Maximizing Return on Socioeconomic Investment in Phase II Proof-of-Concept Trials

Cong Chen¹ and Robert A. Beckman²³

Abstract

Phase II proof-of-concept (POC) trials play a key role in oncology drug development, determining which therapeutic hypotheses will undergo definitive phase III testing according to predefined Go-No Go (GNG) criteria. The number of possible POC hypotheses likely far exceeds available public or private resources. We propose a design strategy for maximizing return on socioeconomic investment in phase II trials that obtains the greatest knowledge with the minimum patient exposure. We compare efficiency using the benefit–cost ratio, defined to be the risk-adjusted number of truly active drugs correctly identified for phase III development divided by the risk-adjusted total sample size in phase II and III development, for different POC trial sizes, powering schemes, and associated GNG criteria. It is most cost-effective to conduct small POC trials and set the corresponding GNG bars high, so that more POC trials can be conducted under socioeconomic constraints. If δ is the minimum treatment effect size of clinical interest in phase II, the study design with the highest benefit–cost ratio has approximately 5% type I error rate and approximately 20% type II error rate (80% power) for detecting an effect size of approximately 1.5δ. A Go decision to phase III is made when the observed effect size is close to δ. With the phenomenal expansion of our knowledge in molecular biology leading to an unprecedented number of new oncology drug targets, conducting more small POC trials and setting high GNG bars maximize the return on socioeconomic investment in phase II POC trials. Clin Cancer Res; 20(7); 1730–4. ©2014 AACR.

Introduction

A critical step in drug development is the "proof-of-concept" (POC) trial, a phase II study designed to provide an initial test of a particular therapy or combination of therapies in a defined population or indication before definitive testing in randomized phase III trials (1). Given that there are typically nearly 1,000 approved and experimental therapies active in oncology clinical trials at any one time, that they can be combined in twos and threes, that different schedules may be used, and that many clinical indications and lines of therapy are available, the number of potential POC trials that could potentially be performed is enormous. The possibilities only increase when one looks at possible subsets defined by biomarker classifiers, in which one must take into account approximately 30,000 genes in the human genome and the genetic instability and consequent heterogeneity of cancer. Although preclinical information offers prioritization of these possibilities, there still remain a very large number of potentially useful POC trials, which in our experience far exceeds the availability of funding from either public or private sources. In addition, the number of available patients for oncology clinical trials is insufficient to investigate the POC hypotheses (2). Statisticians and clinicians traditionally design POC trials with the concepts of type I and II error in mind. These refer, respectively, to the false-positive and false-negative rates due to chance findings as a result of the finite sample size in a POC trial. False-positive results lead to phase III trials that are undertaken in error and will likely lead to negative results at great expense. False-negative results lead to the wrong conclusion that the drug is ineffective for the indication, resulting in a loss of opportunity. In our analysis, we consider not only type I and II errors, but define a type III error, in which a POC trial that would have been successful is not performed. Traditionally, phase II POC trials are designed to have a type I error rate of 5% or 10% and a type II error rate of 10% to 20% (the "power" is 100% minus the type II error, thus, 80%–90% in this case) for detecting a treatment effect size of clinical interest in a qualified surrogate endpoint such as progression-free survival (PFS) in a randomized study or response rate in a single-arm study. So engraved is this tradition that POC trials with less than 80% power are often termed "underpowered," even though the traditional powering still allows significant room for error, and "perfect" POC trials would require infinite sample sizes. But in fact, there is no absolute scientific basis for
selecting particular type I and II error rates in POC trials. These are simply a function of risk tolerance, which is in turn a function of strategy. Indeed, we observe an alternative style of smaller “underpowered” trials being executed in many cases.

Chen and Beckman investigated whether optimal type I and II error rates could be objectively defined for randomized phase II trials by requiring that the efficiency of phase II and phase III development be maximized (3–8). The methodology has been used to address various related issues such as resource allocation across multiple POC candidate trials, futility bars in phase III, design of seamless phase II/III trials, and development of personalized medicines in phase II/III.

In this article, we present the high-level design strategy for individual POC trials and overall POC programs under a fixed budget derived from the same methodology.

Materials and Methods

Consider a typical randomized phase II POC trial of two arms with a 1:1 randomization ratio (study drug vs. placebo, or standard-of-care plus study drug vs. standard-of-care plus placebo). The trial has type II error rate \( \beta \) for detecting a minimum treatment effect size \( \delta \) of clinical interest at one-sided type I error rate \( \alpha \). Although the totality of data will be looked at closely after completion of the trial, a Go decision to a phase III confirmatory trial is generally made if the one-sided \( P \) value from the POC trial based on a test statistic is less than \( \alpha \) favoring the study drug. The empirical Go-No Go (GNG) bar, corresponding to the minimal observed effect size for a Go decision, increases when \( \alpha \) decreases or when \( \beta \) increases. There are infinitely many ways to choose \( \alpha \) and \( \beta \) without changing the sample size. Which choice is optimal? More importantly, what if the sample size itself is also subject to optimization?

Given the likelihood that the possible expenditure on POC trials of interest in oncology usually exceeds available patient and financial resources, we try to optimize \( \alpha \) and \( \beta \) with respect to a benefit–cost ratio defined as the risk-adjusted number of truly effective drug/indication combinations identified by POC trials (benefit) divided by the risk-adjusted number of patients used in phase II and III trials (cost). Risk adjustment takes into account the risks inherent in type I and II errors. The risk-adjusted benefit includes the risk of failing, as a result of a single POC trial, to bring a truly effective drug to phase III (single-trial false negative: type II error). The risk-adjusted cost includes the risk of involving patients in a phase III trial using an ineffective experimental drug that was felt to be effective based on the POC trial (false positive: type I error). The benefit–cost ratio measures how much a patient contributes to development of an active drug, and its inverse measures how many patients it takes to develop an active drug (see Fig. 1 for an example).

The benefit for each POC trial indication \( B \) is the probability that the drug is truly effective \( (1 - \beta) \) multiplied by the probability that the POC trial will detect its effectiveness \( (1 - \beta) \); \( \beta \) is the false-negative rate of the POC trial indication:

\[
B = P(1 - \beta)
\]

The total cost of the trial program \( C \) is the sum of two parts for each drug indication: the size of the phase II POC study \( C(2) \) and the size of the phase III pivotal study \( C(3) \). Cost may directly refer to the number of patients exposed (sample size); alternatively, if measured in monetary terms, one must estimate the cost as a function of the sample size. The probability that the phase III study will occur is the sum of two parts: the probability that the drug is truly effective \( (1 - \beta) \) multiplied by the probability that this will be detected in the POC trial \( (1 - \beta) \), and the probability that the drug is truly ineffective \( (1 - P) \) multiplied by the probability of getting a false-positive result in the POC trial \( (\alpha) \):

\[
C = C(2) + C(3)[P(1 - \beta) + (1 - P)\alpha]
\]

The efficiency of a single POC trial and its associated phase III trial is given by the ratio \( B/C \). For an overall program with multiple POC studies and possible phase III trials, the total benefit and cost are each summed up over all the POC studies and indications in the proposed program, and the ratio of these two sums is taken (3–5).

Results

Required inputs into the benefit–cost ratio analysis are: (i) the probability of the drug being truly active with an effect size of at least \( \delta \) (a component of probability of success). Organizations that develop multiple therapies will often attempt to estimate this parameter as part of designing the program and prioritizing among different POC trials and therapies for resource allocation. It is a subjective assessment based on scientific knowledge and preclinical data. Although it is difficult to estimate this precisely, the results are not highly sensitive to this parameter, and (ii) the relative sample size between phase II and III studies given identical type I/II error rates. [This does not mean that the phase II and III trials will actually be designed to have identical type \( I \) and II error rates. It is a required piece of information for the mathematics, and it relates to the different endpoints of phase II and III. For example (9, 10), it usually takes more patients to investigate an overall survival (OS) endpoint (phase III) than a PFS endpoint (phase II) because the percentage improvement in an OS endpoint will usually be less than the percentage improvement in the corresponding PFS endpoint. This is one reason phase III studies are larger than phase II studies. The other reason is the lower type I and II error rates that are used in phase III. For a typical oncology program, this relative sample size number is about 20% to 30%.] Table 1 provides the optimal design of phase II POC trials when the probabilities of success are in a range from 10% to 50% and the relative sample sizes are in a range from 10% to 30%, both of practical interest.

Under the optimal design with the maximum benefit–cost ratio, the type I error rate ranges from 2% to 8% and the optimal type II error rate ranges from 40% to 52%. Thus, the power, which is 100% minus the type II error, is 48% to 60% at the optimum for detecting an effect size of \( \delta \). The
Figure 1. Diagrams show the logic of a POC program for a drug with a probability of being active of 30% and a minimally clinically significant PFS HR of 0.6. We compare one large POC trial with two small ones. All possible development outcomes are shown. Each arrow is labeled with its associated outcome and probability of occurrence. We assume the drug requires a 600-patient phase III study. On the left is a decision diamond reflecting the unknown effectiveness of the drug. In the upper branch, the drug is truly active; in the lower branch, it is truly inactive. An active drug proceeds to the upper middle diamond. The POC trial either produces a false negative (probability equal to $b$) or a true positive (probability equal to $1/b$). An inactive drug proceeds to the lower middle diamond. The POC trial either produces a false positive (probability equal to $a$) or a true negative (probability equal to $1/a$). The benefits and costs of the associated outcomes are shown in the rectangles on the right. Only true positives produce a benefit. Both true positives and false positives are associated with the additional cost of phase III. The chance of reaching any of the rectangles is the product of all the arrows on the path to that rectangle. The risk-adjusted benefit is the sum of the benefits in each rectangle, each multiplied by the chance of reaching that rectangle. The risk-adjusted cost is the sum of the costs (in patients) in each rectangle, each multiplied by the chance of reaching that rectangle. The efficiency is the risk-adjusted benefit divided by the risk-adjusted cost, A, a single traditional POC trial with 80% power for the minimum effect size $d$ is performed, at a cost of 160 patients. The risk-adjusted benefit is 0.24 true positives developed through phase III. The risk-adjusted cost is 346 patients. (Continued on the following page.)
variability is much smaller than the ranges for the two inputs, suggesting that the results are not strongly dependent on the probability of the drug being truly active or on the relative sample size number. There is a trend to higher optimum error rates if the drugs are likely to be active or if the sample size ratio is closer to 1 (i.e., phase III size is relatively small). Furthermore, the results do not directly depend on the actual value of $\delta$ as long as it is defined as the minimum effect size of clinical interest, such that a drug with true effect size of $\delta$ or more is considered a true positive. However, once the type I and II errors are chosen, the resulting sample size depends on $\delta$.

The optimal power is lower than the traditional phase II power. It determines a new study design with a smaller sample size. The efficiency gain of the proposed design comes from reduced sample size. For example, when a HR of 0.6 in PFS is of clinical interest, the proposed design only needs 42 events. As a comparison, the conventional design with 5% type I error rate and 20% type II error rate will need approximately 95 events. The sample size and type I error determine a power curve: i.e., the larger the effect size, the easier it is to detect and, therefore, the more power. Thus, the optimal design result could be expressed in a different way: Rather than highlighting the fact that the optimal designs have lower power for the traditional effect size, one could ask what effect size these optimal studies can detect at the traditional power. Viewed this way, the optimal designs have the traditional power of approximately 80% for detecting a treatment effect size approximately 50% greater than $\delta$ (i.e., 1.5$\delta$). Thus, the optimal study is designed at traditional power for larger effect sizes (or equivalently at lower power for traditional effect sizes (Table 1).

This has an interesting implication for the actual GNG bar, which is the minimum effect size that one has to actually observe in the study to obtain statistical significance at the level $\alpha$. Ordinarily, a study result can achieve statistical significance with an observed result that is less than the effect size it is powered on. Typically, a study powered on $\delta$ at the traditional power of 80% at 5% type I error will be positive with observed effect sizes as low as 0.66$\delta$. In contrast, the optimal study, which has 48% to 60% power for $\delta$ (or 80% power for approximately 1.5$\delta$), will typically only be positive if the observed effect size is approximately $\delta$.

To implement the optimal designs, one needs to choose an estimate of the probability that the drug is truly effective with an effect size of at least $\delta$ and an estimate of the ratio of study sizes between phase II and III given identical inputs, suggesting that the results are not strongly dependent on the probability of the drug being truly active or on the relative sample size number. There is a trend to higher optimum error rates if the drugs are likely to be active or if the sample size ratio is closer to 1 (i.e., phase III size is relatively small). Furthermore, the results do not directly depend on the actual value of $\delta$ as long as it is defined as the minimum effect size of clinical interest, such that a drug with true effect size of $\delta$ or more is considered a true positive. However, once the type I and II errors are chosen, the resulting sample size depends on $\delta$.

The optimal power is lower than the traditional phase II power. It determines a new study design with a smaller sample size. The efficiency gain of the proposed design comes from reduced sample size. For example, when a HR of 0.6 in PFS is of clinical interest, the proposed design only needs 42 events. As a comparison, the conventional design with 5% type I error rate and 20% type II error rate will need approximately 95 events. The sample size and type I error determine a power curve: i.e., the larger the effect size, the easier it is to detect and, therefore, the more power. Thus, the optimal design result could be expressed in a different way: Rather than highlighting the fact that the optimal designs have lower power for the traditional effect size, one could ask what effect size these optimal studies can detect at the traditional power. Viewed this way, the optimal designs have the traditional power of approximately 80% for detecting a treatment effect size approximately 50% greater than $\delta$ (i.e., 1.5$\delta$). Thus, the optimal study is designed at traditional power for larger effect sizes (or equivalently at lower power for traditional effect sizes (Table 1).

This has an interesting implication for the actual GNG bar, which is the minimum effect size that one has to actually observe in the study to obtain statistical significance at the level $\alpha$. Ordinarily, a study result can achieve statistical significance with an observed result that is less than the effect size it is powered on. Typically, a study powered on $\delta$ at the traditional power of 80% at 5% type I error will be positive with observed effect sizes as low as 0.66$\delta$. In contrast, the optimal study, which has 48% to 60% power for $\delta$ (or 80% power for approximately 1.5$\delta$), will typically only be positive if the observed effect size is approximately $\delta$.

To implement the optimal designs, one needs to choose an estimate of the probability that the drug is truly effective with an effect size of at least $\delta$ and an estimate of the ratio of study sizes between phase II and III given identical $\alpha$ and $\beta$, and find the optimum values of type I and type II error rates using the standard software code previously published (3). One may also find an approximate optimum by directly comparing the benefit–cost ratios under different $\alpha$ and $\beta$ values of interest. Finally, an approximately optimal design may be found for most practical cases by simply powering the study at 80% for detecting an effect size of 1.5$\delta$ at 5% type I error rate. Once optimal $\alpha$ and $\beta$ have been selected, the sample size in patients or number of events is determined.

**Discussion**

The type II error rate under the proposed design is higher than traditionally recommended. Although a low type II error rate is always desirable, it necessitates larger POC trials and, therefore, fewer trials due to fiscal constraints. Some worthy hypotheses will, therefore, fall below the funding line. The opportunity cost of missing POC trials that might have identified a true positive, due to running a small number of large POC trials under a fixed budget, has been termed type III error (7). Although conventional statistics

---

(Continued.) The efficiency is $6.94 \times 10^{-4}$. B, two small POC trials (in different indications of equal merit), each with approximately 60% power for the minimum effect size $\delta$, are performed at a cost of 80 patients each (only one is shown), using the same fixed POC budget as in A. The risk-adjusted benefit is 0.18 true positives developed through phase III for each POC trial, for a total of 0.36 true positives developed through phase III. The risk-adjusted cost, including both small POC trials and all the phase III trials resulting from them, is 460 patients. The efficiency is $7.83 \times 10^{-1}$, 13% higher than that of the program with a single traditional POC trial (A). The second indication would not have been explored in A due to the fixed POC budget. This type III error reduces the efficiency of A, which is less efficient despite its lower type II error per POC trial.
focuses on type I and II errors only, a consideration of type III error is critical to identifying an optimal POC strategy. Although some active indications will be missed because of the higher type II error, this is more than compensated for by the reduction in type III error inherent in testing more POC hypotheses in the total program, often in parallel. By considering all three types of error, we minimize (but do not eliminate) missed active indications in the total system of trials. For highly targeted agents that have only one applicable POC trial, this paradigm may not apply (however, some highly targeted agents may address the same molecular lesion across tumor types and, thus, have several applicable POC trials).

Even for a highly targeted agent with a single POC trial, the paradigm may apply if the agent is part of a portfolio of POC trials across multiple drugs funded by the same finite budget. In this article, we have discussed a single POC trial in isolation, but the same methods can be used to optimize resource allocation across a portfolio of POC trials representing one drug or a portfolio of drugs, including those with hypotheses of unequal merit (3–5). POC trials corresponding to hypotheses with more clinical value or stronger scientific support will get more than their share of the resources when the efficiency function is optimized across a portfolio. In the extreme case, the hypotheses with the greatest merit will be tested in larger trials and the some of the weaker hypotheses will not be tested, mirroring the traditional paradigm. POC trials of highly targeted agents may receive a greater resource allocation if their scientific basis is strong.

The strategy of conducting more small POC trials with high GNG bars may not be new to some drug developers as they may have come to the same conclusion based on experience. However, to our best knowledge, the benefit–cost ratio analysis is the first objective analysis that provides theoretical support of this strategy and quantitative guidance for its application.

The addition of futility analyses in phase III will help control the total program type I error from multiple small trials, and alter the optima presented here slightly (4). Because futility analyses reduce the average expected sample size in phase III, they will tend to lead to higher optimal type I and II error rates in the preceding POC trial (i.e., even smaller trials). POC trials should be considered together with their associated phase III trials as a unit to optimize development efficiency.

In some cases, a larger POC trial may be suggested by other considerations such as multiple objectives or endpoints, or need to acquire a larger safety database. Despite these considerations, the general strategy of conducting more small POC trials with high GNG bars remains essential for drug development to be cost-effective.

With the unprecedented number of new oncology drug targets, it is most cost-effective to conduct small POC trials and set the GNG bars high so that more POC trials can be conducted under socioeconomic constraints. This strategy facilitates maximal drug development results while conserving demands on patient populations and public and private financial resources. When $\delta$ is the minimum treatment effect size of clinical interest in phase II, the study design with the highest benefit–cost ratio will have approximately 5% type I error rate and 20% type II error rate (80% power) for detecting a treatment effect size of approximately 1.5$\delta$. A Go decision to phase III is made when the observed treatment effect size is approximately $\delta$. Broad application of this design will maximize return on socioeconomic investment in phase II POC trials.

Disclosure of Potential Conflicts of Interest

R.A. Beckman reports receiving ownership interest (including patents) in Daiichi Sankyo Pharma Development and Johnson & Johnson. No potential conflicts of interest were disclosed by the other author.

Authors’ Contributions

Conception and design: C. Chen, R.A. Beckman
Development of methodology: C. Chen, R.A. Beckman
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C. Chen
Writing, review, and/or revision of the manuscript: C. Chen, R.A. Beckman
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): C. Chen

Received August 21, 2013; revised January 11, 2014; accepted February 9, 2014; published OnlineFirst February 13, 2014.
Maximizing Return on Socioeconomic Investment in Phase II Proof-of-Concept Trials

Cong Chen and Robert A. Beckman


Updated version

Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-13-2312

Cited articles

This article cites 10 articles, 3 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/20/7/1730.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.