A review of phase 2–3 clinical trial designs

Peter F. Thall

Abstract   This article reviews phase 2–3 clinical trial designs, including their genesis and the potential role of such designs in treatment evaluation. The paper begins with a discussion of the many scientific flaws in the conventional phase 2 → phase 3 treatment evaluation process that motivate phase 2–3 designs. This is followed by descriptions of some particular phase 2–3 designs that have been proposed, including two-stage designs to evaluate one experimental treatment, a design that accommodates both frontline and salvage therapy in oncology, two-stage select-and-test designs that evaluate several experimental treatments, dose-ranging designs, and a seamless phase 2–3 design based on both early response-toxicity outcomes and later event times. A general conclusion is that, in many circumstances, a properly designed phase 2–3 trial utilizes resources much more efficiently and provides much more reliable inferences than conventional methods.

Keywords   Adaptive design · Clinical trial · Design · Phase II clinical trial · Phase III clinical trial · Selection

1 Introduction

When a new treatment for a particular disease is proposed, the key question is whether the treatment is sufficiently safe and efficacious to be adopted by physicians as part of their routine clinical practice. Except for diseases that are completely intractable, addressing this question requires comparison of the new treatment to existing standard therapies. Inevitably, the evaluation process also involves ethical, logistical, economic

P. F. Thall (✉)
Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center,
1515 Holcombe Boulevard, Houston, TX 77030, USA
e-mail: rex@mdanderson.org

and regulatory issues. To deal with this complexity, in recent decades the medical and scientific communities have developed an elaborate infrastructure for conducting clinical trials, which are experiments to evaluate the effects of medical treatments on human subjects. Since there is a high degree of variability between patients in the effects of a given treatment, statistical methods are required for clinical trial design, conduct and analysis. While the simplicity of the conventional "phase 1 → phase 2 → phase 3" paradigm for treatment evaluation and drug development is appealing, unfortunately the validity of this paradigm relies on a number of assumptions that are at odds with good statistical practice. Moreover, in practice clinical trials are often much more complex than the structures assumed by conventional designs, due to the heterogeneity of different diseases, treatment regimes and processes for evaluating patient outcomes. Consequently, the conventional three-phase paradigm often is inadequate or dysfunctional.

Many members of the medical community regard statistics with some degree of skepticism. The quote, attributed to Benjamin Disraeli, that "There are three kinds of lies: lies, damned lies, and statistics" is amusing primarily because it reflects the genuine confusion that many non-statisiticians feel when confronted by statistical arguments. It also reflects the mistrust caused by the not uncommmon practices of dishonestly manipulating a complex statistical argument, or of citing statistics that have no empirical or observational basis, to support a preconceived conclusion. Additionally, statistics relies on probability, which can be very non-intuitive. Consequently, statisticians must provide transparent, convincing explanations of their methods to non-statisticians. Evidently, the latter task has proved to be very difficult in the clinical trial arena, since a large gap remains between the best available statistical methods and what most physicians are willing to use to design and analyze their trials. This is especially true of the family of "phase 2–3" trial designs, which combine aspects of conventional phase 2 and phase 3 trials.

In this paper, I will first review the main ideas that motivate phase 2–3 designs, including scientific flaws in the conventional phase 2 → phase 3 treatment evaluation process. The remainder of the paper will be devoted to describing some particular phase 2–3 designs. What follows pertains primarily to oncology trials, since most of my experience is in this area, and some statements may not true for trials in other diseases.

2 The conventional three-phase paradigm

To begin, I will discuss settings where there is one experimental treatment, E, to be evaluated, and later consider the more complex problem of evaluating several new treatments that are available at the same time. In its simplest form, the three-phase clinical trial paradigm begins with a phase 1 trial to determine a safe dose of E. The dose-finding procedure typically is based on adverse treatment effects ("toxicity") while ignoring anti-disease effect. This is followed by a single-arm phase 2 trial in which each patient's outcome is characterized by a binary variable, Y, indicating whether a desirable response to therapy with E has been achieved. If the observed response rate is sufficiently high, this is used as the rationale to conduct a confirmatory

phase 3 trial. Since the topic of this paper is phase 2–3 trials, I will not elaborate on the many advantages of using efficacy as well as toxicity to determine an optimal dose, or of optimizing treatment administration schedule rather than choosing it arbitrarily. The interested reader may refer to Thall and Cook (2004) or Braun et al. (2007), and the references therein.

Let S denote an established "standard" therapy. Phase 3 trials are considered the most scientifically reliable tool for confirmatory evaluation of new treatments. In phase 3, patients are randomized between E and S in order to obtain unbiased comparisons, which are based on an outcome, T, that characterizes long-term or permanent treatment effect. In most phase 3 trials, T is a time-to-event (TTE) variable such as overall survival (OS) time or disease-free survival (DFS) time. A phase 3 design must control the overall false positive and false negative rates at suitably small values. Usually, one or more interim tests are performed during the trial (Jennison and Turnbull 2000) to avoid continuing to randomize patients after it has been determined that one treatment is superior to the other. Less frequently, the design also includes additional interim "futility" rules to stop the trial early and accept the null hypothesis that there is no difference between E and S. Statistical sample size computations to achieve a given power (probability of correctly detecting a true treatment difference of given magnitude) under a hypothesis testing framework typically require phase 3 trials to be large, time-consuming and expensive. This is the main reason that phase 2 trials are conducted first, in order to reliably screen new treatments before committing the resources required by phase 3.

The phase 2 outcome, Y, should characterize treatment benefit for the particular disease in a convincing manner. Examples include achieving $> 50\%$ shrinkage of a solid tumor; complete remission (CR) of leukemia in terms of the levels of platelets, blastic (immature) blood cells and white blood cells; and reperfusion of blood flow to the site of an acute stroke. From a logistical standpoint, Y must be an "early outcome," observed much sooner than T. In the above examples, solid tumor shrinkage and CR each is scored after a 4–6 week period of therapy, whereas reperfusion after a stroke often is evaluated within 24–48 h. The idea is that, compared to phase 3, a phase 2 trial should be much smaller and completed much more quickly, in order to provide a feasible means to screen E. This is based on the implicit assumption that there is a strong connection between Y and T, although this is formally ignored by most phase 2 and phase 3 designs. I will return to this important point below. The confirmatory phase 3 outcome should be tailored to the particular disease and treatment goals. For example, whereas OS or DFS is appropriate for most phase 3 oncology trials, in evaluating effects of rapid treatment of stroke the phase 3 outcome may be a stroke severity score quantifying long-term neurological or motor function.

A typical phase 2 trial is based on a single parameter, $\theta_E = \Pr(\text{response with } E) = \Pr(Y = 1|E)$. Sometimes the stated goal is simply to estimate θ_E with a given reliability, and sometimes θ_E is compared to $\theta_S = \Pr(\text{response with } S) = \Pr(Y = 1|S)$. A common frequentist approach is to assume a fixed null value of θ_S, base phase 2 on a one-sided test of $H_0 : \theta_E = \theta_S$ vs. $H_1 : \theta_S < \theta_E$ with given size and given power at a targeted alternative $\theta_S + \delta$, usually with $\delta = .10$ to $.20$, and proceed to phase 3 if the test rejects H_0. The design usually includes one or more interim futility tests that stop the trial early if H_0 is accepted (Fleming 1982; Simon 1989). Bayesian

phase 2 designs (Thall and Simon 1994) usually begin with a non-informative prior on θ_E, an informative prior on θ_S obtained by elicitation or from historical data, and base decisions on posterior probabilities of the form $\Pr(\theta_S + \delta < \theta_E | data)$. Frequentist operating characteristics of Bayesian designs, including false positive and false negative rates and sample size distributions, typically are established by computer simulation. More generally, Y may be multinomial in order to include toxicity as well as response, and more complex decision rules that include early stopping due to excessive toxicity may be employed (Bryant and Day 1995; Thall et al. 1995). In any case, the data on Y from the trial of E are used to decide whether E is sufficiently promising, compared to S, to justify proceeding with a phase 3 trial.

3 Problems with conventional methods

The genesis of phase 2–3 designs comes from recognition of several severe problems with the conventional paradigm, all arising from violation of basic statistical principles and practice. Claims to the contrary notwithstanding, phase 2 trials are inherently comparative (Simon et al. 1985; Thall and Simon 1994). If the goal is to determine whether E has any anti-disease activity at all (Gehan 1961) then θ_E is compared to 0. If an established standard therapy exists, then θ_E is compared to θ_S. If the goal is to rank and select among different treatments, $E_1, \ldots, E_k$, then $\theta_{E,1}, \ldots, \theta_{E,k}$ are compared among each other. In any case, one or more comparisons are made whether or not a formal test of hypotheses is done.

It is well-known that a statistical comparison between E and S based on data from a single-arm trial of E and historical data on S confounds the E-versus-S treatment effect $\theta_E - \theta_S$ with between-trial effects. Trial effects may be due to differences in entry criteria, supportive care, the definition of response, dose modification procedures, compliance patterns, or a myriad of unknown ("latent") variables (Estey and Thall 2003; Spiegelhalter et al. 2003). If an estimator $\hat{\theta}_E$ based on a single-arm trial of E and an estimator $\hat{\theta}_S$ based on historical data on S are used to compute $\hat{\theta}_E - \hat{\theta}_S$, this statistic does not estimate the actual treatment effect $\theta_E - \theta_S$. Rather, it estimates the combined effects of E-versus-S and whatever between-trial effects may be at play.

Ignoring variability by treating a statistic as if it were a constant is also a pervasive problem. This is implicit in test-based phase 2 designs. In practice, the assumed fixed value of θ_S used to define a null hypothesis for a one sample test of θ_E is actually a statistical estimator, $\hat{\theta}_S$, based on historical data (c.f. Thall and Simon 1990). Consequently, since the test statistic is based on a comparison of $\hat{\theta}_E$ to $\hat{\theta}_S$ while assuming that var($\hat{\theta}_S$)=0, the variance of the test statistic is under-estimated and thus both the false positive and false negative rates of the test are larger than their nominal values.

An additional source of variability and bias is patient heterogeneity, which generally is substantial in any clinical setting. For many diseases, the combined effects of prognostic covariates such as disease severity, age, performance status, etc., are much larger than any treatment effect. This explains the commonly seen statistical phenomenon that occurs if one computes the sum X_n of a sample $Y_1, \ldots, Y_n$ of binary outcomes, assumes these variables are iid with common probability p, and uses $\overline{Y} = X_n/n$ to estimate p. In such settings, the common statistical estimate $\sum_{i=1}^{n} (Y_i - \overline{Y})^2 / n(n-1)$ of var($\overline{Y}$)

is often much larger than the binomial model-based variance estimate $\overline{Y}(1 - \overline{Y})/n$, a phenomenon sometimes called "extra-binomial variation." In a clinical trial setting where the response rate $\Pr(Y = 1)$ is well-known to vary with patient covariates, extra-binomial variation is due to nothing more than the failure to account for covariate effects by fitting an appropriate binary outcome regression model, such as the logistic, to the available data. The common practice of comparing the observed response rate from an uncontrolled trial of E to an historical rate with S without accounting for patient covariates thus produces an estimate of $\theta_E - \theta_S$, or a test of $\theta_E = \theta_S$, that is likely to depend more on covariate imbalances between trials than any real treatment effect. This problem is manifested if physicians, consciously or unconsciously, choose patients having better prognosis for a single-arm trial of E in order to increase the chance of showing that E is promising compared to S, so-called "cherry picking." Physicians who engage in this practice often do so with the altruistic intention, based on an optimistic prior belief about the efficacy of E, that E should be given the best possible chance to show how well it works, so that it may benefit future patients. The simple model that, on average, [clinical effect] = [treatment effect] + [patient covariate effects] shows the flaw in this reasoning.

In addition to the practice of ignoring covariates, other valuable information often is wasted when making the decision of whether to proceed with phase 3. While the relationship between Y and T is essential to the rationale that a phase 2 trial based on Y should be the basis for deciding whether to conduct a phase 3 trial based on T, this assumption is almost never made explicitly. As I will show in Sect. 6, below, accounting for (Y, T) as a multivariate outcome using a simple mixture model greatly increases the reliability of both phase 2 and phase 3, and leads very naturally to a phase 2–3 design. The other side of this coin sometimes is seen in settings where no observable Y that is related to T is available. In such cases, investigators may use an early outcome, say Y^*, simply because it is available, despite the fact that no relationship between Y^* and T has been established. This is often motivated by the belief that a phase 2 trial based on a such an early outcome is better than no phase 2 trial at all. A common example of this practice is that in which Y^* indicates the presence of a particular biomarker, derived from laboratory experiments showing cancer cell killing in cell cultures or anti-tumor effects in rats, but Y^* is of no value whatsoever in predicting T in humans. The difficult step of validating Y^* by showing that it predicts T in humans is frequently ignored in so-called "translational" or "bench-to-bedside" research, due to the problem that it often is not practical to validate Y^* prior to phase III. It is a truism in modern clinical oncology research, however, that so-called "targeted" therapies very often have biological effects, Y^*, that are later found to be statistically unrelated to either tumor shrinkage or to T. Some examples are given by Thall et al. (Sect. 5.1, 2007). The point is not that new treatments motivated by laboratory-based research are not potentially useful, but rather that there are important distinctions between effects seen in cells, effects seen in animals, pre-clinical effects in humans, and actual clinical improvement in humans.

Consider the harder problem of evaluating several different treatments, $E_1, \ldots, E_k$, simultaneously in phase 2, either in one randomized trial or in separate single-arm trials. If the treatment $E_{[k]}$ having the largest observed response rate among the estimates

$\hat{\theta}_{E,1}, \ldots, \hat{\theta}_{E,k}$ is selected for phase 3 evaluation, the subsequent phase 3 test suffers from selection bias that inflates the false positive rate. In the null case where $\theta_{E,1} = \cdots = \theta_{E,k} = \theta_E$, that is, where the k treatments are equivalent to each other, all $\hat{\theta}_{E,j}$ have mean θ_E but the maximum $\hat{\theta}_{E,[k]}$ has mean $> \theta_E$, and moreover $\hat{\theta}_{E,[k]}$ increases stochastically with k. Intuitively, since one must observe all of the $\hat{\theta}_{E,j}$'s before $[k]$ can be identified, $[k]$ is a statistic that depends on the data from all k treatments. A phase 3 trial of $E_{[k]}$ vs. S that ignores this fact will have a false positive rate larger than its nominal value.

Selection bias also may arise in other, more subtle ways. In the process of developing a single new agent, E, it is a common practice for a pharmaceutical company to conduct phase 2 trials of E in several indications (disease areas), select the indication where E performs best, and then conduct a phase 3 trial in the selected indication. This preliminary selection among indications inflates the risk of a false positive decision. To see this, consider a simple model where the response probability with S in the jth indication is $\theta_{S,j}$ for $j = 1, \ldots, k$. Suppose that, for each j, the phase 2 trial conducted in the jth indication yields estimate $\hat{\theta}_{E,j}$ of $\theta_{E,j}$. Suppose that the best indication, having index denoted by $[k]$, is defined as that maximizing $\hat{\theta}_{E,j} - \theta_{S,j}$, and the company's strategy is to conduct a phase 3 trial of E in the indication $[k]$. If in fact E is completely equivalent to S in all indications, formally if $\theta_{E,j} = \theta_{S,j}$ for all j, then each $\hat{\theta}_{E,j} - \theta_{S,j}$ has expected value 0. However, the maximum of these k differences has expected value > 0. Moreover, $\hat{\theta}_{E,[k]} - \theta_{S,[k]}$ increases stochastically with k, so that the more indications that are tested the more likely it is that the phase 2 selection procedure will produce a false positive result.

Selection bias aside, a small trial can at best produce treatment effect estimators having limited reliability. The common practice of designing phase 2 trials to be as small as possible in order to get to phase 3 as quickly as possible suffers from two severe flaws. The first is that it ignores the simplest statistical principle of all, that the reliability of any valid statistical inference increases with sample size. The second flaw is the optimistic prior assumption, which is wrong much more often than it is right, that E will certainly provide a substantive clinical improvement over S. Such optimism often is given as a rationale for not using early stopping rules in phase 2, since such rules would risk depriving patients of the putative greater benefit of E. Of course, taken to its logical conclusion, one could use prior optimism about E rather than a phase 2 trial as a basis for proceeding to phase 3, then likewise do away with phase 3 since it would be unethical to randomize patients to S, and simply treat all future patients with E. As ridiculous as this may sound from a scientific viewpoint, replacing empirical treatment evaluation with prior optimism is actually a very common practice.

For any combination of these reasons, treatments found to be promising in phase 2 very often perform less well in phase 3, and most phase 3 trials yield negative results. This has resulted in an immense waste of resources, including time, money, drugs and patients. The designs that I will describe below are attempts, in some particular settings, to improve the scientific reliability and efficiency of the clinical evaluation process.

4 Phase 2–3 designs for one experimental treatment

Ellenberg and Eisenberger (1985) pointed out that a small, uncontrolled trial of E provides an estimate of θ_E that has very limited reliability, and also noted the common problem that treatments considered promising based on such studies often perform less well in subsequent phase 3 trials. They proposed a two-stage phase 2–3 design, based on binary outcomes, using a futility early stopping rule after stage 1. In stage 1, $2n$ patients are randomized between E and S. Denoting the numbers of responses in the two arms by $X_{E,n} = \sum_{i=1}^{n} Y_{E,i}$ and $X_{S,n} = \sum_{i=1}^{n} Y_{S,i}$, the trial is terminated early and E is rejected if $X_{E,n} \leq X_{S,n}$, and otherwise it continues to a second stage. The value of n is chosen to control the stage 1 early stopping probability, equivalently the false negative rate in stage 1, at a suitably small value under a given alternative $\theta_E = \theta_S + \delta$. If the trial is not stopped early, then a conventional comparative test based on a randomized trial of E vs. S is done in stage 2. One may regard this as a phase 2–3 trial with stage 1 the "phase 2" portion, and stage 2 the "phase 3" portion. A major departure from conventional phase 2 practice is that stage 1 includes a control arm, with patients randomized between E and S. While only the binary outcome case was presented, the general idea of randomizing in phase 2 and using a futility stopping rule could be applied more generally, e.g. in trials with TTE outcomes.

A formal version of this design that controls overall size and power and optimizes expected overall sample size was given by Thall et al. (1988a). To test the hypotheses $H_0 : \theta_E = \theta_S$ vs. $H_1 : \theta_E > \theta_S$, in stage 1, $2n_1$ patients are randomized between E and S, with H_0 accepted and the trial terminated early if an approximately normal test statistic Z_1 based on the two binomial samples from stage 1 is $\leq y_1$. If $Z_1 > y_1$ then $2n_2$ additional patients are randomized in stage 2. A final test statistic Z_2 based on the pooled data on $N_{max} = 2n_1 + 2n_2$ patients from both stages is computed at the end of stage 2, with H_0 accepted if $Z_2 \leq y_2$ and rejected if $Z_2 > y_2$. Given overall type I and type II error probabilities, the sample sizes n_1, n_2 and test cut-offs y_1, y_2 are chosen to minimize the expected overall sample size $E(N) = 2n_1 + 2n_2 \Pr(Z_1 > y_1)$, with this expectation computed as an equally weighted average $\frac{1}{2}E_0(N) + \frac{1}{2}E_1(N)$ under H_0 and under H_1 at $\theta_E = \theta_S + \delta$. For size .05 and power .80 to detect $\delta = .15$, the optimal designs require maximum samples sizes $N_{max} = 202$–292, with $E(N) = 128$–183 under H_0 and 191–276 under H_1. For $\delta = .20$ these values are much smaller, with $N_{max} = 104$–164. In either case, the design has early stopping probability .60 to .64, and if the trial is stopped early this results in a sample size $2n_1$ anywhere from 42 to 122. This design may be regarded as a randomized version of the widely used Simon (1989) single-arm phase 2 design. For example, an optimal Simon two-stage design to test $H_0 : \theta_E = .20$ with size .05 and power .80 to detect $\theta_E = .40$ requires $n_1 = 13$ patients in stage 1, $n_2 = 30$ in stage 2, with null early stopping probability .75. In comparison, the optimal randomized design requires $2n_1 = 56$ patients in stage 1, $2n_2 = 80$ in stage 2, and stops early with probability .62. The much larger investment of patients with the randomized trial ($N_{max} = 136$ vs. 43) is the price paid for doing away with trial effects and selection bias, accounting for variability in the estimate of θ_S, and thus obtaining unbiased comparisons between E and S. A limitation of the two-stage randomized design is that Y is used as the outcome in both stages, rather than using T in stage 2.

A related design for single-arm phase 2 trials with binary outcomes was given by Chang et al. (1987), who optimized the Fleming (1982) group sequential design using a sequential probability ratio test. To test the same one-sided hypotheses as given above, at each of up to k stages their design allows early stopping with either acceptance or rejection of H_0. Given overall size, power, k and the sample sizes $n_1, \ldots, n_k$ at all stages, they determined interim test cut-offs that minimize the overall expected sample size, equally weighted between H_0 and H_1 as above. They recommended application with $k = 2$ or 3, and small maximum sample sizes in the range 20–60. Under the same set of assumptions, Therneau et al. (1990) used an elegant geometric approach to improve this design. They provided a computational algorithm to ensure that the optimal design is admissible in the sense that it minimizes a linear combination of the type I error, type II error, and expected sample size. While these designs provide a formal way to choose the parameters of a Fleming design, they suffer from the limitations, noted above, of any single-arm design for which a fixed null value of θ_S is assumed.

5 Evaluating several experimental treatments

Suppose that several experimental treatments, $E_1, \ldots, E_k$, are available simultaneously for evaluation. Although in oncology each treatment often is a combination with a very specific schedule of administration involving successive courses, and treatments in other disease areas also may have multi-stage structures, since this can be quite complex I will ignore it here in order to focus on phase 2–3 issues. Thus, for now I will assume that outcome is one-dimensional and characterize the k average outcomes by the one-dimensional parameters $\theta_1, \ldots, \theta_k$. These may be response probabilities, median event times, or some other parameter of primary interest. Let θ_0 denote a fixed null value that is considered "not promising," corresponding to S. Several different but related goals may be addressed in this setting. The first is to identify the best among the E_j's, that is, identify the largest θ_j. A more difficult goal is to determine whether any E_j provides at least a given δ improvement over θ_0, while controlling both the false positive rate and false negative rate of the decision scheme. A much more difficult goal is to identify the specific set of all E_j's that provide a δ improvement. The simple but extreme approach of conducting a $k + 1$ arm randomized trial of $E_1, \ldots, E_k$ and S and performing all k pairwise comparisons of E_j v. S while controlling the overall false positive rate is prohibitively expensive in most cases. This consideration has motivated several phase 2–3 strategies that first select among $E_1, \ldots, E_k$, and then compare the selected treatment(s) to S.

In this setting, Simon et al. (1985) proposed that patients should be randomized fairly among the k treatments in phase 2, using conventional phase 2 sample sizes, and that the treatment arm $E_{[k]}$ having the largest observed response rate should be selected for phase 3 evaluation. In addition to randomizing to avoid bias, they proposed the use of ranking and selection rather than hypothesis testing. Randomized phase 2 trials are sometimes criticized on the grounds that they are underpowered phase 3 trials. The goals of the two types of trials are very different, however. The goal of a randomized phase 2 trial is simply to select the best treatment, $E_{[k]}$, among $E_1, \ldots, E_k$, in

a way that avoids the bias inherent in unrandomized comparisons. This is much less demanding than the goal of demonstrating that the best treatment is better than the other treatments, or better than S, by a predetermined amount δ while also controlling the false positive probability. While one may compute the probability of correct selection (PCS) of E_k in the simple case where $\theta_1 = \cdots = \theta_{k-1} = \theta_0$ and $\theta_k = \theta_0 + \delta$, the aim of a phase 2 selection trial is *not* to achieve this with a conventional power figure and small type I error. For example, if $k = 2$ and $\theta_1 = \theta_2$, then each treatment has a 50% chance of being selected. Because a simple randomized selection trial design does not attempt to control type I error, its conclusion is not that $E_{[k]}$ is substantively better than the other treatments, but rather that, empirically and based on unbiased comparisons, $E_{[k]}$ is the best. That is, the purpose of a randomized phase 2 trial is screening, not confirmatory evaluation.

As noted earlier, if one wishes to follow a randomized phase 2 selection trial by a phase 3 trial of the selected treatment $E_{[k]}$ vs. S, the phase 2 selection process may introduce bias into the subsequent phase 3 comparative test. In the k treatment setting with binary outcomes, Whitehead (1986) addressed the problem of determining how many patients should be treated in each of phase 2 and phase 3, if the total sample size N is predetermined. Denoting the number of patients per arm in phase 2 by n, a total of $N_1 = kn$ are randomized among $E_1, \ldots, E_k$ in phase 2, and $N_2 = N - N_1$ are randomized between $E_{[k]}$ and S in phase 3. Assuming that $\theta_1, \ldots, \theta_k$ are drawn randomly from a beta distribution, by averaging over the conditional power in phase 3 given $\theta_{[k]}$ Whitehead derived the probability that the phase 3 test would reject the null, and used these computations to obtain values of N_2 for given N and k that maximize the probability of rejecting in phase 3. For example, under a beta(2,8) prior with $\theta_0 = 0.20$, $k = 5$ and total sample size $N = 200, 300$ or 500 the corresponding optimal phase 3 sample sizes are 140, 210, and 345. The design strategy proposed by Whitehead did not include the data from stage 1 in the final two-sample test statistic, however.

Thall et al. (1988b) proposed the following two-stage phase 2–3 design, which includes both a k-treatment selection in stage 1 and a two-arm comparison in stage 2 based on the pooled data from both stages while controlling the overall false positive and false negative error probabilities. In stage 1, $(k + 1)n_1$ patients are randomized fairly among $E_1, \ldots, E_k$ and S. If the approximately normal two-sample stage 1 test statistic Z_1 based on $\hat{\theta}_{[k]}$ and $\hat{\theta}_S$ is $\leq y_1$, the stage 1 cut-off, then the trial is terminated and the global null hypothesis $H_0 : \theta_1 = \cdots = \theta_k = \theta_0$ is accepted. If $Z_1 > y_1$ then an additional $2n_2$ patients are randomized between $E_{[k]}$ and S in stage 2, a test statistic Z_2 based on the pooled $E_{[k]}$ and S data from both stages is computed, and a final test is performed with the conclusions that $\theta_{[k]} > \theta_0$ if $Z_2 > y_2$ and H_0 if $Z_2 \leq y_2$. Importantly, Z_1 depends on the data from all $k + 1$ samples in stage 1, and Z_2 depends on all of the data from both stages. To utilize ranking and selection theory (Bechhofer et al. 1995), Thall et al. (1988b) specified fixed $\delta_1 > 0$ to be a marginal or uninteresting improvement and $\delta_2 > 0$ a larger value such that $\theta_0 + \delta_2$ would be a clinically meaningful improvement over θ_0. To account for the hybrid nature of this design, they extended the usual definition of power, as follows. Assume without loss of generality that $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$. Suppose that (i) at least one $\theta_j \geq \theta_0 + \delta_2$ and (ii) no θ_j falls between $\theta_0 + \delta_1$ and $\theta_0 + \delta_2$, since it is not possible to reliably distinguish statistically between arbitrarily close parameter values. Let $\beta(\boldsymbol{\theta})$ denote

the probability of not concluding $\theta_j > \theta_0$ for any j, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$. It can be proved that, under (i) and (ii), the false negative probability $\beta(\boldsymbol{\theta})$ is maximized for $\theta_1 = \cdots = \theta_{k-1} = \theta_0 + \delta_1$ and $\theta_k = \theta_0 + \delta_2$, the *least favorable configuration* (LFC), denoted by $\boldsymbol{\theta}^*$. The *generalized power* is defined as $1 - \beta(\boldsymbol{\theta}^*)$. Given this structure, they chose the four design parameters n_1, n_2, y_1, y_2 to minimize the expected sample size $E(N) = (k+1)n_1 + 2n_2 \Pr(Z_1 > y_1)$, with this expectation the equally weighted average between H_0 and the LFC. Note that this design generalizes the simpler two-treatment design of Thall et al. (1988a) described earlier by accommodating k experimental treatments, rather than one. Depending on θ_0, δ_1, and δ_2, for $k = 2, 3$, or 4 the optimal design has N_{max} varying from 178 to 458, and $E(N)$ between 140 and 415. For example, for $\theta_0 = .40$, $\delta_1 = .05$, and $\delta_2 = .20$ and $k = 3$, and generalized power .75, $4 \times 47 = 188$ patients are required for stage 1, and $2 \times 63 = 126$ for stage 2, so that $N_{max} = 314$, and $E(N) = 281$.

Thall et al. (1989) derived another version of this design that does not include an S arm in stage 1, instead using a fixed stage 1 cut-off λ with stage 2 conducted only if $\hat{\theta}_{[k]} > \lambda$. While this design produces an apparent saving in sample size compared to the design that does include S in stage 1, because stage 1 has no control arm it suffers from the fact that stage 1 does not provide an unbiased comparison of the E_j's to S based on actual data, hence there is a built-in trial effect. Moreover, no stage 1 data are incorporated into the stage 2 test statistic.

Schaid et al. (1990) also proposed a 2-stage phase 2–3 design to screen k experimental treatments $E_1, \ldots, E_k$ in phase 2, followed by randomized comparison to S in phase 3. Their design is similar to that of Thall et al. (1988b), with the two important differences that TTE outcomes are assumed and the design allows more than one of the E_j's to be moved forward to phase 3. Under a proportional hazards assumption, let θ_j denote the hazard ratio between E_j and S and $T^j(u)$ the log rank statistic comparing E_j to S at study time u. In stage 1, $(k+1)n_1$ patients are randomized fairly among $E_1, \ldots, E_k$ and S. For decision cut-offs $C_1 < C_2$, (i) if $\max\{T^1(t_1), \ldots, T^k(t_1)\} < C_1$ then the trial is terminated with all k experimental treatments declared not promising; (ii) if any $T^j(t_1) > C_2$ then the trial is terminated with the conclusion that all such treatments provide an overwhelming survival advantage over S. Otherwise, all $k_2 \leq k$ treatments for which $C_1 \leq T^j(t_1) \leq C_2$ are moved forward to stage 2, with fair randomization of $(k_2 + 1)n_2$ patients among the selected treatments and S. Accrual in stage 2 is terminated at t_a and a final analysis is performed at t_2, where $t_1 < t_a < t_2$. In the final comparisons, it is concluded that E_j is superior to S if $T^j(t_2) > C_3$. Denote the pairwise comparison false positive probability by α and pairwise power by $1 - \beta$, the null hazard with S by λ_0, and the common hazard ratio under the global null by θ. The design parameters are the test cut-offs C_1, C_2, C_3, and the per-arm sample sizes n_1, n_2. Schaid et al. derived the design parameters to minimize the null expected total sample size for given $\alpha, 1 - \beta, \lambda, \theta$, follow-up duration $t_2 - t_a$, and accrual rate, assuming uniform accrual during $[0, t_a]$ and exponential event time distributions, although they indicated how these assumptions may be relaxed. Given the goal of allowing all treatments that are promising to move forward for phase 3 evaluation, this design is highly efficient. For example, for hazard ratio of 1.5, $k = 2$ to 4, approximate overall type I error $k\alpha = .05$ and pairwise power $1 - \beta = .80$, and accrual rate/hazard rate $= 50$, the null expected sample sizes range from 233

to 448; increasing the hazard ratio to 2.0 gives expected sample size range 109–188. While allowing more than one experimental treatment to be moved forward to stage 2 requires a larger sample size than restricting stage 2 to allow only one E_j, an important advantage of this more flexible approach with TTE outcomes is due to the fact that survival differences may not be seen until the second stage. Consequently, this provides protection against false negatives in the initial screening. This design has the general phase 2–3 advantages that stage 1 data are utilized in the stage 2 tests, it avoids the bias of uncontrolled pre-test treatment selection, and it controls the overall error rates.

Schaid et al. (1988) proposed a phase 2–3 design for an oncology setting where k new agents $A_1, \ldots, A_k$ are available, along with an experimental treatment E that previously has been determined to be promising in phase 2 and a standard treatment S. Their design is aimed at newly diagnosed patients who have not yet been treated. At study entry, for frontline therapy patients are randomized fairly among all $k + 2$ treatments. For each patient, at disease progression a patient who received E or S initially is taken off study, whereas a patient who received one of the k new agents as frontline is then crossed over by being re-randomized between E and S as salvage therapy. This design has the advantage that an E-versus-S comparison may be made using both frontline and salvage data. To do this, Schaid et al. first computed the expected number of deaths for patients randomized to E or S as frontline therapy. The probability that a previously untreated patient randomized to $t = E$ or S will die by final follow-up time τ is computed as $p_{1t} = \int_0^a F_t(\tau - u) dG(u)$ where G is the cdf of the accrual time during the interval $[0, a]$ and F_t is the cdf of the survival time for a patient initially randomized to t. If Q is the proportion of all patients who are randomized in phase 3, and P is the proportion of patients in phase 3 who are randomized to E, then E(# deaths in the E arm in phase 3) $= NQPp_{1E}$ and the corresponding value for the S arm is $NQ(1 - P)p_{1S}$. Similar, more complex expressions are given for patients who receive one of $A_1, \ldots, A_k$ as frontline and are then re-randomized to E or S, accounting for the time to progression with the agent A_j given initially. These expressions are then used to compute a stratified log rank test statistic that accounts for the differential frontline treatment effects in a final comparison of survival with E and S. Since this design accounted formally for the two-stage nature of frontline and salvage therapy in cancer, it was a pioneering contribution in the field now known as "dynamic treatment regimes" (Thall et al. 2000; Murphy 2003, 2005; Lavori and Dawson 2004; Thall et al. 2007). While space does not permit a review of this rapidly growing literature, it is closely connected to phase 2–3 designs. A dynamic treatment regime is a set of rules for repeatedly treating and evaluating a patient in a multi-stage fashion, essentially a formalization of what physicians do routinely. If a design comparing dynamic treatment regimes has among its successive outcomes an early "phase 2" outcome Y and a subsequent "phase 3" outcome T, if patients are randomized, and if the design aims to control the overall false positive rate, then it may be regarded as a phase 2–3 design. The following section reviews designs for which the primary outcome is the pair (Y, T), rather than a single variable.

6 Accounting for both early and late outcomes

The relationship between Y and T reminds me of the wedding at which one of the guests discretely asked a member of the bride's family, "Aren't the bride and groom first cousins?" The family member, with equal discretion, replied, "Oh, yes, everybody knows that. We just don't talk about it."

In this section, I will openly discuss the intimate relationship between Y and T, despite the fact that it is nearly ignored in the clinical trial design literature. The rationale for using a phase 2 trial based on an early response indicator Y to decide whether to proceed with a phase 3 trial based on T is that the occurrence of response is likely to increase the value of T, that is, T increases stochastically with Y. This is defined by the inequality $Pr(T > u \mid Y = 1) = \bar{F}_1(u) > \bar{F}_0(u) = Pr(T > u \mid Y = 0)$ for all $u > 0$, where F_y is the cdf of T given $Y = y$ and $\bar{F}_y = 1 - F_y$. In words, responders live longer than non-responders, on average. If this is the case, then it makes perfect sense to use Y to decide whether E is promising. What does not make sense is to ignore the fact that the unconditional distribution of T is the mixture $f(u) = f_1(u)\pi + f_0(u)(1 - \pi)$, where $\pi = Pr(Y = 1)$ and $f_y = F'_y$ is the pdf of $[T \mid Y = y]$. The importance of this simple expression may be seen once treatment effects are considered. For example, suppose that the probability of surviving 12-months is .40 for non-responders and .60 for responders, a 50% improvement, and that $\pi_S = .20$ and $\pi_E = .40$, that is, E doubles the response probability of S. Then the 12-month survival probability with S is

$$\bar{F}_0(12)(1 - \pi_S) + \bar{F}_1(12)\pi_S = .40 \times (1 - .20) + .60 \times .20 = .44,$$

and the 12-month survival probability with E is

$$\bar{F}_0(12)(1 - \pi_E) + \bar{F}_1(12)\pi_E = .40 \times (1 - .40) + .60 \times .40 = .48.$$

That is, *doubling* the early response rate only increases the 12-month survival from .44 to .48, a 9% improvement. A conventional 2-sided group sequential design with one interim test using O'Brien-Fleming bounds, size .05 and power .80 to detect this alternative would require up to 2,850 patients. If one repeats these computations using other numerical values, it quickly becomes apparent that, in order to have any substantive impact on T, an enormous improvement in response rate is required, or response must provide an even larger increase in T, that is, $\bar{F}_1(u)/\bar{F}_0(u)$ must be much larger than 1.5. Even if response doubles 12-month survival, from .40 to .80, and E still doubles response rate, repeating the above computation gives 12-month survival probabilities .48 with S and .56 with E, still only a 17% improvement. In this very optimistic scenario, the maximum sample size required in phase 3 would still be over 900. Given these simple computations, and the many scientific problems with phase 2 trials noted earlier, it is hardly surprising that most phase 3 trials yield negative results. This was shown empirically by Tori et al. (1992), who performed a meta-analysis of 13 randomized ovarian cancer trials in which they assumed T to be a linear function of treatment effect and $\log\{\hat{\pi}_E/(1 - \hat{\pi}_E)\}$ for study-specific estimator $\hat{\pi}_E$ under a random effects model. Based on their analysis, they concluded that a new

treatment must produce a very large improvement in response rate in order to achieve a meaningful improvement in median (T).

Still, actual applications are generally much more complex. Additional elements that may come into play include direct effects of treatment on T not mediated through Y, and effects of patient covariates on both Y and T. The following more general mixture model accounts for these possibilities, and may be used as the basis for a phase 2–3 design. Since a binary response indicator may be inadequate, let Y be a discrete variable taking on c possible values, indexed by $y = 1, \ldots, c$. This accommodates important extensions such as that where Y has possible values (response, no toxicity), (response, toxicity), (no response, no toxicity) and (no response, toxicity). Let $\mathbf{Z} = (Z_1, \ldots, Z_q)$ denote a vector of baseline patient covariates, and let τ denote treatment, which may be $\{E, S\}$ or more generally $\{E_1, \ldots, E_k, S\}$. Denote the early outcome probability by $\pi_y(\tau, \mathbf{Z}) = Pr(Y = y \mid \tau, \mathbf{Z})$. The particular form of $\pi_y(\tau, \mathbf{Z})$ could be, for example, a generalized logistic or a bivariate binary regression model, provided that it accounts for effects of both τ and $\mathbf{Z}$. The distribution of $[T \mid \tau, \mathbf{Z}]$ may be expressed as the mixture

$$f_T(t \mid \tau, \mathbf{Z}) = \sum_y f(t \mid Y = y, \tau, \mathbf{Z}) \, \pi_y(\tau, \mathbf{Z}).$$

This model accounts for direct effects $(\tau, \mathbf{Z}) \longrightarrow Y$ and $(\tau, \mathbf{Z}) \longrightarrow T$ of treatment and covariates on each outcome, as well as the "phase 2 $\longrightarrow$ phase 3" relationship $Y \longrightarrow T$ between the early and late outcomes.

In the numerical examples given earlier, the distribution of T only varied with Y. Extending that example to allow direct effects of treatment on the phase 3 outcome gives

$$\bar{F}(12 \mid \tau) = \bar{F}_0(12 \mid \tau)(1 - \pi_\tau) + \bar{F}_1(12 \mid \tau)\pi_\tau$$

for $\tau = E$ or S. For example, denoting $\mu_{\tau,y} = \mathrm{E}(T \mid \tau, Y = y)$, the overall mean of T under treatment τ is $\mathrm{E}(T \mid \tau) = \mu_{\tau,0}(1 - \pi_\tau) + \mu_{\tau,1}\pi_\tau$, which depends on three parameters, and thus the difference in overall means,

$$\Delta = \{\mu_{E,0}(1 - \pi_E) + \mu_{E,1}\pi_E\} - \{\mu_{S,0}(1 - \pi_S) + \mu_{S,1}\pi_S\},$$

is a function of six parameters. Based on this representation of Δ, a substantive overall treatment effect may be achieved if three effects are at work:

1) $\pi_E > \pi_S$
2) $\mu_{E,1} > \mu_{S,1}$ and $\mu_{E,0} > \mu_{S,0}$
3) $\mu_{E,1} > \mu_{E,0}$ and $\mu_{S,1} > \mu_{S,0}$.

That is, (1) E increases the response probability compared to S, (2) E increases survival compared to S among both responders and non-responders, and (3) response increases survival regardless of which treatment achieved it. This sort of argument may be extended to more general Y. For example, if toxicity reduces survival, a treatment that reduces the toxicity rate while maintaining the response rate may improve

survival. Similarly, covariate effects may play a very important role in the above sort of analysis, especially if there are treatment-covariate interactions in either π or μ.

Inoue et al. (2002) applied a mixture model to derive a Bayesian phase 2–3 design in which patients are randomized between E and S throughout, with comparative decisions based on posterior and predictive probabilities defined in terms of the difference Δ in mean survival time with E and S, computed under a mixture model, so that in particular Δ accounts for all of the effects described above. In their application, a non-small-cell lung cancer trial, Y indicates local control of the tumor evaluated by biopsy at five months, T is survival time, one covariate Z indicating stage III disease is used, and the phase 2–3 goal is improve survival by 25%, assuming direct treatment effects on both Y and T, as well as improvement in survival if local control is achieved. The underlying mixture model accounts for the complication that Y is not observed in patients who die prior to the five-month evaluation. During an interim period in phase 2, the design makes repeated decisions of whether to stop due to futility, continue phase 2, or proceed to phase 3. If phase 3 is begun, the trial is expanded by adding more clinical centers, so the design has the advantage that there is no delay between the two phases. Superiority-continuation rules are applied at four-month intervals throughout phase 3. This could be regarded as a phase 3 trial with intensive monitoring, although the use of a parameter Δ accounting for the combined average treatment effect under a mixture model is far from conventional. To control overall size at .05 and achieve power .80, the design requires up to 900 patients. Over a wide array of cases, compared to conventional methods, on average the Bayesian phase 2–3 design provides a smaller trial, although the intensive monitoring produces much greater variability in sample size.

7 Dose ranging trials

Berry et al. (2001) proposed a Bayesian phase 2–3 design for a dose-ranging trial of a neuroprotective agent for stroke. The agent was administered intravenously as soon as possible after the stroke, with the trial goal to determine the optimal dose in terms of the ED95 (smallest dose providing a 95% response rate). Each patient's dose was assigned adaptively, based on posterior quantities computed from data on patients treated previously in the trial. A normal dynamic linear dose-response model was assumed for each patient's response profile, which was observed longitudinally, with response defined as the change in the patient's stroke score from baseline. In stage 1, the primary focus was to determine whether there existed a dose sufficiently efficacious to warrant a second, confirmatory stage in which patients were randomized between a selected dose and placebo. The possible stage 1 decisions were to stop the trial due to futility, continue dose-finding, or shift to the confirmatory phase. The basis for this decision was the predictive probability, given the current data, that the trial would ultimately show a selected dose to be superior to placebo. A key aspect of the design was that the switch from stage 1 to stage 2 was done "seamlessly," without suspending patient accrual. The duration of stage 1 was allowed to vary depending on the accumulating data on the dose-response curve.

Liu and Pledger (2005) took a very different approach to a phase 2–3 dose-finding trial. They proposed a two-stage design for a placebo controlled trial by generalizing Thall et al. (1988b) and Schaid et al. (1990), with several doses of a new drug playing the roles of the different experimental treatments. Like Berry et al. (2001) and Inoue et al. (2002), their proposed design does not stop accrual between phase 2 and phase 3. The Liu and Pledger design allows the two stages to overlap, with stage 2 begun while the stage 1 patients are still being followed for evaluation. The stage 1 sample size is determined to control the false negative rate in that stage. At an interim analysis, a futility stopping rule is applied, safety is assessed, and the stage 2 sample size is chosen to minimize expected total sample size subject to overall type I error and power constraints. Under the assumption that the test statistics from each stage are normally distributed, the data from the two stages are pooled for a final test using a trend statistic computed under a dose-response model in which the probability of efficacy reaches a plateau as a function of dose.

8 Discussion

Since the number of new treatments becoming available for clinical evaluation is rapidly increasing in many disease areas, while resources for evaluating treatments are limited, the need for efficient trial designs is more pressing than ever. Since it is easy to show that a properly designed phase 2–3 trial utilizes resources much more efficiently than conventional alternatives, whenever feasible, phase 2–3 trials should be adopted as standard practice. While the designs reviewed here provide an array of different methods that have been proposed, there are many other designs that may be called "phase 2–3," or that are closely related. A somewhat broader review than that provided here is given by Rubinstein et al. (2005).

While I have described some statistical problems with conventional clinical trial methods, in my experience the difficulties go beyond purely statistical issues. With regard to what may be gained or lost in a clinical trial, the utilities of physicians, patients in the trial, future patients, pharmaceutical companies and regulatory agencies can be quite different. While, ideally, a clinical trial design should strike a balance among these competing utilities, in practice a much less demanding but more realistic set of goals must be considered.

The relationship between statisticians and physicians plays a critical role in the clinical trial design process. Patients typically choose physicians, not treatments. Once a patient has entrusted a physician with his or her life, the patient may rely on the physician's expertise and simply hope for the best outcome. In my experience, physician-investigators choose statisticians in precisely the same way, entrusting the design of their clinical trials or analysis of their data to a statistician with the hope that it will produce scientifically valid results.

In practice, each trial has its own particular structure and goals. Constructing a design to accommodate the trial at hand becomes both more important and more difficult with more complex trials. This is especially true for trials that have multiple outcomes, multiple disease subtypes, multiple stages of therapy, multiple stages of treatment evaluation, or multiple goals. Tailoring a design to fit particular trial is

time-consuming, often requiring development of new computer software, and the investigators must sell the design to colleagues in their own clinics, to physicians in other medical centers considering participating in a multi-institution study, and to regulatory agencies. Given these difficulties, it is not surprising that many investigators choose a conventional design that may be constructed quickly and easily using an existing computer program. The motivation for an investigator to choose a more complicated but more realistic trial design comes from the belief that there is something substantive to be gained by its use, and the investigator's trust in his/her chosen statistician's abilities. There is a lot to be gained by the use of carefully constructed phase 2–3 designs, but this will occur only to the extent that statisticians earn their physician-collaborators' trust.

References

Bechhofer RE, Santner TJ, Goldsman DM (1995) Design and analysis of experiments for statistical selection, screening and multiple comparisons. John Wiley and Sons, New York, NY

Berry DA, Mueller P, Grieve AP, Smith M, Parke T, Blazek R, Mitchard N, Krams M (2001) Adaptive Bayesian designs for dose-ranging drug trials. In: Gatsonis C, Kass RE, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M (eds) Case studies in Bayesian statistics, V 99–181. New York, NY, Springer-Verlag

Braun TM, Thall PF, Nguyen H, de Lima M (2007) Simultaneously optimizing dose and schedule of a new cytotoxic agent. Clin Trials, (in press)

Bryant J, Day R (1995) Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics 51:1372–1383

Chang MN, Therneau TM, Wieand HS, Cha SS (1987) Designs for group sequential phase II clinical trials. Biometrics 43:865–874

Ellenberg SS, Eisenberger MA (1985) An efficient design for phase III studies of combination chemotherapies. (with discussion). Cancer Treat Rep 69:1147–1154

Estey EH, Thall PF (2003) New designs for phase 2 clinical trials. Blood 102:442–448

Fleming TR (1982) One sample multiple testing procedure for phase II clinical trials. Biometrics 38: 143–151

Gehan EA (1961) The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. J Chronic Diseases 13:346–353

Inoue LYT, Thall PF, Berry DA (2002) Seamlessly expanding a randomized phase II trial to phase III. Biometrics 58:823–831

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman and Hall, New York, NY

Lavori PW, Dawson R (2004) Dynamic treatment regimes: practical design considerations. Clin Trials 1:9–20

Liu Q, Pledger GW (2005) Phase 2 and 3 combination designs to accelerate drug development. J Am Stat Assoc 100:493–502

Murphy SA (2003) Optimal dynamic treatment regimes (with discussion). J Roy Stat Soc Ser B 65:331–366

Murphy SA (2005) An experimental design for the development of adaptive treatment strategies. Stat Med 24:1455–1481

Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy P, Smith M (2005) Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 23:7199–7206

Schaid DJ, Ingle JN, Wieand S, Ahmann DL (1988) A design for phase II testing of anticancer agents within a phase III clinical trial. Controlled Clin Trials 9:107–118

Schaid DJ, Wieand HS, Therneau TM (1990) Optimal two-stage screening designs for survival comparisons. Biometrika 77:507–513

Simon R (1989) Optimal two-stage designs for phase II clinical trials. Controlled Clin Trials 10:1–10

Simon R, Wittes RE, Ellenberg SS (1985) Randomized phase II clinical trials. Cancer Treat Rep 69: 1375–1381

Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. John Wiley and Sons, New York, NY

Thall PF, Cook JD (2004) Dose-finding based on efficacy-toxicity trade-offs. Biometrics 60:684–693

Thall PF, Simon R (1990) Incorporating historical control data in planning phase II clinical trials. Stat Med 9:215–228

Thall PF, Simon R (1994) Practical Bayesian guidelines for phase IIB clinical trials. Biometrics 50:337–349

Thall PF, Simon R, Ellenberg SS, Shrager R (1988a) Optimal two-stage designs for clinical trials with binary response. Stat Med 71:571–579

Thall PF, Simon R, Ellenberg SS (1988b) Two-stage selection and testing designs for comparative clinical trials. Biometrika 75:303–310

Thall PF, Simon R, Ellenberg SS (1989) A two-stage design for choosing among several experimental treatments and a control in clinical trials. Biometrics 45:537–547

Thall PF, Simon R, Estey EH (1995) Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med 14:357–379

Thall PF, Millikan R, Sung H-G (2000) Evaluating multiple treatment courses in clinical trials. Stat Med 19:1011–1028

Thall PF, Wooten LH, Logothetis CJ, Millikan R, Tannir NM Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Stat Med. (in press)

Therneau TM, Wieand HS, Chang M (1990) Optimal designs for a grouped sequential binomial test. Biometrics 46:771–781

Torri V, Simon R, Russek-Cohen E, Midthune D, Friedman M (1992) Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. J Nat Cancer Institute 84(6):407–413

Whitehead J (1986) Sample sizes for phase II and phase III clinical trials: an integrated approach. Stat Med 5:459–464