Statistics and Its Interface Volume 5 (2012) 431–442

Adaptation in clinical development plans and adaptive clinical trial designs

Tze Leung Lai, Olivia Yueh-Wen Liao[*] and Ray Guangrui Zhu

At the planning stage of clinical trials or of an encompassing clinical development plan for drug development, there is usually inadequate information about essential parameters for designing the Phase I, II and III clinical trials or for optimizing the sequence of clinical trials in an overall plan. It is therefore inevitable that strong assumptions need to be made at the planning stage to come up with over-simplified plans and designs. In this paper we describe novel statistical methods that can adapt these "initializing" designs or development plans to the sequential information accumulated during the development process. We show that the adaptive version of the initializing design/plan performs similarly to, or even better than, the initializing counterpart if the underlying assumptions actually hold, but can perform much better if the initial assumptions differ substantially from reality. We also describe how to maintain the prescribed type I error probability in these adaptive designs, thereby removing a major barrier to their use for regulatory approval of a new treatment.

AMS 2000 subject classifications: Primary 62L05; secondary 62F03.

Keywords and phrases: Adaptive design, Clinical development plan, Decision tree, Multi-arm trials, Multiple endpoint, Seamless Phase II-III trials.

1. INTRODUCTION

In the development of a new drug, an important component of the effort and costs involves clinical trials to provide clinical data to support a beneficial claim of the drug, and in case this is not valid, to support the termination of its development. The clinical trials progress in steps and are labeled Phase I, II and III trials. A project team steers the operations in which intensity, cost, and duration increase with the phase; in particular, Phase III often involves over 3,000 professionals, several years to reach completion, and over $100 million in cost. In addition, there is a core team that makes decisions guided by a clinical development plan (CDP). The CDP maps out the clinical development pathway, beginning with first-in-man studies and ending with submission to the regulatory agency or termination of development. It defines the number and type of clinical studies and their objectives, determines the time sequence of the studies, some of which may be carried out in parallel, identifies major risk areas, and sets key decision points and go/no-go criteria. An important objective of the CDP is to build a clinical data package to support a beneficial claim (which we refer to as "success data") or to support termination of further development (referred to as "termination data"). These data should start to be collected from Phase I dose ranging/selection studies, and provide evidence in favor of stopping or continuing at various decision points in subsequent Phase II and III trials.

The creation of a CDP involves team effort, with representation from R&D scientists, biostatisticians, project managers and the marketing department. To create the plan, the team faces many choices such as deciding among different paths, different clinical trial designs and endpoints. There are also choices concerning indications, which are related to disease sub-types and stages and patient subpopulations. Although it may be easier to establish efficacy for some indications, commercial implications should also be considered since earlier approval for the "easier" indications may not be the most profitable. Julious and Swank [20] have noted that statistical methods for clinical trial design have focused primarily on "optimizing individual clinical trials" but are lacking "at a more global level in the optimization of clinical development plans", which consist of sequenced experiments, some of which may be run in parallel, and "go" or "no-go" decisions. They use decision trees to compute the net present value of a clinical development plan (CDP) and thereby provide a method to optimize competing CDPs; see also [9]. In practice, however, it is often difficult to specify in advance the cost of each clinical trial in the sequence and the prior probabilities of a go or no-go decision resulting from the trial. In Section 2, we use ideas from adaptive statistical methods to resolve this and other difficulties in developing CDPs. An important method which is of much current interest is adaptive design of randomized clinical trials. Although confirmatory Phase III trials are built on the information gained from previous phases, such information is often inadequate for designing a Phase III trial because the earlier-phase trials have relatively small sample sizes due to the overall cost and time constraints, besides operational constraints such as study centers and patient accrual. Adaptive designs that can use information acquired during the

[*] Corresponding author.

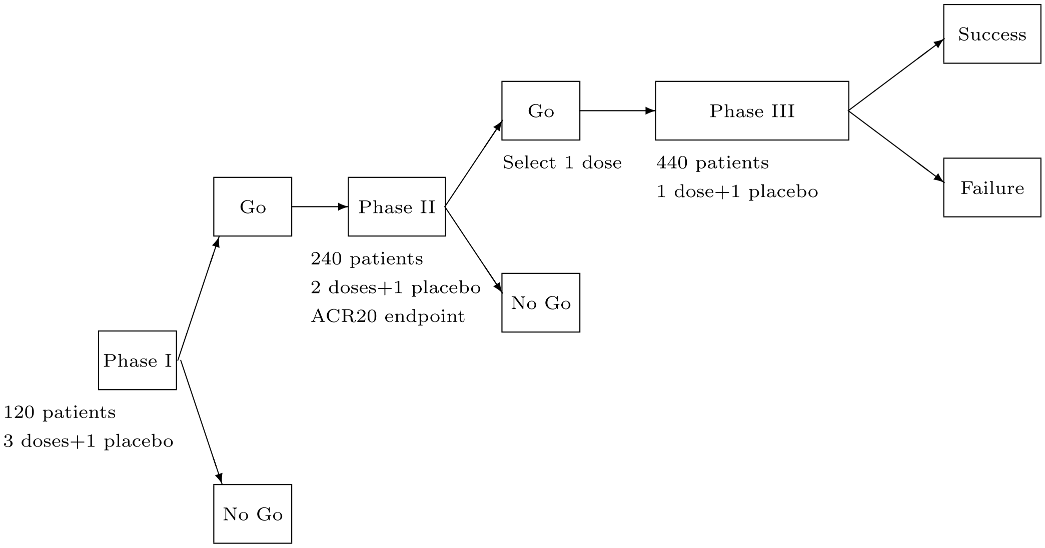

Figure 1. Decision tree of CDP for RA drug development.

course of the trial to update the design features have therefore attracted increasing interest from the pharmaceutical industry. Anticipating increasing use of these designs by the industry, the European Medicines Agency and the Food and Drug Administration in the United States have recently issued guidelines [10, 13] on adaptive trial designs. The main point of the guidelines is that the adaptive features offered by these innovative clinical designs should not inflate the probability of a false positive conclusion, i.e., the type I error probability of a confirmatory Phase III trial. Most of the methodological developments, therefore, have focused on the problem of controlling the type I error probability in adaptive designs. Section 3 gives a brief review of recent developments in seamless Phase II/III adaptive designs and considers one such design in particular, which it uses to illustrate the advantages of including adaptation in a CDP for a rheumatoid arthritis drug.

2. CONSIDERATIONS IN PLANNING CLINICAL DEVELOPMENT

2.1 Identifying the time and cost constraints

The time and cost constraints in clinical development of a new drug should be specified in its associated CDP. Decision trees of the type in [20] have been proposed for CDP teams to make decisions, as illustrated in Figure 1 on the development of a new drug to treat rheumatoid arthritis (RA). The Phase I study assesses the safety and tolerability of the drug in the patient population and compares several doses; it is often called a Phase IIA study. It uses a randomized, double-blind, placebo-controlled, parallel-arm design, with three active dose groups and one placebo group, and is expected to enroll 30 patients per arm. The active drug is an add-on therapy to the Standard of Care. The Phase II study in the plan is a parallel-arm design with two active dose groups and one placebo group, involving 80 patients per arm to achieve approximately 90% power to detect a 20% improvement in response rate over the placebo, which has a 50% response rate. Results from the Phase II study will be used to compare each dose's performance so that one dose can be selected for further testing in the Phase III study. The primary objective of the Phase III study is to demonstrate the efficacy and safety of the drug and file for its approval by the regulatory agency.

The decision tree in [20] also includes estimated costs of the clinical trials, together with prior probabilities of the outcomes of the Phase I, II and III trials. The expected cost of the trial can then be calculated and compared with those of other competing plans. However, despite the simple appearance of this decision tree in Figure 1, these cost estimates and prior probabilities are difficult to pin down. They can be based on previous experience and projections into the future, and on reported results scattered in medical literature and related studies, but there is usually a lack of reliable information at the time when a CDP is devised. No data from human subjects have been collected from the clinical trials yet to be performed, and there are many unknown factors that may affect outcomes and patient accrual. Moreover, during the course of development of the drug, unforeseen problems concerning patient subgroups, safety and efficacy may arise, and assumptions that yield *a priori* estimates may turn out to be overly simplistic. All these uncertainties can lead to inconclusive results at the end of the development process. On the other hand, much of the lacking information at the planning stage will be revealed in the development process. If the plan could be adapted to the accumulating information, the trial would end with conclusive results instead. Taking into consideration the time and cost constraints, an adaptive CDP can take advantage of the accumulating information to optimize the outcome of the overall process subject to these constraints.

2.2 Indications and levels of proof

There are four levels of proof of a drug's benefit: level 1 relates to the target exposure, level 2 to the target mechanism, level 3 to biomarkers that correlate with the clinical endpoints, and level 4 to the clinical or surrogate endpoints accepted by the regulatory agency. Biomarkers can be very helpful in clarifying the scientific understanding of the observed outcomes of the treatment in study subjects. They can be used to demonstrate mechanisms and to obtain early proofs of activity and safety, or lack thereof, and to control variability by allowing a wider range of hypothesis formulation that includes patient types and disease stages. Moreover, unfavorable pharmacokinetic (PK) and pharmacodynamic (PD) results provide evidence for early termination. The overall plan should be able to coordinate multiple studies at different stages after analysis of the data accumulated so far. Sequential learning from these studies can fine-tune dose selection and determine endpoints for the Phase III trial so that the new drug can at least be approved for some indication(s). In an example from our experience in the development of a neurologic drug, potential indications are migraine or neuropathic pain, multiple sclerosis, seizure and Parkinson's disease. Each indication is expensive to test through Phase II and III clinical trials. The mechanism was known for seizure (epilepsy), but there was little knowledge for the other indications. The sponsor tried Parkinson's disease, pain and seizure, but did not get significant results from the trials that were all on low doses. Results from the seizure patients showed no safety issues, suggesting that the dose could have been chosen to be higher. It was then decided to test seizure patients in two Phase III trials, using a 2–3 fold higher dose. Both showed significant results, with acceptable safety. Although much money had been wasted on negative trials with the three indications, the approval of the drug for seizure provided revenue for continuing the study for the other indications.

2.3 Regret and adaptation

Despite the sequential nature of Phase I, II and III trials in an overall development plan, the trials are often planned separately, each trial being treated as an independent study whose design depends on the results of previous studies. An advantage of this is that the reproducibility of the results of the trial can be evaluated on the basis of the prescribed design, without worrying about the statistical variability of the results of the earlier-phase trials that determine the prescribed design. However, an important disadvantage is that the sample sizes of the trials are often inadequate because of the separate planning; moreover, it is difficult to optimize the trial designs subject to the overall cost and time constraints of the drug's development. Experience has shown that many studies end in failure but their results are not negative. A way to address this issue is efficient integration of the sequenced trials in an overall development plan that expands a trial seamlessly from one phase to the next.

As an anecdotal illustration, a CDP was developed for a new drug to treat severe sepsis by suppressing the immune response to allow the body to prepare for the sepsis reaction. The CDP used the traditional Phase I, II and III framework for clinical trials. The Phase II trial was small and single-arm, giving a go-decision for Phase III which was a larger randomized trial involving about 2,000 subjects and took seven years. The protocol did not stipulate early stopping for futility; early stopping was stipulated by the Data and Safety Monitoring Committee only for major safety concerns. The final outcome was negative. After examining the data, this conclusion could have been reached after the first year since the final conclusion was consistent with the first 300 patients accrued. An important finding of the trial was that choosing a covariate-based treatment strategy (such as determining the time to initiate treatment based on the type of infections and other patient characteristics) could have shown the drug to be efficacious. However, this was post-hoc analysis and could only suggest future trials for an improved treatment strategy. After wasting seven years with no positive indication for the drug, the remaining patent life was too short to conduct such trials.

The *regret* of a CDP, which is the difference between the expected reward (market value of the approved drug minus the development cost) of the "oracle CDP", which assumes knowledge of the optimal treatment strategy, and that of the actual CDP can be greatly reduced if adaptive designs were used for the sequenced trials in the CDP, as will be explained in Section 3. The concept of regret was introduced by Lai and Robbins [22] in the classical "multi-arm" bandit problem. Let Π_j, $j = 1, \ldots, k$ denote statistical populations specified, respectively, by the density function $f(x;\theta_j)$, where the θ_j are unknown parameters belonging to some set Θ. Suppose $\mu_\theta = E_\theta(X)$ is finite for all $\theta \in \Theta$. How should we sample $x_1, \ldots, x_n$ sequentially from the k populations in order to maximize the expected value of $S_n = X_1 + \cdots + X_n$? This is called the "multi-arm bandit problem" and the name derives an imagined slot machine with $k \geq 2$ arms. When an arm is pulled, the player wins a random reward. For each arm j, there is an unknown probability distribution Π_j of the reward. The problem is to choose n pulls on the k arms so as to maximize the total expected reward. There is an apparent dilemma between the need to learn from all populations about their parameter values and the objective of the sampling only from the best population. The oracle policy assumes $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ to be known samples from the population with the largest μ_θ and has reward $n\mu(\theta^*)$, where $\theta^* = \text{argmax}_{1 \leq i \leq k} \mu(\theta_i)$. The *regret* of a given policy is defined in [22] by

$$\begin{aligned} R_n(\boldsymbol{\theta}) &= n\mu(\theta^*) - E_{\boldsymbol{\theta}} S_n \\ &= \sum_{j:\mu(\theta_j) < \mu(\theta^*)} \left(\mu(\theta^*) - \mu(\theta_j)\right) E_{\boldsymbol{\theta}} T_n(j), \end{aligned}$$

where $T_n(j)$ is the number of observations that the given policy samples from Π_j up to stage n. Whereas the traditional equal randomization scheme that randomly assigns each pull to the k arms with equal probability $1/k$ has a regret of order n, [22] have shown that an adaptive randomization scheme can attain the asymptotically minimal order of $\log n$ for its regret:

$$R_n(\boldsymbol{\theta}) = \sum_{j:\mu(\theta_j)<\mu(\theta^*)} \left(\frac{\mu(\theta^*) - \mu(\theta_j)}{I(\theta_j, \theta^*)} + o(1) \right) \log n,$$

where $I(\theta, \lambda) = E_\theta \log\left(f(X;\theta)/f(X;\lambda)\right)$ is the Kullback-Leibler information number.

Regarding a CDP as a policy consisting of a sequence of decisions involving sequenced clinical trials that generate data concerning a new drug to support its approval by the regulatory agency or to terminate further development, an optimal adaptive policy in this problem is clearly much more complicated than that in the multi-armed bandit problem. Nevertheless, similar principles still apply and the advantages of adaptation are even more pronounced. Incorporating adaptation in a CDP can self-tune the plan to increasing information about unknown parameters for optimal choice of decisions at various phases of its execution. In Section 3 we describe an adaptive alternative to the CDP in Figure 1 that integrates the Phase II trial involving dose selection into the Phase III trial on the selected dose. By appropriately using the Phase II (internal pilot) data in the final analysis, this design reduces the sample size and saves the study time since it avoids the long time lag between Phase II and Phase III trials. Because of the lack of information on the magnitude and sampling variability of the treatment effect at the design stage, there has also been increasing interest in adaptive designs that can adapt to information acquired during the course of the trial. Incorporating adaptation in a CDP can likewise self-tune the plan to increasing information at various stages of the development process.

3. ADAPTIVE DESIGN METHODOLOGY

As drug development has become increasingly costly and challenging, there has been increasing interest from the biopharmaceutical industry in adaptive designs that can self-tune a clinical trial to the information acquired during the course of the trial. In the FDA document [12] on the critical path to new medical products, it is recognized that today's advances in biomedical science offer new possibilities to treat many diseases but that the number of new drug and biologic applications submitted to the FDA has been declining. Innovative clinical trial designs and improvements in efficiency and cost effectiveness are needed for the biomedical advances to reach their full potential in treating diseases. In this section we first review some recent developments in adaptive designs towards this goal and then introduce a seamless Phase II/III design that leads to an adaptive modification of the CDP in Figure 1.

3.1 Efficiency under type I error and maximum sample size constraints

Most of the literature on adaptive design methodology focuses on the prototypical problem of testing a normal mean or difference between two normal means, beginning with Bauer [4], who introduced adaptive strategies for multiple testing, and Wittes and Brittain [36] who considered internal pilot studies. Depending on the topics covered, the term "adaptive design" in the literature is sometimes replaced by "sample size re-estimation", "trial extension" or "internal pilot studies." Much of the literature is about finding ways to adjust the test statistics after mid-course sample size modification so that the type I error probability is maintained at the prescribed level. In standard clinical trial designs, the sample size is determined by the power at a given alternative, but in practice, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. Although a standard method to address this difficulty is to carry out a preliminary pilot study, the results from a small pilot study may be difficult to interpret and apply, as pointed out by Wittes and Brittain [36] who proposed to treat the first stage of a two-stage clinical trial as an internal pilot from which the overall sample size can be re-estimated.

The basic idea of these two-stage designs dates back to Stein [33] who introduced a two-stage test of the hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative $\mu_X \neq \mu_Y$ for the mean of two independent normal populations with common, unknown variance σ^2, and based on i.i.d. observations $X_1, X_2, \ldots \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \ldots \sim N(\mu_Y, \sigma^2)$. In its first stage, Stein's test samples n_0 observations from each of the two normal populations and computes the usual unbiased estimate s_0^2 of σ^2. In the second stage, the test samples up to

$$n_1 = n_0 \vee \left[(t_{2n_0-2,\alpha/2} + t_{2n_0-2,\beta})^2 \, \frac{2s_0^2}{\delta^2} \right] \tag{1}$$

observations from each population, where α is the prescribed type I error probability, $1-\beta$ is the prescribed power at the alternatives satisfying $|\mu_X - \mu_Y| = \delta$, and $t_{\nu,a}$ denotes the upper α-quantile of the t-distribution with ν degrees of freedom. The null hypothesis $H_0 : \mu_X = \mu_Y$ is rejected if

$$\frac{|\bar{X}_{n_1} - \bar{Y}_{n_1}|}{\sqrt{2s_0^2/n_1}} > t_{2n_0-2,\alpha/2} \tag{2}$$

Stein [33] showed that the use of the initial variance estimate s_0^2 in the final test statistics (2) ensures that the test has type I error probability α and power at least $1-\beta$. However, this feature also diminishes the practical appeal of the test. Instead of s_0^2, it is more efficient to estimate σ^2 by the variance s_1^2 based on $2n_1$ observations. Wittes and Brittain [36] and subsequent authors develop critical values of the modified test statistics so that the type I error probability is approximately satisfied, as reviewed by Shih [31] and Whitehead et al. [37].

Without specifying δ, Fisher [14] proposed an adaptive design that uses the first-stage data to estimate $\theta = \mu_X - \mu_Y$ when the variance σ^2 is assumed known. Without loss of generality, let $\sigma^2 = 1/2$. Suppose the trial is designed to have a fixed sample size, with n observations per group. After rn pairs of observations $(0 < r < 1)$, letting $S_1 = \sum_{i=1}^{rn}(X_i - Y_i)$, we have $n^{-1/2}S_1 \sim N(r\theta\sqrt{n}, r)$. If it is now desired to change the second-stage sample size from $(1-r)n$ to $\gamma(1-r)n$ for some $\gamma > 0$, then letting $S_2 = \sum_{i=rn+1}^{n*}(X_i - Y_i)$, where $n* = rn + \gamma(1-r)n$ is the new total per treatment sample size, we have conditional on the first-stage data

$$(n\gamma)^{-1/2}S_2 \sim N((1-r)\theta\sqrt{\gamma n}, 1-r). \tag{3}$$

Note that under $H_0 : \theta = 0$, (3) has the $N(0, 1-r)$ distribution regardless of the choice of γ, showing that the test statistic $n^{-1/2}\left(S_1 + \gamma^{-1/2}S_2\right)$ has a $N(0,1)$ distribution under H_0. The corresponding test has been called the *variance spending test* because the variance $1 - r$ of (3) is the remaining part of the total variance 1 not spent in the first stage. Shen and Fisher [30] gave a multistage version of this procedure based on $S_1, S_2, \ldots, S_k$ in which the sample size updated at each stage may be data-dependent. Working in terms of the z-statistics that divides S by its standard deviation, Proschan and Hunsberger [27] noted that any non-decreasing function $C(z_1)$ with range $[0,1]$ can be used as a conditional type I error function to define a two-stage procedure, as long as it satisfies

$$\int_{-\infty}^{\infty} C(z_1)\phi(z_1)dz_1 = \alpha, \tag{4}$$

and suggested certain choices of C. Having observed the first-stage data Z_1, $H_0 : \theta = 0$ is rejected in favor of $\theta > 0$ after stage two if $Z_2 > \Phi^{-1}(1 - C(z_1))$. The condition (4) ensures that the type I error probability of any test of this form is α. The test proposed earlier by Bauer and Kőhne [6] can be represented in this common framework, as noted by Posch and Bauer [26].

Cui, Hung and Wang [11] discussed the issue of increasing the maximum sample size after interim analyses in a group sequential trial. They cited a study protocol, which was reviewed by the Food and Drug Administration, involving a Phase III group sequential trial for evaluating the efficacy of a new drug to prevent myocardial infarction in patients undergoing coronary artery bypass graft surgery. During interim analyses, the observed incidence for the drug achieved a reduction that was only half of the target reduction assumed in the calculation of the maximum sample size M, resulting in a proposal to increase the maximum sample size to $\tilde{M}$ ($N_{\max}$ in their notation). Cui, Hung and Wang [11] and Lehmacher and Wassmer [24] extended the sample size re-estimation approach to adaptive group sequential trials by adjusting the test statistics as in [27] and allowing the future group sizes to be increased or decreased during interim analyses so that the overall sample size does not exceed $\tilde{M}(>M)$ and the type I error probability is maintained at the prescribed level.

Jennison and Turnbull [19] found from simulation studies that the two-stage tests of [11], [14] and [30] perform poorly in terms of efficiency and power in comparison to group-sequential tests. Tsiatis and Mehta [34] independently came to the same conclusion, attributing this inefficiency to the use of non-sufficient weighted statistics $S_1 + r^{-1/2}S_2$ to combine data from the two stages. Bartroff and Lai [2, 3] used efficient generalized likelihood ratio (GLR) statistics to address this issue in testing the composite null hypothesis $H_0 : u(\theta) \le u_0$ in a multiparameter exponential family $f_\theta(x) = e^{\theta^T x - \psi(\theta)}$, where u is a smooth real-valued function such that the Kullback-Leibler information number

$$I(\theta, \lambda) = (\theta - \lambda)\dot{\psi}(\theta) - \{\psi(\theta) - \psi(\lambda)\} \tag{5}$$

is increasing in $|u(\lambda) - u(\theta)|$ for every fixed θ. Here and in the sequel, we use $\dot{\psi}$ to denote the gradient vector of the partial derivative of ψ with respect to the components of θ and $\ddot{\psi}$ to denote the Hessian matrix of second partial derivatives. The GLR statistic $\Lambda_{i,j}$ for sample size n_i at stage i has the form

$$\Lambda_{i,j} = \inf_{\theta:u(\theta)=u_j} n_i I(\hat{\theta}_{n_i}, \theta). \tag{6}$$

It estimates the unknown parameters in the likelihood ratio statistics by maximum likelihood, or constrained maximum likelihood estimates under the null hypothesis. To adjust for the uncertainties in these estimates, [2] and [3] used a 3-stage test that stops and rejects H_0 at stage $i \le 2$ if

$$n_i < M, \quad u(\hat{\theta}_{n_i}) > u_0 \quad \text{and} \quad \Lambda_{i,0} \ge b. \tag{7}$$

Early stopping for futility (accepting H_0) can also occur at stage $i \le 2$ if

$$n_i < M, \quad u(\hat{\theta}_{n_i}) < u_1 \quad \text{and} \quad \Lambda_{i,1} \ge \tilde{b}. \tag{8}$$

The test rejects H_0 at stage $i = 2$ or 3 if

$$n_i = M, \quad u(\hat{\theta}_M) > u_0 \quad \text{and} \quad \Lambda_{i,0} \ge c, \tag{9}$$

accepting H_0 otherwise. The sample size n_2 of the three-stage test is given by

$$n_2 = m \vee \{M \wedge \lceil (1 + \rho_m) n(\hat{\theta}_m) \rceil \}, \tag{10}$$

with

$$n(\theta) = \min\left\{ \frac{|\log \alpha|}{\inf_{\lambda:u(\lambda)=u_0} I(\theta,\lambda)}, \frac{|\log \tilde{\alpha}|}{\inf_{\lambda:u(\lambda)=u_1} I(\theta,\lambda)} \right\},$$

where $\rho_m > 0$ is an inflation factor to adjust for uncertainty in $\hat{\theta}_m$. As pointed out in [3], the right-hand side of (10) with $n(\hat{\theta}_n)$ replaced by $n(\theta)$ is an approximation to Hoeffding's lower bound [15] for the expected sample size, at the parameter value θ, of any test that has type I error probability $\le \alpha$ and type II error probability $\le \tilde{\alpha}$ at alternatives λ

satisfying $u(\lambda) = u_1$. Details on the choice of $b, \tilde{b}$ and c are given in [2, 3], and [3] also developed adaptive designs for mid-course modification of the maximum sample size in a group sequential trial, using efficient GLR statistics instead of the weighted statistics of [14] and [11].

3.2 Seamless Phase II-III designs involving multiple endpoints

The usefulness of seamless Phase II-III designs for clinical trials with bivariate short-term response and long-term survival endpoints is widely recognized, but how to design such trials has been a long-standing problem. Although randomized Phase II trials are common in other therapeutic areas, in oncology the majority of Phase II studies leading to Phase III studies are single-arm, and the most commonly used Phase II designs are Simon's single-arm two-stage designs [29] for testing $H_0 : p \leq p_0$ versus $H_1 : p \geq p_1$. Whether the new treatment is declared promising in a single-arm Phase II trial, however, depends strongly on the prespecified p_0 and p_1. As noted by [38], uncertainty in the choice of p_0 and p_1 can increase the likelihood that (a) a treatment with no viable positive treatment effect proceeds to Phase III, for example, if an artificially smallp_0 is chosen to inflate the appearance of a positive treatment effect when one exists; or (b) a treatment with positive treatment effect is prematurely abandoned at Phase II, for example, if an optimistically large p_1 is chosen. To circumvent the problem of choosing p_0, Vickers et al. [38] and Rubinstein et al. [28] have advocated randomized Phase II designs. In particular, it is argued that randomized Phase II trials are needed before proceeding to Phase III trials. However, the major barriers to randomization in Phase II cancer trials are that randomized designs typically require a much larger sample size than single-arm designs and that there are multiple research protocols competing for a limited patient population. Being able to include the Phase II study as an internal pilot for the confirmatory Phase III trial may be the only feasible way for a randomized Phase II cancer trial of such a sample size and scope to be conducted. Although tumor response is an unequivocally important treatment outcome, the clinically definitive endpoint in Phase III cancer trials is usually time to event, such as time to death or time to progression. The go/no-go decision to Phase III is typically based on tumor response because the clinical time-to-failure endpoints in Phase III are often of long latency, such as time to bone metastasis in prostate cancer studies. These failure-time data, which are collected as censored data and analyzed as a secondary endpoint in Phase II trials, can be used for planning the subsequent Phase III trial. Furthermore, because of the long latency of the clinical failure-time endpoints, the patients treated in a randomized Phase II trial carry the most mature definitive outcomes if they are also followed in the Phase III trial.

Although seamless Phase II-III designs can overcome the major barriers to randomization in Phase II cancer trials and have additional advantage mentioned above, only Bayesian methodologies introduced by Inoue et al. [18] and Huang et al. [16] have been developed until the recent work of Lai et al. [21]. The Bayesian approach assumes a parametric model that relates survival to response. Let Z_i denote the treatment indicator (0 = control, 1 = experimental), T_i denote survival time, and Y_i denote the binary response for patient i. The Bayesian approach assumes that the responses Y_i are independent Bernoulli variables and the survival times T_i given Y_i follows an exponential distribution, denoted Exp(λ) in which $1/\lambda$ is the mean:

$$Y_i | Z_i = z \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\pi_z), \tag{11}$$

$$T_i | \{Y_i = y,\, Z_i = z\} \overset{\text{i.i.d}}{\sim} \text{Exp}(\lambda_{z,y}), \tag{12}$$

and also assumes independent prior gamma distributions for $\lambda_{z,0}$ and $\lambda_{z,1}$ $(z = 0, 1)$ and beta distributions for π_0 and π_1. Each interim analysis involves updating the posterior probability $\hat{p} = P(\mu_1 > \mu_0 |\text{ data})$ and checking whether $\hat{p}$ exceeds a prescribed upper bound p_U or falls below a prescribed lower bound p_L, which is less than p_U. If $\hat{p} > p_U$ (or $\hat{p} < p_L$), then the trial is terminated, rejecting (accepting) the null hypothesis that the experimental treatment is not better than the standard treatment; otherwise the study continues until the next interim analysis or until the scheduled end of the study. The posterior probabilities are computed by Markov chain Monte Carlo, and simulation studies of the frequentist operating characteristics under different scenarios are used to determine the maximum sample size, study duration and the thresholds p_L and p_U.

Note that (11) and (12) imply that the conditional distribution of T_i given Z_i is a mixture of exponentials:

$$T_i | Z_i = z \overset{\text{i.i.d}}{\sim} \pi_z \text{Exp}(\lambda_{z,1}) + (1 - \pi_z)\text{Exp}(\lambda_{z,0}), \tag{13}$$

for which the hazard ratio of the treatment group $(z = 1)$ to the control group $(z = 0)$ is not constant over time. The assumption of exponential survival times in (11) and (12) in the Bayesian approach seems overly restrictive and semiparametric methods such as Cox regression are often preferred because of relatively large sample sizes in Phase III studies. Lai et al.[21] recently replaced (12) and (13) by the Cox regression model

$$\lambda(t|Y, Z) = \lambda_0(t) \exp(\alpha Y + \beta Z + \gamma Y Z). \tag{14}$$

The exponential model (11) and (12) is a special case of (14), with $\lambda_0(\cdot)$ being the constant hazard rate of an exponential distribution. Let $\pi_0 = P(Y = 1|Z = 0)$, $\pi_1 = P(Y = 1|Z = 1)$, and let $a = e^{\alpha}$, $b = e^{\beta}$ and $c = e^{\gamma}$. A commonly adopted premise in the sequenced trials to develop and test targeted therapies of cancer is that the treatment's effectiveness on an early endpoint such as tumor response would translate into long-term clinical benefit associated with a survival endpoint such as progression-free or overall survival, and conversely, that failure to improve the early endpoint would translate into lack of definitive clinical benefit. This explains why the go/no-go decision to Phase III made in a conventional Phase II cancer trial is based on the response endpoint. Under this premise, [21] also shows that the complement of the set of parameter values defining an efficacious treatment corresponds to the null hypothesis $H_0 = H_0^R \cup H_0^S$, where $H_0^R : \pi_0 < \pi_1$ and

$$H_0^R : \pi_0 < \pi_1 \quad \text{and} \quad \pi_0 a + (1-\pi_0) \le \pi_1 abc + (1-\pi_1)b.$$

In view of this decomposition of the null hypothesis, [21] uses a group sequential design that focuses on testing H_0^R in the interim analyses and then switches to testing H_0^S after H_0^R is rejected. It uses Lai and Shih's modified Haybittle-Peto test [23] involving GLR statistics to test H_0^R and extends the methodology to maximum partial likelihood ratio statistics to test H_0^S after H_0^R is rejected.

3.3 A CDP using Phase II-III design

As a chronic and symptomatic disease, rheumatoid arthritis (RA) can lead to various outcomes with diverse severities and therapeutic effects. Typically, the first claim to be tested for labeling and regulatory approval of a RA treatment is a reduction in signs and symptoms, which can be evaluated by the American College of Rheumatology 20% criterion (ACR20). ACR20 is a binary endpoint that categorizes patients' response to a treatment within a six-month follow-up period as success or failure. Another potential claim of the clinical benefits of the treatment in terms of disease modification is the prevention of structural damage. The claim can be supported by testing a radiographic index, such as the Sharp score which measures inhibition of radiographic progression by comparing the measurements taken one (or two) year(s) after treatment with those taken prior to treatment. As mentioned in the FDA's guidelines for the pharmaceutical industry, the inhibition of radiographic progression according to the Sharp score is usually considered after the treatment benefit has been demonstrated by ACR20 to relieve signs and symptoms of RA.

The testing problem, therefore, is similar to that in Section 3.2. Both problems involve a binary endpoint and consider the other endpoint only after the treatment has been shown to have significant benefits for the binary endpoint. However, there are three major differences between the two problems. First, the survival outcome in Section 3.2 requires long-term follow-up and is the clinically definitive endpoint in the sense that the treatment cannot gain regulatory approval without demonstrating survival benefits. In contrast, the binary endpoint ACR20 is the clinically definitive endpoint for RA treatments and the 1-year Sharp score only needs 6 additional months of follow-up time for each subject and is fully observable, unlike survival times that may be censored. Secondly, while survival trials need a Data and Safety Monitoring Committee (DSMC) to monitor the trial periodically and a group sequential design is natural as interim analysis can be carried out at the monitoring meetings, the RA trials do not need monitoring by DSMC and are usually designed as fixed sample size trials rather than group sequential trials, as in Figure 1. Thirdly, unlike the Phase II-III trial in Section 3.2 that uses only one dose, which is the maximum tolerated dose determined at the end of the Phase I cancer trial, the Phase II RA trial in Figure 1 involves dose selection for Phase III. In fact, for RA treatments, dose selection is a difficult but important problem and needs comparison of large enough samples of ACR20 scores of patients treated at several doses.

A seamless Phase II-III design that aims at interweaving the Phase III trial with the Phase II trial, in which several doses of a new drug are compared to a control or placebo with the goal of deciding whether to stop or to continue development with a selected dose for Phase III, is an active area of current research in adaptive design of clinical trials. The Bayesian approach uses posterior probabilities to select arms in an adaptive multi-arm trial that starts with multiple treatment arms and makes a mid-course decision concerning which arm is appropriate to carry forward for confirmatory testing; see [7, Ch. 5]. Instead of using posterior probabilities, the frequentist approach selects mid-course the apparently better of the two leading arms to carry forward to the scheduled end of the trial, and either uses the asymptotic normality of a weighted statistic of the type (3), or the exact or approximate sampling distribution of a more efficient test statistic under the null hypothesis to control the type I error probability of falsely rejecting the null hypothesis that the new treatment with the selected dose is better than the control (or placebo); see [5, 8, 17, 32, 35].

In summary, an adaptive counterpart of the separate Phase II and Phase III trials, which involve only the ACR20 endpoint in Figure 1 for the development of the RA drug, is a seamless Phase II-III design involving ACR20 and Sharp score as a bivariate endpoint, similar to that in Section 3.2, and carrying out mid-course dose selection as in references cited in the preceding paragraph. However, unlike these references, we use efficient test statistics and stopping rules similar to those introduced by Bartroff and Lai [2, 3] described in Section 3.1. Figure 2 gives the decision tree of the adaptive modification of the CDP in Figure 1.

3.4 A Phase II-III design to test ACR20 and Sharp score improvement

The Phase II-III trial in Figure 2 begins with two doses for the new treatment. Since ACR20 is a binary variable taking the value of 1 or 0, the associated test statistic for the treatment with dose j, which we shall call treatment arm j, is the normalized difference

$$(15) \qquad \Delta_{i,j} = \frac{\hat{p}_{ij} - \hat{p}_{i0}}{\{\hat{p}_{ij}(1-\hat{p}_{ij})/n_{ij} + \hat{p}_{i0}(1-\hat{p}_{i0})/n_{i0}\}^{1/2}},$$

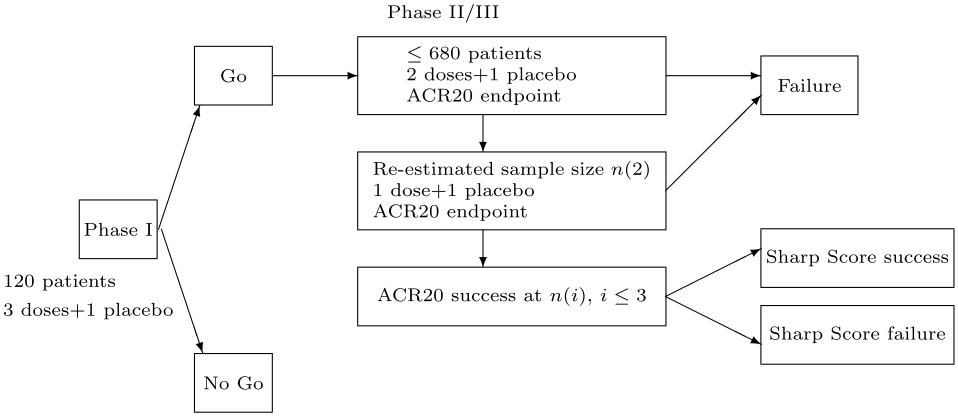

Figure 2. Decision tree of adaptive CDP.

where n_{ij} is the sample size of treatment arm j at the ith interim analysis and $\hat{p}_{ij}$ is the proportion of ACR20 responders to treatment arm j among the n_{ij} subjects, in which treatment arm 0 refers to the placebo. In practice, there is an upper bound M on the total sample size of the Phase II-III trial because of funding and time constraints and because there are other trials that compete for patients, investigators, and resources. Lai and Shih [23, p. 511] have pointed out that M implies constraints on the alternatives that can be used in power calculations to determine the sample size. Specifically, here we use the alternative δ "implied" by M, in the sense that M can be determined as the total sample size of the level-α test with power $1-\tilde{\alpha}$ at δ, using the normal approximation for (15) in testing the null hypothesis $H_0^R : \max(p_1 - p_0, p_2 - p_0) \le 0$; here the superscript R stands for ACR20. Alternatively, δ can be specified as a clinically relevant or anticipated effect size based on previous experimental or observational studies.

Although we use stopping rules similar to those in (7)–(9) to test H_0^R, we use in (7) and (9) the Wald statistics (15) instead of GLR statistics (6). The first stage of the Phase II-III design is the same as Phase II design in Figure 1, with a total of $n(1) = n_{10} + n_{11} + n_{12} = 240$ patients randomized to the three arms. The first interim analysis, performed at the end of this stage, chooses the apparently better dose $\mathbf{J} = \operatorname{argmax}_{j=1,2}\Delta_{1,j}$ to test the treatment in subsequent stages. Based on the effect size of dose $\mathbf{J}$, the analog of (10) is $n(2) = (n(1) - n_{1\mathbf{J}'}) \vee \{(M - n_{1\mathbf{J}'}) \wedge \lceil (1+\rho_m)\hat{n}\rceil\}$, where $\mathbf{J}' = \operatorname{argmin}_{j=1,2}\Delta_{1,j}$ is the dose eliminated from further study and

$$\hat{n} = 2\min\left\{\frac{|\log\alpha|}{(\hat{p}_{1\mathbf{J}} - \hat{p}_{10})^2/\hat{v}}, \frac{|\log\tilde{\alpha}|}{(\hat{p}_{1\mathbf{J}} - \hat{p}_{10} - \delta)^2/\hat{v}}\right\}, \tag{16}$$
$$\hat{v} = \hat{p}_{1\mathbf{J}}(1 - \hat{p}_{1\mathbf{J}}) + \hat{p}_{10}(1 - \hat{p}_{10}).$$

The $n(2) - n(1)$ subjects in the second stage are randomized to treatment (with dose $\mathbf{J}$) and placebo. This explains the factor 2 in the first line of (16).

At stage $i = 1$ or 2, the test rejects H_0^R if

$$n(i) < M - n_{1\mathbf{J}'} \quad \text{and} \quad \Delta_{i,\mathbf{J}} > b. \tag{17}$$

The Wald statistic corresponding to $\Lambda_{i,1}$ in (8) is

$$\tilde{\Delta}_{i,\mathbf{J}} = \frac{\hat{p}_{i\mathbf{J}} - \hat{p}_{i0} - \delta}{\{\hat{p}_{i\mathbf{J}}(1 - \hat{p}_{i\mathbf{J}})/n_{i\mathbf{J}} + \hat{p}_{i0}(1 - \hat{p}_{i0})/n_{i0}\}^{1/2}}. \tag{18}$$

The test accepts H_0^R and terminates the trial for futility at stage $i \le 2$ if

$$n(i) < M - n_{1\mathbf{J}'} \quad \text{and} \quad \tilde{\Delta}_{i,\mathbf{J}} < \tilde{b} < 0. \tag{19}$$

If neither (17) nor (19) occurs, the test does not stop early and rejects H_0^R if

$$\Delta_{3,\mathbf{J}} > c, \tag{20}$$

accepting H_0 otherwise. The Wald statistic $\Delta_{3,\mathbf{J}}$ in (20) is based on $M - n_{1\mathbf{J}'}$ observations and the thresholds $b, \tilde{b}$, and c are defined by the equations

$$\begin{aligned} &P_{\max(p_1,p_2)=p_0+\delta}\{(19) \text{ holds for } i = 1 \text{ or } 2\} = \tilde{\epsilon}\tilde{\alpha}, \\ &P_{p_1=p_2=p_0}\{(17) \text{ holds for } i = 1 \text{ or } 2\} = \epsilon\alpha, \\ &P_{p_1=p_2=p_0}\{(17) \text{ does not hold for } i \le 2, \text{ and } (20) \text{ holds}\} \\ &\quad = (1-\epsilon)\alpha, \end{aligned} \tag{21}$$

which can be computed by using the asymptotic joint normality of the Wald statistics and recursive numerical integration; see the next section for details.

As pointed out in Section 3.3, besides the six-month ACR20, the one-year Sharp scores are also measured from patients in the trial. If H_0^R is rejected, the CDP would proceed to check if the Sharp scores data support a beneficial claim of the treatment's efficacy in inhibiting structural damage. Note that in this connection smaller Sharp scores are associated with better outcomes. Because Sharp scores are markedly non-normal, the Wilcoxon test based on the sum S of the ranks of the Sharp scores from the treatment group in the combined sample (consisting of both the treatment and the placebo groups) is often used to test the null hypothesis of no improvement in Sharp scores for the treatment. The standardized Wilcoxon statistic at the ith interim analysis after selecting dose $\mathbf{J}$ for the treatment is

$$W_{i,\mathbf{J}} = (n_{i0}n_{i\mathbf{J}}/2 - S_{i,\mathbf{J}}) / \left\{ n_{i0}n_{i\mathbf{J}}(n_{i0}+n_{i\mathbf{J}})/12 \right\}^{1/2}, \tag{22}$$

which can be used to replace the GLR statistics in (7) and (9). Futility stopping is based on ACR20 and not on Sharp scores, so we do not have to consider the counterpart of (8) for Sharp scores. Hence, to test $H_0^S : \max(p_1, p_2) > p_0$ and $P(Y \le X_j) \ge 1/2$ for $j = 1, 2$, where Y is a randomly selected Sharp score from the placebo population and X_j is that from the treatment population (receiving dose j), let I be the stage at which the three-stage trial stops and let w_{obs} be the observed value of $W_{I,\mathbf{J}}$. After rejecting H_0^R, we reject H_0^S if $w_{obs} > 0$ and

$$P_{\hat{p}_0,\hat{p}_1,\hat{p}_2;H_{0+}^S}(W^*_{I^*,\mathbf{J}^*} > w_{obs}) \le \alpha. \tag{23}$$

The probability measure in (23) refers to that which substitutes the unknown parameters p_0, p_1, p_2 by their maximum likelihood estimators at stage I and assumes $H_{0+}^S : P(Y \le X_1) = \frac{1}{2} = P(Y \le X_2)$. W use asterisks for the random variables I^*, $\mathbf{J}^*$and $W^*_{I^*,\mathbf{J}^*}$ to denote that these random variables are generated from the aforementioned probability measure, independently of how the observed sample $(I, \mathbf{J}, W_{I,\mathbf{J}})$ was generated. Note that $(\hat{p}_0, \hat{p}_1, \hat{p}_2)$ satisfies $\max(\hat{p}_1, \hat{p}_2) > \hat{p}_0$, otherwise the trial would not have proceeded to test H_0^S whose $\max(p_1, p_2) > p_0$ requirement is already satisfied by $(\hat{p}_0, \hat{p}_1, \hat{p}_2)$. The probability in (23) can be evaluated by recursive numerical integration after using the normal approximation for the normalized statistics $\Delta_{i,j}$ and $W_{i,j}$; details are given in the next section.

3.5 Implementation and a simulation study

The probability in (23) can be decomposed into

$$P_{\hat{p}_0,\hat{p}_1,\hat{p}_2;H_{0+}^S}(W^*_{I^*,\mathbf{J}^*} > w_{obs}) = \sum_{i=1}^{3}\sum_{j=1}^{2} P_{\hat{p}_0,\hat{p}_1,\hat{p}_2;H_{0+}^S}(I^* = i,\, \mathbf{J}^* = j,\, W^*_{i,j} > w_{obs}),$$

in which each summand can be evaluated by a trivariate extension of the recursive numerical integration algorithm of Armitage, McPherson and Rowe [1] using the joint asymptotic normality of $\Delta_{i,j}$ and $W_{i,j}$. For example, for $i = j = 1$, the summand can be expressed as $P_{\hat{p}_0,\hat{p}_1,\hat{p}_2;H_{0+}^S}(\Delta^*_{1,1} > \Delta^*_{1,2},\, \Delta^*_{1,1} > b,\, W^*_{1,1} > w_{obs})$, in which $(\Delta^*_{1,1}, \Delta^*_{1,2}, W^*_{1,1})$ is asymptotically normal with covariance matrix $(v_{hk})_{1 \le h,k \le 3}$ such that $v_{kk} = 1$, $v_{12} = 1/2$, $v_{13} = r_1$ and $v_{23} = r_2$. For $j = 1, 2$, r_j can be estimated by the sample correlation coefficient of ACR20(j) - ACR20(0) and $I_{X_0 \le X_j}$, in which ACR20(j) and X_j are the ACR20 value and Sharp score, measured at stage i, of a subject receiving the treatment at dose j, and ACR20(0) and X_0 are those of a subject receiving the placebo. Note that the correlation coefficient is computed from $n_{Ik}n_{I0}$ pairs of subjects from the treatment (with dose j) and placebo groups.

We can use a similar decomposition to compute the probability in (21). For example, $P_{p_1=p_2=p_0}\{(17)$ holds for $i = 1$ or $2\}$ can be decomposed into

$$\sum_{j=1}^{2} \left[P_{p_1=p_2=p_0}\{\Delta_{1,j} \ge \Delta_{1,j'} \text{ and } \Delta_{1,j} > b\} + P_{p_1=p_2=p_0}\{b \ge \Delta_{1,j} \ge \Delta_{1,j'} \text{ and } \Delta_{2,j} > b\} \right], \tag{24}$$

in which j' denotes the other dose that is not j. The first term in the sum can be evaluated by using the asymptotic normality of $(\Delta_{1,1}, \Delta_{1,2})$. The second term in the sum can be evluated by a bivariate extension of the recursive numerical algorithm of [1]; see [3, p. 260]. Specifically, using the normal approximation, [3] shows how $p(x) = P_{p_1=p_2=p_0}\{\Delta_{2,j} > b | \Delta_{1,j} = x\}$ can be expressed in terms of a normal cumulative distribution function (cdf). Moreover, the density function of $\Delta_{1,j}$ is approximately that of a normal density ψ. Hence, the second term in (24) can be expressed as

$$\int_{-\infty}^{b} P_{p_1=p_2=p_0}\{\Delta_{1,j'} \le x | \Delta_{1,j} = x\} p(x)\psi(x)dx,$$

in which the conditional probability can be approximated by that of a normal cdf.

We use the preceding implementation for a simulation study comparing the clinical trial design in the CDP in Figure 1, which will be called Plan A, with the Phase II-III design in Section 3.4 for the CDP in Figure 2, which will be called Plan B. The simulation study assumes that the Sharp score is conditionally $N(\mu, \sigma^2)$ given ACR20 = 1 and $N(\tilde{\mu}, \tilde{\sigma}^2)$ given ACR20 = 0. It considers five scenarios, the first of which is labeled S1 and assumes that the placebo and the two treatment groups have the same $\mu, \sigma, \tilde{\mu}$ and $\tilde{\sigma}$ and also the same probability p of ACR20 taking the value 1. The other four scenarios, labeled S2, ..., S4, let $p, \mu, \sigma, \tilde{\mu}, \tilde{\sigma}$ vary over the placebo and treatment arms 0, 1, 2. These values are shown in Table 1, in which S2 uses the results in Table 4 of [25]. For each scenario, patients are assumed to arrive uniformly with a recruitment rate of 200 patients per year. For the Phase II-III design in Section 3.4, the significance level is set at $\alpha = 0.05$ and the power is $1 - \tilde{\alpha} = 0.8$. The maximum sample size constraint $M = 680$ leads to $n(1) = 240$, which is the same as that of the Phase II trial in Plan A, and $n(3) = 680$. We choose $\epsilon = 1/2$ and $\tilde{\epsilon} = 1/3$ in (21), following [2, 3]. Of particular interest for the CDP in Figure 2 is the probability of success of the trial, in which success means a positive claim for the ACR20 endpoint, with probability $P(R)$ shown in Table 2; a bigger success would be a

Table 1. Parameter settings for five scenarios

		p	μ	σ	$\tilde{\mu}$	$\tilde{\sigma}$
S1		0.5	0.7	5.5	2.6	10.7
S2	0	0.5	0.7	5.5	2.6	10.7
	1	0.5	1.5	7.2	1.1	4.7
	2	0.62	0.1	3.8	0.2	3.4
S3	0	0.45	0.2	4	0.75	4
	1	0.55	0.3	4	0.75	4
	2	0.55	0.3	4	0.75	4
S4	0	0.5	0.5	4	3.5	4
	1	0.5	−0.57	4	2.43	4
	2	0.5	−0.57	4	2.43	4
S5	0	0.45	0.5	2	3.5	2
	1	0.45	0.5	2	3.5	2
	2	0.57	−0.14	2	2.86	2

Table 2. Probabilities of positive claims, $E(N)$ and $E(T)$ for Plans A and B

		$P(R)$	$P(R\&S)$	$E(N)$	$E(T)$
S1	A	0.04	0.003	593	4.27
	B	0.05	0.003	522	3.26
S2	A	0.75	0.51	670	4.83
	B	0.80	0.50	542	3.74
S3	A	0.68	0.04	676	4.87
	B	0.83	0.04	516	3.62
S4	A	0.04	0.03	594	4.28
	B	0.05	0.04	529	3.31
S5	A	0.73	0.73	617	4.84
	B	0.82	0.68	538	3.73

positive claim for both ACR20 and Sharp score, with probability $P(R\&S)$ shown in the table. The efficiency of the trial is measured by savings in expected sample size $E(N)$ and the expected duration $E(T)$ over the separate fixed sample size Phase II and III trials in the CDP of Figure 1. Table 2 also gives $E(N)$ and $E(T)$ in the five scenarios. Each result is based on 5,000 simulation runs.

For comparison, Table 2 also considers the CDP in Figure 1, labeled Plan A. The total sample size is $M = 680$ if the Phase II trial ends with a "go" decision for Phase III; it is 240 if the Phase II trial ends with a "no-go" decision, which is guided by the same futility stopping criterion (19) used by Plan B. Plans A and B use the same maximum sample size for comparison purpose. Table 2 also gives the total expected sample size $E(N)$ for Plan A, taking into account of potential "no-go" decisions for Phase III. The analog of $E(T)$ is more subtle. Usually there is a long delay in starting a Phase III trial after analyzing the Phase II data, and the time delay adds to the total time needed to carry out the CDP from the Phase I trial to the new drug application. Since the seamless Phase II-III trial does not incur this time delay, we define T for Plan A as the duration from the beginning of the Phase II trial to the end of the Phase III trial, minus the time delay between the two trials, if the Phase II trial ends with a go decision; we define T as the duration of the Phase II trial if it ends with a no-go decision. Unlike the adaptive design in Figure 2, the Phase III trial in Figure 1 is intended to test the Sharp score as a secondary endpoint, similar to traditional designs. This means separate tests for the null hypotheses $p_{\mathbf{J}} \leq p_0$ and $P(X_0 \leq X_{\mathbf{J}}) \geq \frac{1}{2}$, where $\mathbf{J}$ is the dose selected at the end of the Phase II trial. No Bonferroni or other multiple testing adjustments are made to maintain the overall type I error probability. The probability $P(R\&S)$ in Table 2 for Plan A should be interpreted in this more liberal sense of separate tests for the primary and secondary endpoints, which the regulatory agency may not accept as sufficient evidence in favor of the claim that the treatment improves the Sharp score when the secondary analysis rejects the null hypothesis $P(X_0 \leq X_{\mathbf{J}}) \geq \frac{1}{2}$.

4. DISCUSSION

Table 2 shows the advantages of the seamless Phase II-III trial over conventional Phase II and Phase III trials. The scenarios S1 and S4 belong to H_0^R. While the two treatment doses in S1 have the same distribution of Sharp scores as that of the placebo population, they both have smaller (and therefore better) mean Sharp scores than the placebo in S4. The adaptive design (Plan B) maintains the type I error probability of 5% in both scenarios, and so does the traditional design (Plan A). For the other three scenarios, Plan B has a higher probability of reaching a positive claim and yet a smaller expected sample size than Plan A. Note that in S3, the Sharp scores of both treatment arms tend to be larger (worse) than those of the placebo, explaining why $P(R\&S)$ is small. On the other hand, in S5, treatment arm 2 has marked improvements in the mean Sharp score for both ACR20 responders and non-responders, and Plan B has an about 70% chance of positive claims for both ACR20 and Sharp score endpoints and an 80% chance of a positive claim for ACR20 alone. The probability of positive claims for both endpoints drops to 50% in S2 in which the treatment arm 2 has smaller improvements in the mean Sharp score over the placebo. In S2 and S5, both plans almost always choose dose 2, the actual better dose, when making a positive claim for ACR20. The reduction in the duration of the development process in all scenarios is an even greater advantage of Plan B over Plan A. The two plans have similar chances of stopping at $n(1)$ for futility (i.e., for the no-go decision) since they use the same stopping rule (19). However, Plan B has a markedly smaller expected sample size because it carries out efficient sample size re-estimation for the second stage. As pointed out in the last paragraph of Section 3.5, the $E(T)$ of Plan A shown in Table 2 does not take into consideration the time delay between Phase II and Phase III trials. For the RA drug, this was estimated to be at least one year. The time savings were most appealing to the CDP team.

In the literature reviewed in Section 1, most existing methods, such as [32] and [35], deal with only one endpoint for multiple arms. Our design deals with bivariate endpoints besides multiple arms. It should be noted that, unlike the decision trees in [20], no cost estimates are attached to Figures 1 and 2. In practice, for the team to arrive at the CDP choice and to present it to senior management, costs and prior probabilities of success together with estimates of the market value of the drug if approved have to be incorporated into the decision tree, as in [20]. At the planning stage, there is a lot of uncertainty in these estimates and in those for planning the trials. Standard clinical trials designs, such as those in Figure 1, which are well understood and relatively simple to present may be a good way to start. Because of the uncertainties and the strong assumptions underlying the "simplified" CDP, its adaptive version can be used for the development process which involves adaptive designs in place of the standard designs in the simplified CDP. Table 2 and other simulation studies not reported here show that the adaptive modification of the original CDP performs similarly to, or better than, the original CDP if the assumptions underlying the CDP actually hold, but can perform much better if they differ substantially from reality. A major barrier to the acceptance of adaptive designs by regulatory agencies is that adaption may substantially inflate the type I error probability. However, as we have shown in Section 3, important advances have been made in the past decade to remove this barrier, and the time is now ripe for using adaptive methods not only in clinical trials but also in clinical development plans for drug development.

ACKNOWLEDGMENTS

Lai's research was supported by the National Science Foundation grant DMS-085879 and the National Cancer Institute grant 1 P30 CA 124435-01. Liao's research was supported by the National Institutes of Health grant 4R37EB002784.

Received 26 September 2012

REFERENCES

[1] Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132** 235–244. MR0250405

[2] Bartroff, J. and Lai, T. L. (2008a). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. Med.* **10** 1593–1611. MR2420330

[3] Bartroff, J. and Lai, T. L. (2008b). Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Sequential Analysis* **27** 254–276. MR2446902

[4] Bauer, P. (1989). A sequential elimination procedure for choosing the best population(s) based on multiple testing. *Journal of Statistical Planning and Inference* **21** 245–252. MR0985461

[5] Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Stat. Med.* **18** 1833–1848.

[6] Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50** 1029–1041.

[7] Berry, S. M., Carlin, B. P., Lee, J. J. and Müller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Chapman & HaiVCRC Biostatistics Series, Vol 38. CRC Press, Boca Raton, FL. MR2723582

[8] Bretz, F., Schmidli, H., Konig, F., Racine, A. and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: General concepts. *Biometrical Journal* **4** 623–634. MR2247048

[9] Burman, C. F. and Senn, S. (2003). Examples of option values in drug development. *Phar. Stat.* **2** 113–125.

[10] Committee for Medicinal Products for Human Use (CHMP) (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency, available at http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf.

[11] Cui, L., Hung, H. M. J. and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Ann. Math. Statist.* **55** 853–857.

[12] Food and Drug Administration (2004). Innovation/Stagnation: Challenge and Opportunity in the Clinical Path to New Medical Products. FDA report from March 2004, available at http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

[13] Food and Drug Administration (2010). Guidelines for Industry: Adaptive Design Clinical Trials for Drugs and Biologics. FDA guidance from February 2012, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf.

[14] Fisher, L. D. (1998). Self-designing clinical trials. *Stat. Med.* **17** 1551–1562.

[15] Hoeffding, W. (1960). Lower bounds for the expected sample size and the average risk of a sequential procedure. *Ann. Math. Statist.* **31** 352–368. MR0120750

[16] Huang, X., Ning, J., Li, Y., Estey, E., Issa, J. P. and Berry, D. A. (2009). Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Stat. Med.* **12** 1680–1689. MR2675244

[17] Hung, J. H. M. and Wang, S. J. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharm. Statist.* **19** 1–11. MR2655585

[18] Inoue, L. Y. T., Thall, P. F. and Berry, D. A. (2002). Seamlessly expanding a randomized Phase II trial to Phase III. *Biometrics* **58** 823–831. MR1945019

[19] Jennison C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat. Med.* **22** 971–993.

[20] Julious, S. A. and Swank, D. J. (2005). Moving statistics beyond the individual clinical trial: Applying decision science to optimize a clinical development plan. *Pharmaceutical Statistics* **4** 37–46.

[21] Lai, T. L., Lavori, P. W. and Shih, M. C. (2012). Sequential design of phase II-III cancer trials. *Stat. Med.* **31** 1944–1960. MR2956028

[22] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22. MR0776826

[23] Lai, T. L. and Shih, M. C. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* **91** 507–528. MR2090619

[24] Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55** 1286–1290.

[25] Lipsky, P. E. et al. (2000). Infliximab and methotrexate in the treatment of rheumatoid arthritis: Anti-tumor necrosis factor trial in rheumatoid arthritis with concomitant therapy study group. *N. Engl. J. Med.* **343** 1594–602.

[26] Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biom. J.* **41** 689–696.

[27] Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51** 1315–1324.

[28] Rubinstein, L., Crowley, J., Ivy, P., LeBlanc, M. and Sargent, D. (2009). Randomized phase II designs. *Clin. Cancer Res.* **15** 1883–1890

[29] Simon, R. (1989). Optimal 2-stage designs for phase II clinical trials. *Contr. Clin. Trial* **10** 1–10.

[30] Shen, Y. and Fisher, L. D. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55** 190–197.

[31] Shih, W. J. (2001). Sample size re-estimation: Journey for a decade. *Stat. Med.* **20** 515–518.

[32] Shun, Z., Lan, K. K. G. and Soo, Y. (2008). Interim treatment selection using the normal approximation approach in clinical trials. *Stat. Med.* **27** 597–618. MR2418468

[33] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16** 243–258. MR0013885

[34] Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90** 367–378. MR1986653

[35] Wang, Y., Lan, K. K. G., Li, G. and Ouyang, S. P. (2011). A group sequential procedure for interim treatment selection. *Stat. Biopharm. Res.* **3** 1–13.

[36] Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9** 65–72.

[37] Whitehead, J., Whitehead, A., Todd, S., Bolland, K. and Sooriyarachchi, M. (2001). Mid-trial design reviews for sequential clinical trials. *Stat. Med.* **20** 165–176.

[38] Vickers, A. J., Ballen, V. and Scher, H. I. (2007). Setting the bar in phase II trials: The use of historical data for determining “go/no go” for definitive phase III testing. *Clin. Cancer Res.* **13** 972–976.

Tze Leung Lai
Department of Statistics
Stanford University
California
USA
E-mail address: lait@stanford.edu

Olivia Yueh-Wen Liao
Department of Statistics
Stanford University
California
USA
E-mail address: yuehwen@stanford.edu

Ray Guangrui Zhu
Allergan Pharmaceuticals
Irvine, California
USA
E-mail address: rayzhuhome@gmail.com