

Statistical Issues in Clinical Trial Design

Kenneth R. Hess, PhD

Corresponding author

Kenneth R. Hess, PhD

Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 447, Houston, TX 77030-4009, USA.

E-mail: khess@mdanderson.org

Current Oncology Reports 2007, 9:55–59

Current Medicine Group LLC ISSN 1523-3790

Copyright © 2007 by Current Medicine Group LLC

This paper reviews some of the more salient recent developments in statistical aspects of clinical trial design methodology and suggests that more emphasis be placed on rigorous assessments of the new methodologies to place them in the context of existing methods, demonstrate their claimed advantages, and fully disclose their remaining limitations.

Introduction

Oncology clinical trials have traditionally been classified according to phases (Table 1). Phase I studies identify dose-limiting toxicities and maximum tolerated doses. Phase II studies yield preliminary assessments of treatment efficacy, typically in an uncontrolled fashion using subjective, short-term endpoints. Phase III studies are large randomized trials with definitive endpoints such as survival. Although this phase nomenclature is increasingly inadequate to categorize clinical oncology trials and may act to inhibit innovation, it facilitates communication and presents a useful baseline for discussion. The basic I → II → III sequence in drug development remains largely unchanged, but a blurring of the distinctions is becoming more common. Phase I studies may include efficacy assessments as part of the dose escalation, phase II studies may include randomization to multiple treatments, and phase III studies may use subjective endpoints such as progression-free survival.

Hundreds of papers have been written over the past few decades on new statistical developments in clinical trial methodology. This paper reviews some of the more salient recent advances and suggests that more emphasis be placed on rigorous assessments of the new methodologies to clearly place them in the context of existing methods, demonstrate their claimed advantages, and fully disclose remaining limitations.

Phase I Trials

The classic phase I oncology study seeks to assess the toxicity of a new treatment by sequential dose escalation. The most common approach follows what is called a “3+3” design. At each step in the process, actions are explicitly prescribed for each potential outcome (eg, if 0 out of 3 patients at a given dose have significant toxicity, escalation proceeds). Despite scores of papers by statisticians pointing out the poor statistical properties of this algorithm, it remains by far the most common design employed. This seeming paradox likely reflects a fundamental difference in how statisticians and clinicians view the goals of phase I studies. Statisticians tend to think in terms of the desire for precise estimates of the maximum tolerated dose (ie, the lowest dose at which patients experience the highest acceptable risk of significant toxicity—generally around 30%). In contrast, clinicians seem to be satisfied with having sufficient assurance that the selected dose is reasonably safe (but not so low as to be ineffective) and that the most common dose-limiting toxicities have been identified. These two similar but distinct goals of estimation versus selection seem to support different designs [1]. For clinicians, the added flexibility and precision of newer model-based designs do not seem to outweigh their complexity. The “3+3” design, though not optimal, is simple, easy to understand, and yields acceptable values for the maximum tolerated dose. As more emphasis is placed on biologic endpoints and early assessments of toxicity-efficacy trade-offs, the model-based designs may play a larger role.

Phase II Trials

A classic phase II oncology study is a single-arm study with a binary endpoint and two stages (ie, a single interim analysis approximately midway through patient enrollment). This approach was popularized first by Gehan [2] and then by Simon [3]. An increasingly common variation is to include two single-arm studies for the same patient population in the same protocol and to randomize patients between the arms. This design is seen as preferable to running the two trials sequentially because randomization acts to reduce differences in patient characteristics between the arms. Sometimes the second arm in a randomized phase II study is a control.

Table 1. Phases of oncology clinical trials

- I. Initial toxicity assessment
- II. Initial efficacy assessment
- III. Definitive efficacy assessment
- IV. Post-market surveillance

This is done when there is concern that the historical values used for comparison may not be appropriate either because of possible changes in elements, such as supportive care, or because of differences in patient characteristics (either of which might affect patient outcome—even more than novel but unproven experimental treatments). Randomization is a key component of sound scientific inference and is increasingly used in phase II studies but often without sufficient patient numbers to yield valid comparisons between the arms of the study [4].

The traditional endpoint for phase II studies has been tumor response (ie, marked tumor shrinkage or complete tumor disappearance), but increasingly alternatives, such as progression-free survival, are used instead [5]. These newer endpoints are particularly popular for newer so-called targeted agents (eg, tyrosine kinase inhibitors), which may prove to delay tumor progression rather than shrink tumors outright. Although, traditionally, the emphasis in endpoint assessment has been antitumor activity (ie, efficacy), increasing attention is being paid in the phase II setting to including toxicity as a co-primary endpoint [6]. By including toxicity as an endpoint, we not only allow formal rules for stopping a trial early because of excessive toxicity but we have quantified the probabilities for making incorrect decisions with respect to excessive toxicity (ie, the probability of false-negative and false-positive conclusions). Designs providing for dose escalation using rules based on monitoring both toxicity and efficacy have also been developed [7].

A particularly salient criticism of the traditional two-stage design is the dilemma faced when the treatment has failed the first dozen or so patients [5]. Most traditional two-stage designs have 15 to 20 patients in the first stage. Investigators are faced with the troubling situation of needing to fill the initial stage of the design to fulfill the protocol requirements along with the increasing realization that the odds of the treatment being effective in this patient population might be unacceptably low. An increasingly common solution to this problem is to use a more sophisticated design that allows for more frequent monitoring and early stopping [8]. Usually referred to as a Bayesian approach because it borrows heavily from Bayesian statistical methodology, this approach offers substantial improvements in flexibility at the cost of increased complexity. In this approach, emphasis is placed on the distribution of possible success

Table 2. Basic steps in single-endpoint Bayesian clinical trial designs

- 0 Formalize endpoint information as probability distribution
- 1 Specify initial distribution for results of standard treatment (S)
- 2 Specify initial distribution for results of experimental treatment (E)
- 3 Specify desired improvement (d) for E over S
- 4 Let P be the probability that E exceeds S by at least d (ie, $P = \text{Prob}[E > S + d]$)
- 5 Specify a lower limit (L) for P , such that when $P < L$, the trial is stopped
- 6 Update distributions for S and E as trial result data accumulate
- 7 Monitor P , stopping if P falls below L

(Adapted from Thall and Simon [8].)

probabilities for the experimental treatment (Table 2). Initially this distribution is taken to encompass a large range of possibilities in order to reflect our uncertainty in the success of the new treatment. The distribution quantifies our uncertainty, and the mathematical machinery that underlies this approach allows us to update this distribution as data accumulate [9••]. At any point in the trial we can compute the probability that the new treatment exceeds the success rate of the historical comparison by some specified amount. When this probability falls below a specified lower limit, we stop the trial [8]. This approach to frequent monitoring of trial results can easily be applied to data from conventional trials in order to facilitate decision making and possibly terminate a study of an ineffective treatment as soon as sufficient data become available to support such a decision. The same basic approach can be used if we wish to take the historical values for the standard treatment as constant (the approach used in the conventional Simon and Gehan phase II designs) or if we wish to include an honest assessment of our uncertainty in the data for the standard treatment (but the probability calculations are a bit more complicated in the latter case).

An added advantage of the Bayesian approach is that it easily adapts to the use of a time-to-event endpoint [10•]. For endpoints such as progression-free survival this is a particular advantage because these endpoints are often conceptualized in terms of time to failure rather than the occurrence or non-occurrence of the event by some specified time after start of treatment. This Bayesian approach has also been adapted for situations in which multiple primary endpoints are desired (eg, toxicity and efficacy) [11]. These enhanced designs not only allow us to quantify the probabilities of making incorrect decisions but also allow us to stop the trial due to either excessive toxicity or excessive

Table 3. Suggested guidelines for establishing initial validity evidence for new statistical methodology in clinical trial design

- 1) State exactly what the method is intended to do or what properties it is intended to have in objectively testable terms
- 2) State explicitly the assumptions under which these properties or outcomes are evaluated and the limitations of these assumptions
- 3) Provide evidence that the method has the claimed performance or properties from simulation studies and outcomes of actual trials
- 4) In the absence of compelling evidence as described in point 3, state clearly that the claimed properties are conjectured and await substantiation
- 5) Where alternative methods already exist, compare the properties of the new method with those of the existing methods (under both assumed circumstances and realistic circumstances that diverge from those assumed)

(Adapted from Mehta et al. [15••].)

ineffectiveness (ie, insufficient tumor control) as soon as sufficient data become available to make the decision with reasonable confidence.

The opportunity for frequent monitoring and early termination is a natural adjunct to situations in which multiple treatments might be assessed simultaneously. By including multiple arms with random allocation between arms, we can generate a design that acts to select from a variety of candidate treatments [4]. These so-called selection designs can have either binary endpoints or time-to-event endpoints. Although they are clearly more complex than mounting a single-arm study, such designs can be useful for facilitating decision making when multiple experimental treatments are available for a single patient population. Unlike conventional methods, the typical basic Bayesian methods for monitoring a multi-arm study are the same as for monitoring a single-arm study. In the case of the single-arm study, the data for the standard treatment are not updated, whereas for a multi-arm study (with standard treatment as one arm) they are. These designs can be further enhanced by adding such elements as covariate-adjusted allocation to improve balance between the treatment arms with respect to potentially confounding patient characteristics [9••]. They can even be augmented with outcome-adaptive randomization such that patients are preferentially allocated to arms with higher estimated success probabilities [9••].

An additional attraction for this approach is the capacity to test a particular experimental treatment in a variety of different patient populations that at least potentially may have similar outcomes [12]. The two usual options in such a circumstance are either to run two separate studies (one in each population) or to combine both populations into a single study. By creating a design that allows us to borrow information between arms when making decisions, we can facilitate decision making in this rather complicated situation. Although these designs are still rarely used in practice, with this new approach to trial design we can find an appropriate middle ground between the two extremes of a single

combined trial or two independent trials. As usual, this added flexibility comes at the cost of added complexity.

Phase III Trials

Traditional phase III oncology trials are randomized trials with survival endpoints, most often between an experimental treatment and a standard treatment for a particular patient population. Increasingly, however, the distinction between phase II and phase III is blurred. The US Food and Drug Administration (FDA) is allowing randomized trials with progression-free survival endpoints to be used to support drug approvals [13]. The FDA is also supporting the use of flexible Bayesian designs in such trials [14]. As indicated previously, such designs allow for frequent monitoring, covariate-adaptive randomization, and outcome-adaptive randomization [9••]. In fact, all of the issues raised in this review concerning multi-arm phase II studies apply to phase III studies as well (with an emphasis on randomization and time-to-event endpoints).

The Bayesian approach allows the prediction of future results in a trial with appropriate attention to the uncertainty of these predictions [9••]. These so-called “predicted probabilities” represent an ultimate in flexibility/complexity trade-offs. Although they are based on extraordinarily complicated mathematics, they offer the potential for huge savings in early trial termination, among other advantages.

Clinical Trial Methodology Evaluation

As new statistical designs for clinical trials are developed, they are presented at meetings and in papers aimed at applied biostatisticians (ie, those chiefly concerned with applying existing methodology as opposed to developing new methodology) and interested clinical investigators. Such developments should be viewed with guarded optimism. This is increasingly true as proponents of the Bayesian approach become more enthusiastic (and less cautious) in their advocacy of these methods.

It would be helpful to establish guidelines for evaluating these new clinical trial designs and to request that developers of new designs meet minimum requirements in terms of evidence that their innovations behave as advertised before they are accepted into wide use (Table 3). Mehta et al. [15••] have attempted to lay such a foundation for statistical methods developed for high-dimensional biology (eg, genomics and proteomics). They emphasize that explicit standards are needed to assess the validity of proposed methods. They also point out that, whereas the choice of which properties or outcomes of a method to study may be subjective, once these have been chosen, they can and should be evaluated objectively. Although the properties for a proposed design may apply over a wide set of circumstances, specific claims need to be supported by specific evidence and given appropriate qualifications.

Computer simulations are often used to establish properties of new designs, and the data scenarios under which the properties are assessed need to be explicitly identified and should be sufficiently broad to be clinically relevant. For example, invariably the data used to simulate the outcomes must come from a specified data distribution. The nature of this distribution is a fundamental assumption underlying the validity of the properties evidenced for the proposed method (eg, exponential distributions for time-to-event data). The assumptions underlying the development or assessments of the proposed design must be stated explicitly. The limitations of these assumptions (eg, when data do not exist to support or refute these assumptions, as with dose-toxicity relationships for phase I studies) should be discussed.

When one or more procedures already exist for the purpose of interest (eg, selecting a starting dose for an experimental treatment), it is not only essential to compare the newly proposed methods with the existing methods but also necessary to carefully describe the relative advantages of the new methods and at what cost (eg, greater complexity or logistical requirements) these advantages are achieved. Where one or more alternative methods already exist, the properties of the new method must be compared with those of the existing method(s). Emphasis should be given to comparing how the new and existing methods operate when the assumptions are violated (in clinically realistic ways). For example, Thall et al. [10•], compare their new approach to existing methods and examine what happens when monitoring becomes less frequent and when the data follow more complicated (and perhaps more realistic) distributions than might be assumed in the design.

One recurring theme in papers describing the properties of new methods is the limited nature of the clinical scenarios and data distributions used to assess the methods. An explanation for these limitations is the computational burden needed to yield operating characteristics for a proposed design over a wide range of clinically rel-

evant scenarios. Some of the newer methods require days or even weeks of computing time on even very fast computers. Nonetheless, it is incumbent on methodologists to substantiate their claims of superior performance with appropriate evidence and to demonstrate the robustness of their methods.

The Latest Bayesian Revolution

In the mid-1960s and again in the mid-1980s, Bayesian statisticians pushed to get their ideas and methods more widely used in clinical trial design. We are currently in the midst of a third wave of enthusiasm. Stakeholders (eg, investigators, funding agencies, regulating agencies, pharmaceutical companies, and insurance companies) will likely be more receptive this time around because, among other factors, more statisticians are familiar with Bayesian methods and the computational hurdles have been largely overcome. However, two major hurdles remain: 1) the added complexity and planning required for these designs; and 2) the reluctance of the proponents of these methods in admitting and overcoming these difficulties. Berry [9••], in his recent overview of Bayesian clinical trials, claims that the “mathematics of the Bayesian approach are quite simple.” Well, this is clearly not the case. But the increased difficulty is in proportion to the increased benefits accrued from these designs: while the mathematics may not be “quite simple,” they are not unduly complicated, especially in consideration of the benefits provided by the adoption of these methods. If the advocates of these methods wish to see their methods used more widely, they would be well-advised to face up to the added complexity and the need for far more extensive planning and logistics in establishing the value-added contributions of their approaches over existing methods.

When it is explained clearly and comprehensively (as by Berry [9••]), the added complexity of the Bayesian approach is much less daunting. Users of traditional approaches, such as the Simon two-stage design, do not need to think about the underlying mathematics of the statistical inference that drives the design. Similarly, once investigators become used to thinking in terms of a probability distribution for the success rate of a new treatment and how this can be updated as data from treated patients are accumulated, they do not need to think much about the fact that these methods are based on using Bayes’ formula to update a beta distribution. Nevertheless, many investigators are not yet comfortable with thinking in terms of probability distributions. Advocates of Bayesian methods must address this need for clearer and more basic descriptions of their methods if they want them to be widely embraced.

A key question is whether these designs must be customized by experts for each study or whether they can be used “off-the-shelf” as are simpler designs. That is, does each trial need to be custom built from scratch by local

methodology experts, or can they be crafted by applied biostatisticians by combining “off-the-shelf” components and using standardized simulation programs for testing and refinement? This situation is complicated by the need for specialized software to conduct the more complicated (adaptive) studies. One reason for the popularity of the Simon two-stage design is that easy-to-use software for creating new designs can be downloaded from the Internet or even accessed as Web-based tools. With these programs, the user simply inputs the required parameters and an appropriate design (if one exists) is output in a matter of seconds. Using boilerplate language for the statistical considerations section of a new protocol, a new design can be created and inserted into a new protocol in as short a time as 15 minutes. Far more time is spent in deciding the appropriate endpoints and corresponding parameters than in crafting a design once these preliminaries are decided.

Even for the simplest Bayesian design for continuous monitoring (eg, single-arm, binary endpoint), no widely available, user-friendly software exists. The most widely used software (Multc-99) is written for trials with multiple endpoints (but can be used for single-endpoint trials) and is not particularly user friendly [16]. The newer Multc-Lean program [16] is considerably more user friendly but is still unduly complicated because it addresses trials with multiple rather than single endpoints. The new program, TTEConduct, for designing single-arm, time-to-event monitoring trials is even more user friendly [16]. Thus, although a great deal remains to be done in terms of making user-friendly software available, considerable strides have been made. But there is a limit to the extent that these designs can be generated with off-the-shelf software. Also, adaptive designs actually need specially constructed software to run the trial. Not only is increased intellectual effort and planning required but also increased computer expertise and infrastructure. However, as with the increased complexity of the mathematics, these difficulties are in proportion to the increased advantages. The bottom line is that there is no free lunch. The advantages of these flexible Bayesian designs come at a price, but it is a price that stakeholders seem increasingly willing to pay [14].

Conclusions

Increasingly useful innovations in the statistical design of clinical trials are being developed and disseminated. These innovations greatly enhance the flexibility and acceptability of modern oncology clinical trials. Advocates of these new methods must address the need for 1) clear tutorial descriptions of their methods; 2) honest assessment of the increased complexity, logistics, and infrastructure required to run these trials; 3) widely available, user-friendly software; and 4) rigorous validation of their methodology if they want their methods to be as widely embraced as they deserve to be.

References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:

- Of importance
 - Of major importance
1. Rosenberger WF, Haines LM: **Competing designs for phase I clinical trials: a review.** *Stat Med* 2002, **21**:2757–2770.
 2. Gehan EA: **The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent.** *J Chron Dis* 1961, **13**:346–353.
 3. Simon R: **Optimal two-stage design for phase II clinical trials.** *Control Clin Trials* 1989, **10**:1–10.
 4. Estey E, Thall P: **New designs for phase 2 clinical trials.** *Blood* 2003, **102**:442–448.
 5. Schlesselman JJ, Reis IM: **Phase II clinical trials in oncology: strengths and limitations of two-stage designs.** *Cancer Invest* 2006, **24**:404–412.
 6. Bryant J, Day R: **Incorporating toxicity considerations into the design of two-stage phase II clinical trials.** *Biometrics* 1995, **51**:1372–1383.
 7. Thall PF, Cook JD: **Dose-finding based on efficacy-toxicity trade-offs.** *Biometrics* 2004, **60**:684–693.
 8. Thall PF, Simon R: **Practical guidelines for phase IIB clinical trials.** *Biometrics* 1994, **50**:337–349.
 9. •• Berry DA: **A guide to drug discovery: Bayesian clinical trials.** *Nat Rev Drug Discov* 2006, **5**:27–36.
An excellent overview of recent developments in Bayesian clinical trial methodology, this paper also includes an introduction into how Bayesian updating works in the context of a single binary endpoint. Only downside is limited attention paid to admitting and addressing issues related to increased complexity of these designs and the need for additional planning, logistics, and infrastructure to support them.
 10. • Thall PF, Wooten LH, Tannir NM: **Monitoring event times in early phase clinical trials: some practical issues.** *Clin Trials* 2005, **2**:467–78.
Extends the now-classic methods presented in the Thall-Simon 1994 paper [8] from binary endpoints to time-to-event endpoints. Methodology evaluation is more comprehensive than most, and user-friendly software is widely available [16].
 11. Thall PF, Simon RM, Estey EH: **Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes.** *Stat Med* 1995, **14**:357–79.
 12. Thall PF, Wathen JK, Bekele BN, et al.: **Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes.** *Stat Med* 2003, **22**:763–780.
 13. Johnson JR, Williams G, Pazdur R: **End points and United States Food and Drug Administration approval of oncology drugs.** *J Clin Oncol* 2003, **21**:1404–1411.
 14. Temple R: **FDA perspective on trials with interim efficacy evaluations.** *Stat Med* 2006, **25**:3245–3249; discussion 3326–3347.
 15. •• Mehta T, Tanik M, Allison DB: **Towards sound epistemological foundations of statistical methods for high-dimensional biology.** *Nat Genet* 2004, **36**:943–947.
An outstanding contribution to the field of statistical methodology. The authors emphasize that explicit standards are needed to assess the validity of proposed methods. They make several bold proposals that can easily be adapted to applications other than genomics. We have adapted their ideas to clinical trials methodology.
 16. Cook J: **Collection of software to design, assess, and conduct Bayesian clinical trials.** <http://biostatistics.mdanderson.org/SoftwareDownload>. Accessed 7/3006.