# Adaptive Trial Designs

Tze Leung Lai,[1,2] Philip William Lavori,[1,2] and Mei-Chiung Shih[2,3]

[1]Department of Statistics, Stanford University, Stanford, California 94305; email: Lavori@Stanford.edu

[2]Department of Health Research and Policy, Stanford University, Stanford, California 94305

[3]Cooperative Studies Program, U.S. Department of Veterans Affairs, Mountain View, California 94043



Annu. Rev. Pharmacol. Toxicol. 2012. 52:101–10

First published online as a Review in Advance on August 11, 2011

The *Annual Review of Pharmacology and Toxicology* is online at pharmtox.annualreviews.org

This article's doi: 10.1146/annurev-pharmtox-010611-134504



0362-1642/12/0210-0101$20.00

## Keywords

group-sequential, variance spending, Bayesian methods, randomization

## Abstract

We review adaptive designs for clinical trials, giving special attention to the control of the Type I error in late-phase confirmatory trials, when the trial planner wishes to adjust the final sample size of the study in response to an unblinded analysis of interim estimates of treatment effects. We point out that there is considerable inefficiency in using the adaptive designs that employ conditional power calculations to reestimate the sample size and that maintain the Type I error by using certain weighted test statistics. Although these adaptive designs have little advantage over familiar group-sequential designs, our review also describes recent developments in adaptive designs that are both flexible and efficient. We also discuss the use of Bayesian designs, when the context of use demands control over operating characteristics (Type I and II errors) and correction of the bias of estimated treatment effects.

## INTRODUCTION

**FDA:** United States Food and Drug Administration

**AWC:** adequate and well-controlled

**Type I error:** rejecting a true null hypothesis; in a clinical trial, it is equivalent to falsely claiming a treatment effect

**Type II error:** accepting a false null hypothesis; in a clinical trial, it is equivalent to missing a true treatment effect

**GS:** group-sequential

In its 2010 Draft Guidance for Industry (1), the United States Food and Drug Administration (FDA) defines adaptive design features as "changes in design or analyses guided by examination of the accumulated data at an interim point in the trial" (p. 1). In a recent review, Coffey & Kairalla (2) cite a definition for an adaptive trial developed by a Pharmaceutical Research and Manufacturers of America working group: "a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial." The aspects of a trial that may be changed or affected include the planned sample size; the dose of drug; the decision to move from Phase II to Phase III (seamless designs); the probabilities used in the randomization of new subjects; and other aspects including the primary outcome, test statistic, and inclusion/exclusion criteria. Both the Draft Guidance and the review by Coffey & Kairalla emphasize the importance of planning for all such changes from the outset, so that the adaptive features are specified in the protocol. Therefore, these reports (and others, but not all) regard the design as fixed, whereas the trial's aspects or features adapt as directed by the design.

There are several reasons for the considerable interest in such trials—noted in the FDA guidance (1) and Coffey & Kairalla review (2)—including the possibility of reducing the duration of trials or the number of patients required for study, the increased power to detect effects, or other increases in information value. Noting that there are well-understood, familiar methods of adaptive trial design as well as less familiar, novel approaches, the FDA document describes aspects of adaptive design trials that "deserve special consideration" (1, p. 2). The focus of the document is on so-called adequate and well-controlled (AWC) trials that form the basis for marketing a drug or biologic and have the greatest impact on FDA regulatory review. In these late-phase trials, it is emphasized that avoiding Type I errors (false positives) is critical. The acceptable probability of a Type I error (conventionally denoted by $\alpha$) is usually fixed by the trial planner. For this reason, sponsors are urged to consider adaptive features in the exploratory studies that begin the clinical development process (e.g., Phase I and Phase II trials), when avoiding Type II errors may be more important. The acceptable probability of a Type II error [which depends on the effect size (see below) and is conventionally denoted by $\beta$] is also defined by the trial planner. Coffey & Kairalla also briefly mention some adaptive features of interest in the exploratory phase, including Bayesian alternatives to the traditional 3 + 3 design for dose finding as well as to the traditional analysis of variance (ANOVA) design for dose ranging.

In this review, we concentrate on the design of AWC trials, particularly to show how some of the goals of the less familiar, novel approaches can be achieved by using the full strength of standard techniques—particularly group-sequential (GS) methods. In particular, we take the need to avoid false positives in AWC trials as a basic necessity, which leads us to a gentle critique of the current enthusiasm for Bayesian methods. We discuss only features that threaten to increase the Type I error rate substantially, effectively limiting our focus to methods that use interim analyses of the treatment effect (unblinded analyses comparing outcomes on study treatments). Other features [such as reestimation of the variance of measures (3, 4) to recalculate sample sizes] that do not substantially affect the Type I error rate therefore are not discussed here. They have been well covered in previous reviews (2, 3, 4, 5). Thus, from here on, an adaptive trial is defined as one that uses the current estimates of treatment effect, based on interim analyses, to change some aspect of trial conduct.

Adaptively altering the trial sample size has the greatest potential of offering benefit and posing risk to an AWC trial. Of all the aspects of a trial that can be adaptively changed, it is the one that has garnered the most attention in the recent methods literature (6, 7). We also discuss Bayesian design, which goes beyond sample-size reestimation to adaptively altering the randomization probability.

## WHY ADAPT THE SAMPLE SIZE OF A TRIAL?

In the design of a standard fixed-sample-size clinical trial, the specification of the required sample size is one of the most critical decisions to be made. The trial's cost, duration, feasibility, multisite scale, and logistical complexity depend on the total sample size. However, the sample size is connected to the power and the size of the treatment effect targeted by the trial. Specifically, for a fixed power and significance level, the required sample size scales with the inverse square of the effect size (in appropriate units, in a conventional fixed-sample-size design).

**Power:** the probability of not committing a Type II error (i.e., 1-β)

**Effect size (denoted δ):** the expected difference or ratio (in appropriate units) in the outcomes associated with the two treatments in a trial

For example, a conventional fixed-sample-size study to detect a 40% difference in response rate, from 20% (control treatment) to 60% (experimental treatment), with conventional 5% Type I and 20% Type II error rates (i.e., 80% power) requires 23 subjects per group, whereas the same study with half of the effect size (from 20% to 40%) requires approximately 82 subjects per group (nearly a fourfold increase). The prior uncertainty about the size of the effect often covers a wide range, with at least some hope that a large effect is possible, but the planners also might be justified in thinking that the lower range of effects would still be of clinical or commercial interest. Initially, it may be hard to commit to a study of sufficient size to guarantee power at the smallest clinically important and at least plausible effect size (say, the 164-patient study designed to detect a 20% difference in response rates described above). Therefore, the planners may want to start with a study that is small (perhaps the 46-patient study designed to detect an effect at the optimistic end of the continuum of plausible differences in response, e.g., a 40% difference). Then at some interim point (perhaps halfway, at 20 patients or so), the difference is observed, and by some criterion the final sample size is adapted to that interim result.

Tsiatis & Mehta (6) put it this way:

> Statisticians are often involved in the design of clinical trials where there is no clear criterion for what constitutes a clinically important treatment difference. Thus, the idea that there exists a design where a trial is started based on some "rough" guess of an anticipated treatment difference but allows the option of adaptively changing the design using the emerging treatment difference has a great deal of appeal (6, p. 375).

For example, if the current response rates halfway through recruitment are 10% in the control group and 30% in the experimental group, the difference so far is only 20%, lower than the effect that the study was designed to detect but perhaps still of clinical or commercial interest. The investigators then might want to increase the total sample size to obtain a reasonable chance that the final analysis will reject the null hypothesis of "no treatment effect" (i.e., be statistically significant). Armed with the encouraging interim result, they may be successful in garnering the sponsor's support for such an increase. Andersen & Liu (7) [who are quoted by Jennison & Turnbull (8)] refer to this tactic as "start small and ask for more."

Unfortunately, a naively applied strategy such as that described above has the fatal effect of inflating the true Type I error rate beyond the nominal rate used in the routine statistical significance calculation at the end of the study because it does not adjust for the adaptive decision. In short, an investigator who does many studies in this way and who uses a standard significance criterion of 5% in the final test will reject a higher-than-5% proportion of true null hypotheses (i.e., claim a treatment effect when there is none more than 5% of the time). This long-known and elementary fact prohibits the use of such naive adaptive strategies in AWC trials. A variety of ways to fix the Type I error inflation problem have been proposed over the past decade, some of which have generated a great deal of attention, accounting for much of the recent interest in adaptive designs (10, 11). These methods initially had great appeal, as indicated by the Tsiatis

& Mehta (6) quote, because they seem to offer a highly flexible way to adapt the sample size to evolving knowledge about the true effect size, even to the point of being self-designing—a term used by Fisher (10)—while still preserving the Type I error rates that are so crucial to AWC trials. However, over the past 10 years, each of the new adaptive methods has been subjected to methodological scrutiny and compared in performance with standard, well-understood methods that also offer an alternative to the fixed-sample-size design (8, 9, 12).

**Stopping for effect (or futility):** halting a trial at an interim analysis and deciding in favor of a treatment effect (or the absence of one)

The well-understood methods that we describe and discuss are the familiar GS methods of interim analysis, which yield the possibility of early stopping for effect or futility (12). The GS methods have a firm statistical underpinning and a complete asymptotic theory of optimal design. The control of the Type I error is built into the foundation of GS designs. Years of successful development and application have provided a rich source of examples and templates; there are commercial-grade software packages, such as East® and PEST (Planning and Evaluation of Sequential Trials) (13), and many expositions at all levels of technical sophistication (9, 12). One main theme of this review is that GS methods can be used to achieve remarkably good results compared with less well-understood and less widely used approaches, and thus GS designs deserve a first look from trial planners trying to achieve the goals described in the FDA Draft Guidance (1).

## THE GROUP-SEQUENTIAL DESIGN

During the Second World War, in response to demands for more efficient sampling inspection of weaponry, Wald (14) introduced a way to test which of two hypotheses to accept and which to reject on the basis of updates to the statistical comparison after every observation. For example, in a clinical trial, suppose an investigator wanted to test whether a new treatment had the same response probability p as the standard treatment ($p = p_0$) or a larger probability ($p = p_1$). Then after each patient's response was observed (on the new treatment), the likelihood of the observed history of responses would be recalculated under the assumption that $p = p_1$ and again under the assumption that $p = p_0$. If the ratio of these likelihoods exceeds the boundary $c_1$, the hypothesis that $p = p_1$ is accepted (meaning $p = p_0$ is rejected), and the study stops (vice versa if the likelihood ratio falls below $c_0$). The constants $c_1$ and $c_0$ are chosen to guarantee that the probability of getting the answer wrong is $\alpha$ when $p = p_0$ and $\beta$ when $p = p_1$. The sequential probability ratio test (SPRT), as it was termed, had a desirable property: Of all the statistical testing procedures with the same error probabilities, it has the smallest expected sample size (the sample size is a random variable in the SPRT procedure). However, even though the idea of optimally efficient tests is highly appealing in the medical context, the SPRT did not see much use because it applied only when the likelihood ratio could be recomputed after each observation and was efficient only for the comparison of two definite hypotheses (among other less critical difficulties). In clinical trials, results may not be instantaneously available relative to the accrual of subjects, so it was necessary to generalize to the now familiar GS calculation, whereby the updating of the statistic is done in batches of patients. These batches usually coincide with the schedule of interim analyses done for the ethical monitoring of a trial by a Data and Safety Monitoring Board.

The GS design has been developed to include many important features of relevance to clinical trials, including symmetric and asymmetric stopping rules (allowing stopping for futility as well as efficacy), rules that allow flexible scheduling of interim analyses ($\alpha$ spending rules), and rules that depend on the observation of events in survival trials (information-driven trials). There are GS designs appropriate for just about any kind of outcome (binary, time-to-event, repeated-measure, and so on). The comprehensive and authoritative text by Jennison & Turnbull (12) is an excellent reference. However, the second problem with the SPRT mentioned above (that its optimality

depends on the specification of two definite alternatives) has been a harder problem to solve. In clinical trials, one hypothesis often is quite definite (the null hypothesis, which describes no difference between the treatments in survival, response, or other outcomes of interest), whereas the other hypothesis (the alternative) often is less definite, as discussed above. Thus, the GS design confronts the same problem that motivated the adaptive designs described above.

**M:** the sample size implied by the smallest clinically or commercially important treatment effect size (which is denoted $\Delta$)

**ASN:** average sample number

The way the GS trial designer solves the problem reverses the logic of the "start small and ask for more" approach of Andersen & Liu (7). Instead, the GS designer asks for the smallest clinically interesting effect ($\Delta$) to be identified up front; the total upper sample size M is then defined as the size of the fixed-sample-size study necessary to detect the effect $\Delta$ with adequate (specified) power and Type I error rate. In fact, the sample size is inflated slightly (often by as little as 10%) to the maximal sample size $M^*$, for technical reasons that need not concern us here. The GS design then specifies the number of interim looks (the groups) and their sizes, as well as the criteria for stopping or continuing. These choices are tuned (before the study starts) to achieve a given power curve, which is the probability of declaring statistical significance at the end of the study. Note that the end of the study could occur at any look. The power curve is expressed as a function of the true effect size $\delta$, which ranges from 0 through $\Delta$ to $L\Delta$, the largest effect size for which there is any substantial hope ($L > 0$). Typically, $L = 2$ or a bit more. Jennison & Turnbull (8) refer to such GS designs as nonadaptive because the sequence of group sizes and boundary values is fixed at the outset and does not depend on the values of the difference statistic observed during the study, as in the adaptive design described above.

At $\delta = 0$ the null hypothesis holds, and the power is the Type I error rate, which is pinned at the desired nominal significance level (5%, for example). The desired power (say, 80%) is pinned at $\Delta$, whereas the rest of the curve describes how the power to detect any particular effect size $\delta$ grows with $\delta$. The part of the curve the investigator should care about is between $\delta = \Delta$ and $\delta = L\Delta$; outside that range, the effect is either too small to be interesting or too large to be plausible. This GS design might best be described as "ask for what you might need, and hope you need less." As in the adaptive design, the actual final sample size of the GS design is a random variable, which can range over the possible group sizes at which the analyses take place, up to the maximal sample size $M^*$. The performance of a particular design can be assessed by the distribution of that sample size, especially by its mean (the average sample number, or ASN).

Jennison & Turnbull (8) point out that "traditional nonadaptive group-sequential tests and more recent adaptive designs are fundamentally similar in form" (p. 919), describing their implementation as a sequence of steps: Randomize and obtain outcomes on the first group with its specified group size, decide whether to stop or continue on the basis of whether a statistic has crossed a boundary, randomize the second group, decide whether to stop or continue, and so on. The GS design (which Jennison & Turnbull refer to as nonadaptive) specifies the group sizes and boundaries in advance, whereas the (so-called) adaptive design allows the subsequent group sizes and stopping boundaries to depend on the current outcome data. They make the important observation that the performance that really matters is the true overall power curve. That is, one begins with an unknown state of nature and applies a design method for planning a trial to decide if the true state is "no treatment effect." That method (whatever its moving parts) then has a definite probability of success at each true value of the treatment effect (the power curve) and a corresponding ASN. Because the required power curve should be defined by the problem and resources at hand, it does not depend on the method. Rather, it is used as input to the design method ("I need nominal 5% power at the null and 80% power at the smallest interesting effect, and I want the power curve to rise sharply with larger effects"). The ASN (or other measure of the design's efficiency) should be used to compare methods. Thus, two designs with a similar power curve can be compared on the number of patients required, on average, to achieve their results.

Because the ASN is a function of the effect size, it is useful to summarize performance over several effect sizes; Jennison & Turnbull (8) use the average ASN at the null, at the smallest effect, and at the larger effect that investigators hope for, among other metrics to assess. With this setup, they compare four classes of designs: (*a*) the equal-group-size GS trial, (*b*) the GS trial that chooses the best first group size optimally and the rest as equal, (*c*) the optimal nonadaptive GS trial, and (*d*) the optimal adaptive GS trial. These four classes increase in complexity and performance (with a lower ASN) because they employ more and more optimization. Jennison & Turnbull (8) find that class *b* does appreciably better than does class *a*, but there is little gain in going to class *c* or *d*. Thus, the best possible design that includes adaptation does not perform much better than a simpler nonadaptive GS trial performs. However, what about the particular adaptive trial designs that have been proposed and described?

## INEFFICIENT ADAPTIVE DESIGNS

As each adaptive trial design has been proposed, its performance has been studied. The results are sobering. Tsiatis & Mehta (6) showed that all the so-called adaptive designs based on conditional power as well as the standard GS designs could be represented with the same mathematical form. The way they do this helps reveal the problem. The adaptive design stops the study to accept (reject) the null hypothesis at the first look if the test statistic is below the lower boundary (above the upper boundary). Therefore, sample-size expansion occurs if, at the first look, the test statistic lies in between the two boundaries. The sample-size expansion is a function of the current value of the test statistic; it is smallest if the statistic is at the upper boundary and largest if the statistic is at the lower boundary. Therefore, the possible values of that sample-size expansion define a partition of the interval between the boundaries. However, we can consider this a multistage GS design with no chance of stopping until the desired second-stage sample size is reached and with boundaries at that point defined by the need to control the error rates. Tsiatis & Mehta then showed that any adaptive trial could be improved through the use of a GS design:

> For any adaptive design, one can always construct a standard group-sequential test based on the sequential likelihood ratio test statistic that, for any parameter value in the space of alternatives, will reject the null hypothesis earlier with higher probability, and for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability (6, p. 375).

Tsiatis & Mehta (6), as well as Jennison & Turnbull (8), believe that some of the inefficiency of the proposed adaptive designs lies in the way that they reweight observations to control the Type I error rate. Such reweighting implies that the test statistic is no longer a function of the so-called sufficient statistics for the data, and this reweighting, by standard theory, causes a loss of efficiency (the GS designs use fully efficient statistics, such as the generalized likelihood ratio). Bartroff & Lai (15, 16) add another source of difficulty: reliance on the noisy early estimate of treatment effect through the calculation of a conditional power. One can see why this is true by considering the representation of the adaptive trial as a multilook GS trial with only one look (after the first) that has an actual chance of stopping. If the true effect is moderately different from the estimate at the first stage, then forcing the only effective look to be at the particular sample size defined by the conditional power will be inefficient. By contrast, the GS design explores other potential sample sizes, in particular those that come earlier, allowing it to stop earlier if the effect is actually larger than the first-stage estimate. The emerging realization is that the new adaptive designs are in fact inefficient versions of the general GS design.

Jennison & Turnbull (8) note that proposed adaptive designs based on conditional power do not have rules that resemble those of the optimal GS design, and others have noted dramatic

underperformance in such designs (6). For example, Jennison & Turnbull (8) describe a simple GS (class *a*) trial with a maximal sample size of 606, whose average ASN over the null, the smallest effect, and four times that effect is 295, which is reduced to 272 by optimal choice of the first group size (to capitalize on the ability to stop very early if the true effect is large). Further improvement is negligible: Optimizing without adaptation reduces the ASN to 264; optimizing with adaptation, to 261. By contrast, the two-stage adaptive trial described by Shun et al. (11) has an ASN exceeding 314 in this context.

The original appeal of the variance spending approach (10) was to be able to do an unplanned analysis, without needing to specify in advance the rules for deciding whether to increase the sample size and by how much, and still preserve the Type I error rate. The variance spending method proposes that investigators start with a small sample size based on an optimistic projection of the treatment effect, and if interim results are less optimistic but still potentially worthwhile, that they increase the size of the study. By contrast, a GS design has to be more specific about the rules, but it can accomplish the same goal. The GS design proposes that investigators start with the smallest clinically important design, which translates into a maximal sample size M, and that they then perform interim monitoring for early stopping, either for futility or for a greater-than-projected effect. By turning the variance spending method on its head, one can convert the approach of "start small and ask for more" into "start large, and if you can, stop early."

The GS method also can use the interim result to self-tune, but in a more subtle way. The interim estimates can be used to change the timing of subsequent analyses or to change the way that the Type I error is spent. These adjustments tend to concentrate the power of the trial in the vicinity of the true effect and thus to approach optimality. Lai & Shih (17) have developed a class of GS designs termed modified Haybittle-Peto designs that have a nearly optimal ASN and a power that is subject to constraints on the Type I error probability and maximum sample size. Whereas other approaches on optimal GS designs are developed exclusively in the framework of normally distributed outcome variables, Lai & Shih's GS design can be applied to generalized likelihood ratio statistics in much more general situations, including multivariate and multiarm settings.

## EFFICIENT ADAPTIVE DESIGNS

In their survey article on adaptive designs, Burman & Sonesson (18) point out that previous criticisms of the statistical properties of adaptive designs may be unconvincing in situations in which being flexible and not having to specify parameters that are unknown at the beginning of the trial are more imperative than efficiency. The optimal adaptive designs proposed by Jennison & Turnbull (8, 19) assume normally distributed outcome variables with known variances. However, as pointed out by Hung et al. (20), "very often, statistical information also depends on nuisance parameters (e.g., the standard deviation of the response variable)." Bartroff & Lai (15, 16) have developed a new approach that can serve to bridge the gap between the two camps in the adaptive design literature. One camp focuses on efficient designs, under restrictive assumptions, that involve sufficient statistics and optimal stopping rules; the other camp emphasizes flexibility to address the difficulties of coming up with realistic alternatives and guessing the values of nuisance parameters at the design stage. The approach of Bartroff & Lai yields adaptive designs that can fulfill the seemingly disparate requirements of flexibility and efficiency.

The adaptive designs described by Bartroff & Lai (15, 16) are similar to Lai & Shih's (17) modified Haybittle-Peto GS designs in that they use efficient test statistics of the generalized likelihood ratio type, which have an intuitively adaptive appeal by replacing unknown parameters with their maximum likelihood estimates. They are GS designs with three groups and the same

early stopping boundaries, for futility or efficacy, as those of the modified Haybittle-Peto designs. Their novel feature is that the sample size of the second group is based on the effect size estimated from the data available at the first interim analysis. Although this calculation is similar to the mid-course sample-size reestimation procedures of Wittes & Brittain (3), Birkett & Day (21), Proschan & Hunsberger (22), Posch & Bauer (23), and Li et al. (24), a major difference is that these two-stage designs do not incorporate the uncertainties in the effect size estimates, whereas Bartroff & Lai's (15, 16) adaptive designs include a third stage to account for such uncertainties. Another important innovation of the Bartroff & Lai (15, 16) approach is its algorithm, based on the Markov property of the observations generated by these adaptive designs, to compute the error probabilities so that the Type I error is maintained at the prescribed level and there is little loss in power due to futility stopping to accept the null hypothesis.

Another way to think of these adaptive designs, which focus exclusively on adapting the sample size to data accumulated in the first stage of the trial, is that they use the first stage of the trial as an internal pilot study to learn about the effect size so that the sample size for the trial can be determined from data rather than from a priori guesses. In addition to the sample size, other trial characteristics such as dose and endpoint also can be determined from an internal pilot.

## BAYESIAN DESIGN

Trial designs based on Bayesian principles provide a rich source of alternatives to the standard frequentist designs described above. Bayesian methods start from a different premise about the nature of inference and the statistical weighing of evidence of treatment effects. Instead of characteristics such as reproducibility of findings in hypothetical repeated trials—which are at the basis of concepts such as Type I and II errors, confidence intervals, and so on—the Bayesian trial designer values a certain coherence of the logic of scientific inference and focuses on the way that uncertainty in knowledge changes as data accumulate. This difference in what is thought to be of primary scientific value gives the Bayesian a great deal of freedom in study design, essentially because "the design doesn't matter." To be more specific, the Bayesian starts with a prior distribution for the treatment effect (reflecting what is thought to be true before the study) and a probability model giving the likelihood of the data as a function of the parameters. Then the entire inference process consists of updating the state of knowledge about the treatment effect, i.e., the posterior probability distribution of the treatment effect (the probability that the effect is at least $x$, for each value $x$) given the data. From the Bayesian point of view, all the information about the treatment effect is contained in the posterior. In particular, the way the data are generated (the design) does not add anything (as long as the likelihood is specified correctly). The Bayesian can deploy a wide variety of designs, including stopping rules and assignment mechanisms that depend on the observed treatment effects and the use of posterior distributions. Bayesian inference implies that one does not have to make adjustments for early stopping or mid-course adaptation (19).

Of course, the laws of nature as they apply to operating characteristics have not been repealed. Thus, the Bayesian designs generally will not guarantee good Type I and II errors or unbiased treatment effects, and they also rely on the choice of the prior distribution; the latter can be problematic in confirmatory trials, when trial designers are trying to convince the scientific community and regulatory agencies of the value of the trial. Therefore, if the investigator desires to know (and control) Type I error rates, the Bayesian must attempt to estimate them, usually by simulations under some assumed parameter configurations. As the models get more elaborate, this process poses computational challenges, but the biggest challenge is the need to assume a parametric model for the data, which is used to simulate trial results. Then the estimates of error rates are only as good as the model. At this point, there is considerable resistance in regulatory

agencies to the reliance on such methods for AWC trials, especially late-phase, pivotal trials. For earlier exploratory studies (in which the consequences of error are not visited upon the public), there is more encouragement for the use of Bayesian designs, particularly in Phase II testing of biomarker-guided treatment strategies and of other components of personalized medicine. In this realm, the Bayesians have taken the lead.

### SUMMARY POINTS

1. Late-phase studies (adequate and well-controlled studies, or AWC studies, in the FDA terminology) require control over the Type I error and adjustment for bias of treatment estimates.
2. There is a growing consensus that the goals of adaptive trials can be met within the confines of the well-understood group-sequential (GS) design, with no loss in efficiency.
3. Adaptive designs that are based on the calculation of interim estimates of conditional power can be inefficient owing to the high variability of the estimates and the use of reweighting methods that do not use the sufficient statistics.
4. Bayesian methods must be used carefully when operating characteristics are important.
5. There is a need for better understanding of the costs and benefits of these newer methods, relative to new improvements on more traditional GS designs.



## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors acknowledge support from grant P30 CA124435 from the National Cancer Institute and from the Clinical and Translational Science Award UL1 RR025744 for the Stanford Center for Clinical and Translational Education and Research (Spectrum) given to Stanford University by the National Institutes of Health's National Center for Research Resources.

## LITERATURE CITED

1. U.S. Food and Drug Administration. 2010. *Draft Guidance for Industry: adaptive design clinical trials for drugs and biologics*. **http://www.fda.gov/downloads/DrugsGuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf**
2. Coffey CS, Kairalla JA. 2008. Adaptive clinical trials: progress and challenges. *Drugs R D* 9:229–42
3. Wittes J, Brittain E. 1990. The role of internal pilots in increasing the efficiency of clinical trials. *Stat. Med.* 9:65–71; discussion 71–72
4. Gould AL, Shih WJ. 1992. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Commun. Stat. Theory Methods* 21:2833–53
5. Chow S, Chang M. 2007. *Adaptive Design Methods in Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC
6. Tsiatis AA, Mehta C. 2003. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 90:367–78

7. Anderson K, Liu Q. 2004. *Optimal adaptive versus optimal group sequential design*. Presented at Conf. Adapt. Des. Clin. Trials, March 4–5, Philadelphia
8. Jennison C, Turnbull BW. 2006. Efficient group sequential designs when there are several effect sizes under consideration. *Stat. Med.* 25:917–32
9. Mehta C, Gao P, Bhatt DL, Harrington RA, Skerjanec S, Ware JH. 2009. Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* 119:597–605
10. Fisher LD. 1998. Self-designing clinical trials. *Stat. Med.* 17:1551–62
11. Shun Z, Yuan W, Brady WE, Hsu H. 2001. Type I error in sample size re-estimations based on observed treatment difference. *Stat. Med.* 20:497–513; commentary 515–18; rejoinder 519–20
12. Jennison C, Turnbull BW. 2000. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC
13. Wassmer G, Vandemeulebroecke M. 2006. A brief review on software developments for group sequential and adaptive designs. *Biom. J.* 48:732–37
14. Wald A. 1947. *Sequential Analysis*. New York: Wiley
15. Bartroff J, Lai TL. 2008. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. Med.* 27:1593–611
16. Bartroff J, Lai TL. 2008. Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Seq. Anal.* 27:254–76
17. Lai TL, Shih MC. 2004. Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* 91:507–28
18. Burman CF, Sonesson C. 2006. Are flexible designs sound? *Biometrics* 62:664–69; discussion 670–83
19. Jennison C, Turnbull BW. 2006. Adaptive and nonadaptive group sequential tests. *Biometrika* 93:1–21
20. Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. 2006. A regulatory view on adaptive/flexible clinical trial design. *Biom. J.* 48:565–73
21. Birkett M, Day S. 1994. Internal pilot studies for estimating sample size. *Stat. Med.* 13:2455–63
22. Proschan M, Hunsberger S. 1995. Designed extension studies based on conditional power. *Biometrics* 51:1315–24
23. Posch M, Bauer P. 1999. Adaptive two stage designs and the conditional error function. *Biom. J.* 41:689–96
24. Li G, Shih WJ, Xie T, Lu J. 2002. A sample size adjustment procedure for clinical trials. *Biostatistics* 3:277–87

Annual Review of Pharmacology and Toxicology

Volume 52, 2012

# Contents

**Indexes**



**Errata**

An online log of corrections to *Annual Review of Pharmacology and Toxicology* articles may be found at http://pharmtox.annualreviews.org/errata.shtml