one has shown that none of the candidate biomarkers is of interest. In scenarios 2–4, one has shown that the treatment is potentially of interest, whether it is matched to a promising subtype (scenario 2) or in general (scenario 3), or both (scenario 4). For the purposes of simplicity, we consider only the case in which a single candidate marker is identified as promising.

**Design of stage 2**

If the overall response rate is not promising and there is no promising biomarker, the study has been completed. It may be appropriate to archive responding tumors for future testing should new candidate biomarkers become known. Indeed, the observation of a number of responding tumors that expressed none of the candidate biomarkers might prompt a search for biological similarities among them. However, classical statistical considerations would not allow one to attach any significance to a promising biomarker identified in such a manner.

If the overall response rate is promising but there is no promising biomarker, the situation is analogous to a promising stage 1 outcome of a classical phase II trial, and a
stage 2 sample size can be calculated according to one of these designs. Similarly, if the overall response rate is not promising but there is a promising biomarker, a stage 2 sample size can be calculated according to a classical design but powered to estimate the prevalence of tumors that are both biomarker-positive and responding. An alternative would be to proceed to a new trial limited to tumors prospectively characterized as promising biomarker positive. This is analogous to the approach of previously proposed adaptive designs.

If the overall response rate is promising and there is a promising biomarker, it is recommended that stage 2 be formulated as a test of the hypothesis that the response rate among biomarker-positive tumors ($p_{R|M}$) is significantly greater than the overall response rate ($p_R$). This approach requires an estimate of the prevalence of the promising biomarker in the overall tumor population, $p_M$.

One can estimate the probability of biomarker positivity in the overall tumor population in 1 of 3 ways: (1) characterize the nonresponding tumors (this resource-intensive approach yields the most precise, unbiased estimate possible, and therefore the greatest power); (2) characterize a sample of the nonresponding tumors (this yields a less precise, unbiased estimate and less power); or (3) characterize an unrelated sample of similar tumors (this might allow one to use existing data or to increase the precision of the estimate, but any such estimate would be subject to whatever bias was introduced by the different methods used to select the different tumor samples).

To design a hypothesis-testing stage 2, the investigator also must provide a minimum difference of interest between the response rate among biomarker-positive tumors and the overall response rate. The number of responses needed to detect this difference is calculated by using the marginal

---

**Figure 1. Stage 1 flow diagram.**

Specify:
A minimum overall response rate of interest
AND
A list of candidate predictive biomarkers (or subtypes)
AND
A minimum subtype response rate of interest and a minimum subtype prevalence of interest, OR
A minimum prevalence of responding tumors with the subtype of interest

- Treat patients
- Count $R$, the number of responding tumors
- Identify a promising biomarker by subtyping responding tumors
- Count $J$, the number of responding tumors of the subtype of interest

If $R \geq n_R$?

- No
  - Is $J \geq n_J$?
  - Yes
    - Classical stage two: estimate $p_{M,R}$
    - Stop
  - No
    - Stop

- Yes
  - Classical stage two: estimate $p_R$
  - Classical stage two: estimate $p_{M,R}$

- $J \geq n_J$?

- No
  - Is $J \geq n_J$?
  - Yes
    - Obtain estimate of $p_M$
    - Hypothesis-testing stage two: $p_{M,R} > p_R$
  - No
    - Stop

- Yes
  - Hypothesis-testing stage two: $p_{M,R} > p_R$

---

**Statistics in CCR**

© 2011 American Association for Cancer Research
binomial distribution of $J$, the number of biomarker-positive, responding tumors. Then, for a specified significance level $\alpha$ and power $(1 - \beta)$, using the marginal binomial $[(N, p_{M,R})]$ distribution of $J$, exact one-sample methods (8) can be used to determine the required sample size. Sample sizes used in Figs. 2 and 3 were calculated using PROC POWER in SAS with the TEST = EXACT option for testing the null and alternate hypotheses stated in terms of the minimum biomarker prevalence at 5% significance level and 80% power.

**The issue of multiple testing**

Because we consider only cases in which a single biomarker is designated as promising, no statistical adjustments for multiple testing are required in stage 2. As in classical designs, stage 1 is an early stopping rule for futility. The efficiency of this stopping rule is a function of the scope of the candidate markers. For example, if virtually all tumors of the tumor type mark positive for at least 1 candidate biomarker, and there are any responses in stage 1, it is a virtual certainty that the trial will proceed to stage 2.

**Results**

Stage 1 sample sizes are largely an inverse function of the minimum biomarker-positive, responding tumor prevalence of interest and have little dependence on the minimum overall response rate of interest (Fig. 2). We think it would be unusual for a stage 1 sample size to be $< 20$, which is, for example, the number necessary to exclude a biomarker with a response rate of 50% and prevalence of $< 34\%$. For a common tumor, a stage 1 sample size of $\sim 50$ might be common. Such a stage 1 sample size would routinely detect a biomarker with a prevalence of 10% and a response rate of 50%.

Stage 2 incremental sample sizes ($N^+$) vary considerably according to the goal of stage 2. In a situation in which the overall response rate is promising but there is no promising biomarker ($R \geq n_R$ and $J < n_J$), the stage 2 incremental sample size would be smaller, and the total ($N + N^+$) sample size would be identical to those employed in classical phase II designs, that is, tens to scores of patients. In a situation in which the overall response rate is not promising but there is a promising biomarker ($R < n_R$ and $J \geq n_J$), the stage 2 incremental sample size and the total sample size would be larger than for classical phase II designs. We anticipate that such trials might involve 100 patients or more. Such a trial would expose additional patients to a treatment that is not likely to be effective for most (see “Discussion” below). In a situation in which the overall response rate is promising and there is a promising biomarker ($R \geq n_R$ and $J \geq n_J$), the stage 2 incremental sample size would be much larger than that of classical phase II designs, and the total sample size would approach that of classical phase III trials for metastatic disease, that is, hundreds of patients (Fig. 3).
Discussion

Comparison with previous designs

Investigators have approached the problem of response heterogeneity using adaptive or flexible (e.g., Bayesian) phase II designs (4–7). All of these designs incorporate inefficiencies that render them unpopular. All require stratified patient enrollment based on prospective characterization of tumor biomarkers. This is cumbersome for a variety of reasons:

- Prospective testing frequently causes treatment delays of days or weeks that are frustrating for both patients and clinicians, and may discourage enrollment.
- Although in most cases batch processing of tumor samples would be more efficient and more accurate, tumors are usually tested one at a time to avoid longer treatment delays.
- Although experience with a new candidate biomarker often is limited to a research laboratory, if treatment decisions (for example, determining whether a patient is eligible) are based on test results, that test must be performed within a Clinical Laboratory Improvement Amendments–certified laboratory.
- In the case of multiple candidate biomarkers, it is cumbersome to enroll sufficient numbers of patients into each stratum.

Our design, which does not require prospective tumor testing, addresses all of these problems. In addition, in the case of negative studies, our design uses fewer resources because initial testing to identify a promising biomarker is performed only on responding tumors, and testing of nonresponding tumors is only necessary if a promising biomarker is tentatively identified.

Our design fills an unmet need in the spectrum of oncology clinical trials. Currently, tumor subtypes are matched with treatments largely on the basis of preclinical findings and retrospective analysis of phase III trials. Experience to date indicates that preclinical findings are poorly predictive of tumor response type (see “Historical examples” below). Retrospective analyses of phase III trials also are a poor platform for accomplishing this objective. Phase III trial sample sizes may be unnecessarily large or too small to answer clinically relevant questions concerning tumor subtypes. Phase III trials typically do not include prospective tumor banking. They may bank tumor samples from only a subset of enrolled patients, which raises the question about appropriate sample size, or include optional tumor banking, which renders the sample size unknown and may introduce unknown biases. Phase III trials generally are limited to treatments that have shown efficacy in phase II. A subtype-specific treatment that is never appropriately matched to its subtype in phase II is

Figure 3. Hypothesis-testing stage 2, incremental sample sizes (\(N^+\), at 5% significance level and 80% power) for the stage 1 outcome of \(n_J = 1\) and \(R > n_J\) at a false-negative rate of 5%. The total sample size for a study is the sum of the stage 1 sample size (\(N\)) and the stage 2 sample size (\(N^+\)). The lines interrupted by numbers represent the stage 2 sample size (\(N^+\)). Example: When \(p_{M,R} \sim 0.045\) and \((p_{M,R} - p_R) \sim 0.04\), the incremental stage 2 sample size is 246. If the stage 1 sample size is 60, the total sample size needed is 306.
unlikely to be studied in phase III and therefore unlikely to be appropriately matched in any phase. Finally, once the sample sizes are estimated by means of the proposed method, it is important to consider the corresponding operating characteristics. We are currently exploring the operating characteristics through simulations and we intend to publish our findings in a future work.

**Historical examples**

The examples that follow illustrate how our design may be both more effective and more efficient than current approaches.

**Example 1.** CALGB 500104 was a classical phase II trial of the farnesyltransferase inhibitor tipifarnib in cutaneous melanoma (NCT00060125 at ClinicalTrials.gov, updated March 17, 2010). No responses were observed in 14 patients, and thus the trial showed with sufficient confidence that tipifarnib is not active in the tumor type cutaneous melanoma. At the time the trial was initiated, however, it was known that ~15% of cutaneous melanomas harbor activating mutations in N-ras, a farnesylated signaling protein (9). It would have been quite rational, then, to propose that the N-ras mutation defines a responding subtype. If mutated N-ras is associated with a response rate of ~50%, the false-negative rate for this classically designed trial (i.e., the probability that tipifarnib is active in mutated N-ras cutaneous melanoma) is ~33%. Although interest in tipifarnib in melanoma was stimulated by knowledge of N-ras mutations, this study was too small to effectively rule out tipifarnib activity in a relevant subtype of melanoma.

**Example 2.** In 2004, Lynch and colleagues (10) described a subtype of non–small cell lung cancer, which they defined as a family of activating mutations of the tyrosine kinase domain of the epidermal growth factor receptor (EGFR) that exhibit unusual responsiveness to the EGFR inhibitor gefitinib. Two unusual circumstances led to this discovery. First, although the authors observed an overall response rate of only 8%, which generally would signify an inactive drug, the size of the patient population was unusually large (n = 275). This allowed them to identify 25 patients with responses. Second, they were fortunate to have archived tumor specimens from 9 of these patients, and EGFR gene mutation analysis revealed tyrosine kinase domain mutations in 8 of these 9 tumors. An analysis of samples of non–gefitinib-treated non–small cell tumors and non–small cell tumor cell lines suggested that the frequency of these mutations in the overall tumor population was ~5%. From this, they concluded that the activating tyrosine kinase domain EGFR mutation is a likely biomarker for gefitinib responsiveness. The 2 unusual circumstances that facilitated this discovery (treatment of an unusually large number of patients and availability of archived tumor samples for retrospective analysis) are explicit features of our proposed design. However, the approach taken by the investigators was inefficient. Assuming the observed parameters (i.e., an overall response rate of 9%, a subtype prevalence of 8%, a subtype response rate of 89%, and accepting a 5% false-positive rate), discovery of the EGFR mutation would have required a stage 1 minimum sample size of 36 patients. The expected number of responses would have been 2, and 1 of the responding patients would have had the EGFR mutation. This would correspond to our scenario 2, in which the overall response rate is not promising but there is a promising biomarker (in our notation, R < nR and J ≥ n1). A nonhypothesis-testing stage 2 would estimate the prevalence of biomarker-positive, responding tumors (pM,R). Although in theory one might object that pursuing stage 2 in the face of a less than promising overall response rate as medically or ethically suspect, this is an example of a case in which investigators were willing to do that. To complete stage 2, an additional 62 patients (N+), for a total of 98 patients (N+ N+), would be required. This is about one third the number of patients treated in the absence of our trial design.

Because our design relies on testing a single tumor specimen, it incorporates the simplifying assumptions that tumor testing is error free and individual patients’ tumors are biologically homogeneous. Biological tumor heterogeneity both within a tumor mass and between/among a primary tumor and metastases is a well-described, potentially complex phenomenon and a problem for all study designs, including ours. We anticipate that an appropriate adjustment for less-than-perfect test sensitivity would be to increase the sample size. If the sensitivity is known, this adjustment can be calculated; if it is unknown, the adjustment might be arbitrary. Less-than-perfect test specificity might result in false-positive trial results. This might lead to further, futile investigations, but eventually the facts would become known.

**Example 3.** A limitation of our design is that it allows only a single candidate biomarker to proceed to stage 2. A third example highlights the need for extensions of our trial design to accommodate multiple biomarkers. Current clinical practice guidelines for the treatment of metastatic colon cancer include the combination of the EGFR-targeted monoclonal antibody cetuximab with the cytotoxic agent irinotecan (NCCN Clinical Oncology Practice Guideline v2.2011). This combination is recommended regardless of the presence or expression of overexpression of the putative cetuximab target, EGFR, but only in the absence of activating mutations of the K-ras oncogene. Neither of these recommendations reflects a result from a prospective stratified trial. The presumption that cetuximab would be active only in EGFR-expressing tumors was so strong that patients with EGFR-nondeexpressing tumors were excluded from the study that led to approval (11). Retrospective analysis of this study showed no relationship between the degree of EGFR expression and clinical benefit. The original recommendation not to test for EGFR or to consider the results of EGFR testing was a nonevidence-based extrapolation of this analysis to patients with tumors that do not express EGFR (NCCN Clinical Oncology Practice Guideline v2.2007). Because K-ras is downstream of EGFR, a logical inference would be that cetuximab is inactive in tumors harboring activating K-ras mutations. This inference has been confirmed and incorporated into guidelines. This confirmation, however, is based exclusively on retrospective
analysis, as prospective typing for K-ras mutation status was not incorporated into clinical trials of cetuximab.

Thus, optimal tumor subtyping may require multiple biomarkers, which raises problems involving multiple testing and potential increases in sample size. On the other hand, there may be interactions among biomarkers (e.g., biomarkers may be mutually exclusive or nested) that, if recognized, can reduce the real number of tests. Future work that incorporates multiple, possibly interacting biomarkers in the trial design is desirable.

Although our presentation and examples focus on tumor response heterogeneity, genetic or environmentally determined differences among patients that affect treatments (e.g., in terms of pharmacokinetics or pharmacodynamics) may also contribute to response rates. Our trial design can be easily adapted to discover patient biomarkers that predict response to a particular therapy.

Although classical designs originally included a binary tumor response endpoint, they have since been adapted to consider other endpoints, including various time-to-event endpoints such as time to progression (12–15). An advantage of tumor response is that, because most tumors do not spontaneously shrink, it uniquely reflects a tumor–treatment interaction; a disadvantage is that transient tumor shrinkage may occur without any corresponding symptomatic or a time-to-event benefit. An advantage of time-to-event endpoints is that they measure something that is almost always relevant to patients; a disadvantage is that they reflect both prognostic biomarker effects and tumor–treatment interactions. In a single-arm study using a time-to-event endpoint exclusively, it is impossible to distinguish prognostic effects from tumor–treatment interactions for which a biomarker is predictive. We anticipate that extension of our design to time-to-event endpoints would be feasible through randomization of patients in 2 or more study arms. Additional study arms could involve observation, a placebo treatment, a standard treatment, and/or a rational solution to the intractability of the problem of matching tumors and treatments, our design complements classical phase II and phase III designs.

When the numbers of agents and tumor types were small, empiric exploration of the tumor–treatment matrix was tractable and largely accepted. Increasing understanding of cancer biology and advances in technology have increased both the number of recognizable tumor types and the number of available treatments. These increases have rendered empiric exploration of the tumor–treatment matrix intractable. As a targeted agent moves from the laboratory to the clinic, the hypothesis implicit in the label “targeted” is that the agent has been appropriately matched to its target. This has encouraged many to consider testing of a targeted agent in nontarget-marked tumors as unnecessary and empiricist. We take a different view: Testing of a targeted agent in both target and nontarget-marked tumors, that is, hypothesis testing, is both the essence of the scientific method and necessary given the imperfection of current preclinical models and limitations of retrospective analysis. Our design can be considered an efficient approach to such hypothesis testing and/or a rational solution to the intractability of purely empiric exploration of the tumor–treatment matrix.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The authors acknowledge the encouragement of Dr. Daniel Sullivan.

Grant Support

National Institutes of Health (P30 CA016659).

Received October 18, 2010; revised May 25, 2011; accepted June 13, 2011; published OnlineFirst July 7, 2011.

References

Phase II Trials Powered to Detect Tumor Subtypes

Phase II Trials Powered to Detect Tumor Subtypes

John D. Roberts and Viswanathan Ramakrishnan

Clin Cancer Res 2011;17:5538-5545. Published OnlineFirst July 7, 2011.