Adaptive Design of Confirmatory Trials: Advances and Challenges

Tze Leung Lai[a,*], Philip W. Lavori[b], Ka Wai Tsang[c],

[a]*Department of Statistics, Stanford University, Stanford, CA, USA*
[b]*Department of Health Research and Policy, Stanford University, Stanford, CA, USA*
[c]*Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA*

Abstract

The past decade witnessed major developments in innovative designs of confirmatory clinical trials, and adaptive designs represent the most active area of these developments. We give an overview of the developments and associated statistical methods in several classes of adaptive designs of confirmatory trials. We also discuss their statistical difficulties and implementation challenges, and show how these problems are connected to other branches of mainstream Statistics, which we then apply to resolve the difficulties and bypass the bottlenecks in the development of adaptive designs for the next decade.

Keywords: Adaptive design, Adaptive randomization, Bayesian inference, Early stopping, Hybrid resampling, Multi-arm bandits

1. Introduction

Because of the lack of information on both the magnitude and the sampling variability of the treatment effect of a new treatment at the design stage, there has been increasing interest from the biopharmaceutical industry in adaptive designs that can adapt to the information collected during the course of the trial. Beginning with Bauer [1], who introduced sequential

*Corresponding author at: Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA, Tel: +1 6507232622
Email address: lait@stanford.edu (Tze Leung Lai)

adaptive test strategies over a planned series of separate trials, and Wittes and Brittain [2] who considered internal pilot studies, a large literature has grown on adaptive design of clinical trials. In Section 2 we review several directions of development and basic methodologies in that literature. Despite the vibrant research activities and the attractiveness of adaptive designs that provide a promising alternative to and major advance over standard clinical trial designs which are handicapped by insufficient information at the planning stage, these adaptive designs are fraught with statistical and implementation difficulties which have been impediments to their widespread use. Section 3 discusses these difficulties and reviews in this connection related aspects of the FDA Draft Guidance for Industry on Adaptive Design, for drugs and biologics, in 2010.

In Section 4 we describe some new advances in adaptive designs to address these difficulties and to respond to certain issues raised by the FDA Draft Guidance. We also use an adaptive clinical trial currently being planned at the Stanford Stroke Center to illustrate the new methodologies and their implementation. Section 5 gives some concluding remarks and further discussion of the challenges and opportunities of adaptive designs for Phase III clinical trials in drug development.

2. Adaptive designs: Overview of methods and developments

In this section we give an overview of the developments of adaptive design of clinical trials together with the associated statistical methods that have been used or introduced. The overview is divided into two parts, the first of which is on frequentist methods, reviewed in Sections 2.1 and 2.2. The second part is on Bayesian adaptive designs, which are reviewed in Section 2.3 and which are arguably the most active area of clinical trial innovations for testing cancer treatments.

2.1. Sample size re-estimation

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 [2], 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 specific problem considered by [2] as an example of internal pilots actually dated back to Stein's two-stage procedure [3] introduced in 1945 for testing hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative $\mu_X \neq \mu_Y$ for the means of two independent normal distributions with common, unknown variance, and based on i.i.d. observations $X_1, X_2, \cdots \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \cdots \sim N(\mu_X, \sigma^2)$. Let $t_{\nu,\alpha}$ denote the upper α-quantile of the t-distribution with ν degrees of freedom. In its first stage, Stein's procedure samples n_0 observations from each of the two normal distributions and computes the usual unbiased estimate s_0^2 of σ^2. In the second stage, it samples up to

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

observations from each population, where α is the prescribed type I error probability, and $1-\beta$ is the prescribed power at the alternatives satisfying $|\mu_X - \mu_Y| = \delta$. The null hypothesis $H_0 : \mu_x = \mu_Y$ is then rejected if $|\bar{X}_{n_1} - \bar{Y}_{n_1}| > t_{2n_0-2,\alpha/2}\sqrt{2s_0^2/n_1}$. Stein's two-stage procedure is modified in [2, 4] as follows. Viewing $|\bar{X}_{n_1} - \bar{Y}_{n_1}|/\sqrt{2s_1^2/n_1}$ as a fixed-sample test statistic based on a sample of size n_1 from each population, the test statistic has the non-central t-distribution with $2n_1 - 2$ degrees of freedom and non-centrality parameter $\delta\sqrt{n_1/(2s_1^2)}$ at the alternative $\mu_X - \mu_Y = \delta$. Fixing α, β and δ, let $n(\sigma^2)$ denote the smallest n_1 for which the probability exceeds $1-\beta$ that an observation from this distribution exceeds the critical value $t_{2n_1-2,\alpha/2}$. An estimate of the total desired sample size based on a pre-trial estimate σ_0^2 of σ^2 is $n(\sigma_0^2)$. Following a pilot study of size n_0 per arm, which results in the variance estimate s_0^2, the total sample size can be re-estimated as $n(s_0^2)$. At this point there are many options for how to proceed. In particular, [2] recommends taking the maximum of $n(\sigma_0^2)$ and $n(s_0^2)$ as the new total sample size, while [4] recommends retaining $n(\sigma_0^2)$ unless $n(s_0^2)$ is substantially larger.

The aforementioned papers and subsequent refinements [5, 6, 7] represent the "first generation" of adaptive designs. The second-generation adaptive designs adopt a more aggressive viewpoint of re-estimating the sample size from the estimate of δ (instead of the nuisance parameter σ) based on the first-stage data, starting with Fisher [8] for the case of normally distributed outcome variables with known common variance σ^2, which can be assumed to equal $1/2$ without loss of generality. If n is the original sample size per treatment, then after rn pairs of observations ($0 < r < 1$), $n^{-1/2}S_1 \sim N\left(r\delta\sqrt{n}, r\right)$,

where $S_1 = \sum_{i=1}^{rn}(X_i - Y_i)$. 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 conditional on the first-stage data, $(n\gamma)^{-1/2}S_2 \sim N\left((1-r)\delta\sqrt{\gamma n}, 1-r\right)$, where $S_2 = \sum_{i=rn+1}^{n^*}(X_i - Y_i)$ and $n^* = rn + \gamma(1-r)n$ is the new total sample size per treatment. Note that under $H_0 : \delta = 0$, $(n\gamma)^{-1/2}S_2$ has the $N(0, 1-r)$ distribution regardless of the (data-dependent) choice of γ, thus Fisher's test statistic

$$n^{-1/2}\left(S_1 + \gamma^{-1/2}S_2\right) \tag{2}$$

has a $N(0,1)$ distribution under H_0. The corresponding test has been called a *variance spending test* because $1-r$ is the remaining part of the total variance 1 not spent in the first stage. Denne [9] proposed a test that also allows data-dependent updates of the total sample size but maintains the type I error probability by a seemingly different method. Denne's test chooses a critical value for S_2 that maintains the conditional type I error rate $P_{\delta=0}(S_1 + S_2 > z_\alpha\sqrt{n} \mid S_1 = s_1)$. Jennison and Turnbull [10] showed that this test is actually equivalent to Fisher's test, which they found to perform poorly in terms of expected sample size and power in comparison to group-sequential tests. Tsiatis and Mehta [11] independently came to the same conclusion, attributing this inefficiency to the use of the non-sufficient "weighted" statistic (2).

Working in terms of the z-statistic that divides a sample sum by its standard deviation, Proschan and Hunsberger [12] 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

$$\int_{-\infty}^{\infty} C(z_1)\phi(z_1)\,dz_1 = \alpha, \tag{3}$$

and suggested certain choices of $C(\cdot)$. Having observed the first-stage data Z_1, $H_0 : \delta = 0$ is rejected in favor of $\delta > 0$ after the second stage if $Z_2 > \Phi^{-1}(1 - C(z_1))$. Condition (3) ensures that the type I error probability of any test of this form is α. The tests proposed earlier by Bauer and Köhne [13] can be represented in this framework, as noted by Posch and Bauer [14]. The basic idea underlying these representations dated back to Bauer [1] who used it to develop sequential adaptive test strategies over a planned series of separate trials.

Assuming normally distributed outcomes with known variances, Jennison and Turnbull [15] introduced adaptive group sequential tests that choose the

jth group size and stopping boundary on the basis of the cumulative sample size n_{j-1} and the sample sum $S_{n_{j-1}}$ over the first $j-1$ groups, and that are optimal in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values, subject to prescribed error probabilities at the null and a given alternative hypothesis. They showed how the corresponding optimization problem can be solved numerically by using backward induction algorithms. They also showed in [16] that standard (non-adaptive) group sequential tests with the first stage chosen approximately are nearly as efficient as their optimal adaptive tests.

A new approach was developed by Bartroff and Lai [17, 18] in the general framework of multiparameter exponential families. It uses efficient generalized likelihood ratio (GLR) statistics in this framework and adds a third stage to adjust for the sampling variability of the first-stage parameter estimates that determine the second-stage sample size. The possibility of adding a third stage to improve two-stage designs dated back to Lorden [19]. Whereas Lorden used crude upper bounds for the type I error probability that are too conservative for practical applications, Bartroff and Lai overcame this difficulty by using new methods to compute the type I error probability, and also extended the three-stage test to multiparameter and multi-armed settings, thus greatly broadening the scope of these efficient adaptive designs.

2.2. Seamless Phase II/III trials with hypotheses selection at interim

Bretz and his collaborators [20, 21] at Novartis have extended Bauer's seminal ideas in [1] to develop a second generation of adaptive designs that are of much greater interest to drug development than sample size re-estimation. Highlighting the need for more efficient and effective drug development processes to translate the ongoing revolution in biomedical sciences to breakthroughs in treating diseases, [20] notes the inefficiency of contemporary Phase III trials that are "stand-alone confirmatory trials, ignoring information from previous phases," and argues for innovation through seamless Phase II/III designs that "aim at interweaving these (phases) by combining them into one single study conducted in two stages." The advantages of these adaptive seamless designs (ASDs), noted in [20, p. 624], are that they

(i) reduce the time to decide on, plan and implement the next phase,
(ii) save costs through the combination of evidence across two studies, and
(iii) get long-term safety data earlier as a direct consequence of following up the Phase II patients.

The basic idea underlying the ASDs in [20] is to extend to multiple testing of k hypotheses $H_0^1, \ldots, H_0^k$ the methods used in [1], [13] and [14] for combining the p-values, over the two stages in a two-stage procedure, of a directional null hypothesis on treatment effects. Here k represents the number of dose levels, or treatment regimens, of a new drug considered in a typical Phase II trial. At the interim analysis after the first stage (which corresponds to the Phase II component of the ASD), only one dose level (or treatment regimen), say the jth one, is selected for continuation in the second stage (corresponding to the Phase III component) of the ASD. Bretz et al. [20] apply the closed testing principle to all intersection hypotheses involving H_0^j, using the Simes method [22] to define for each intersection hypothesis the adjusted first stage p-values of the hypotheses in the intersection, and thereby keeping the family-wise error rate (FWER) controlled at a prespecified level. Although it controls the FWER, this way of combining the first- and second-stage data is very inefficient, as pointed out in [10, 11]. More efficient two-stage designs have been introduced by Stallard and Todd [23] and Wang et al. [24] for the case of normally distributed outcome variables with known variances, and nearly optimal ASDs have recently been developed by using deeper concepts and more powerful techniques that are described in Section 4.1. The procedures proposed in [20, 21], however, are widely regarded as being more flexible than those to [23, 24] and easier to extend to more complicated settings. In fact, Brannath et al. [25] have extended them to survival outcomes, and Jenkins, Stone and Jennison [26] provide a further extension by allowing an intermediate endpoint (such as progression-free survival) to be used to guide hypothesis selection at the end of the first stage, while using overall survival as the primary endpoint for the second stage. Because of the complexity of the problem, [26] does not discuss the inefficiency of combining the p-values from the two separate stages, even though one of its authors has advocated to use more efficient group sequential test statistics for considerably simpler testing problem in [10, 15].

2.3. Bayesian approach

Adding confusion to the debate between the "efficiency camp" represented by [10, 11, 15] on efficient designs, under restrictive assumptions, that involve sufficient statistics and nearly optimal stopping rules, and the "flexibility camp" that focuses on combining information in a flexible way from different stages of the trial to tackle complicated settings, is the "Bayesian camp" that purportedly has both efficiency and flexibility. Since Bayesian

inference is based on the posterior distribution, it does not need adjustments for learning and adaptation during the course of the trial. Acknowledging that the statistical benchmark to gain regulatory approval of a new treatment is to have a statistically significant result in the frequentist sense, at a specified Type I error, Bayesian adaptive designs for Phase III trials rely on Monte Carlo simulations, under a chosen parameter configuration belonging to the null hypothesis, but there is no guarantee that the Type I error is maintained by this approach at other parameter configurations for a composite hypothesis. Another regulatory issue with Bayesian designs is the choice of the prior distribution, which may be difficult to justify.

Despite the regulatory difficulties with the Bayesian approach, it is arguably the most active area of development in adaptive design of clinical trials. Berry [27] and Berry et al. [33] point out the flexibility and natural appeal of applying the Bayesian approach to midcourse adaptation in a trial, such as dropping an unfavorable arm of the new treatment being tested or modifying the randomization scheme, incorporation of historical and other related information, treatment of multiple endpoints and multiple sub-groups, missing data and deviations from the original study plan. Since Bayesian inference can be updated continually as data accumulate and is not tied to the design chosen, early stopping for safety and futility or efficacy are allowed. Besides early stopping based on the posterior probability of an efficacious outcome, Bayesian adaptive designs also use the posterior probabilities to determine randomization proportions to improve the outcomes of patients accrued to a trial. This idea dated back eight decades ago to Thompson [29] who proposed to randomize patients to one of two treatments with probability equal to the current posterior probability that it is the better treatment. He was motivated by the ethical consideration of exposing a minimal expected number of patients in the trial to the inferior treatment. Meuer, Lewis and Berry [30] point out how this outcome-adaptive randomization can be used to close the "therapeutic misconception" gap for recruiting patients to a trial, which is particularly important for "trials in time-sensitive conditions that require rapid decision making by patients or surrogates and by physicians":

> Some trial participants and family members believe that the goal of a clinical trial is to improve their outcomes — a misconception often reinforced by media advertising of clinical research. Clinical trials have primarily scientific aims and rarely attempt to

> collectively improve the outcomes of their participants.... Any benefit to an individual trial participant is a chance effect of randomization and the true, but unknown, relative effects of treatments.... Thus, even though serving as a research participant is essentially an altruistic activity, many clinical trial volunteers do not participate in research out of altruism. An adaptive clinical trial design can be used to increase the likelihood that study participants will benefit by being in a clinical trial.

Adaptive randomization based on posterior probabilities features prominently in many Bayesian adaptive oncology trials, particularly those at the M.D. Anderson Cancer Center. One such trial is the BATTLE (Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination) trial of personalized therapies for non-small cell lung cancer (NSCLC). It uses an adaptive randomization scheme to select $K = 5$ treatments for $n = 255$ NSCLC patients belonging to $J = 5$ biomarker classes. Let y_{mjk} denote the indicator variable of disease control of the mth patient in class j receiving treatment k. The adaptive randomization scheme is based on a Bayesian probit model for $p_{jk} = P(y_{mjk} = 1)$. The posterior mean $\gamma_{jk}^{(t)}$ of p_{jk} given all the observed indicator variables up to time t can be computed by Gibbs sampling. Letting $\hat{\gamma}_{jk}^{(t)} = \max(\gamma_{jk}^{(t)}, 0.1)$, the randomization probability for a patient in the jth class to receive treatment j at time $t+1$ is proportional to $\hat{\gamma}_{jk}^{(t)}$. Moreover, a refinement of this scheme allows suspension of treatment k from randomization to a biomarker subgroup. Simulations of the frequentist operating characteristics at some parameter configurations are used to determine the threshold for treatment suspension [31]. The BATTLE design, which "allows researchers to avoid being locked into a single, static protocol of the trial" that requires large sample sizes for multiple comparisons of several treatments across different biomarker classes, can "yield breakthroughs, but must be handled with care" to ensure that "the risk of reaching a false positive conclusion" is not inflated, as pointed out in an April 2010 editorial in *Nature Reviews in Medicine*, on such designs.

Besides BATTLE, another design mentioned in the editorial is that of I-SPY2 [32]. The I-SPY1 and I-SPY2 (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis) trials represent innovative approaches to multi-center clinical trials for the development and testing of biomarker-guided therapies for breast cancer. I-SPY1 involved ten cancer centers and the National Cancer Institute (NCI SPORE

program and cooperative groups) to identify the indicators of response to chemotherapy that would best predict the survival of women with high-risk breast cancer. During the four-year period 2002–2006, I-SPY1 monitored 237 breast cancer patients serially using MRI and tissue samples to study the cancer biology of responders and non-responders to standard chemotherapy in a neoadjuvant (or pre-surgery) setting. It found that tumor response evaluated in this manner was a good predictor of the patients' overall survival, and that tumor shrinkage during the treatment was a good predictor of long-term outcome and response to treatment. The findings of I-SPY1 set the stage for the I-SPY2 trial, an ongoing adaptive clinical trial of multiple Phase 2 treatment regimens (in combination with standard chemotherapy) in patients with tumors with varying genetic signatures. I-SPY2 was launched in 2010 as a collaborative effort between the NCI and cancer centers, the FDA and industry. The overall trial design includes a multi-institutional, multi-arm, adaptively randomized framework with therapeutic arms introduced as older arms graduate, or are dropped due to their high, or low, Bayesian predictive probability of being more efficacious than standard therapy, with pathologic complete response as the study endpoint. The study tests novel therapeutic arms against a standard therapy arm. Drugs that are found during the trial to have a sufficiently low predictive probability of being successful in a subsequent confirmatory phase III study are dropped from the study. As more mature treatment/biomarker signature combinations drop out for futility or graduate to a subsequent confirmatory phase, newer arms are designated to take their place, thereby generating increased efficiency in a dynamic and flexible framework.

Berry [27, p.33] acknowledges that "the flexibility of the Bayesian approach can lead to complicated trial designs" and that "institutional review boards and others involved in clinical research, including regulators when the trial is for drug or medical device registration, require knowing the trial design's operating characteristics." Markov chain Monte Carlo (MCMC) simulations are used to compute these operating characteristics, but there is a delicate computational issue because convergence of the MCMC algorithm to the stationary distribution, which is the target (posterior) distribution, is a theoretical concept that has not been accompanied by error bounds for practical implementation. Although there are diagnostics as described in [33, p.48–49], one can only implement simple checks in the simulation program to compute the operating characteristics, for which an upper bound on the number of iterations has also to be imposed to avoid getting into an infi-

nite loop. Therefore, although Berry [27, p.34] argues that "one can use the Bayesian approach to build a design and modify it to deliver predetermined frequentist characteristics, such as 5% false positive rate and 90% power at a particular difference in treatment effects," resulting in an "essentially frequentist" modified Bayesian design, the complex and somewhat fuzzy Type I error claim of this approach is not easy to gain regulatory acceptance. For more complicated trials such as those involving censored survival data, these frequentist modifications become formidable and the usual frequentist analysis without modification for data-dependent adaptation is carried out in these Bayesian adaptive designs, as in [33] that reports a trial comparing relapse-free survival for standard chemotherapy to that for capecitabine, using the hazard ratio for disease recurrence or death of the capecitabine group to the standard chemotherapy group. The trial was discontinued for futility at the first interim analysis and usual p-values and confidence intervals were reported in [33], ignoring that this was a group sequential design with a Bayesian stopping rule.

3. Statistical and regulatory issues

As we have reviewed in Sections 2.1 and 2.2, the flexibility and other advantages of adaptive designs over traditional clinical trial designs gained increasing acceptance, during the period 1990–2010, by the pharmaceutical industry for which a major concern was regulatory acceptance of these novel designs and the associated analyses. Accordingly much effort was devoted to maintaining the Type I error probability of the confirmatory test comparing the new treatment to an active control or placebo based on data from a data-dependent adaptive design. An "efficiency camp" of biostatisticians from academia subsequently raised issues with the efficiency of the methods introduced and questioned whether the inefficiency of most of these methods to protect the Type I error probability (e.g., by weighting the test statistics or p-values at different stages of the adaptive design) might outweigh the advantages of adaptation, and proposed to use standard group sequential designs instead. A Bayesian camp, also from academia, who felt that working with Bayesian posterior probabilities could deliver both flexibility and efficiency, then emerged. Although Bayesian designs were used in some highly visible oncology trials of biomarker-guided personalized therapies, the pharmaceutical industry was leery of using them for new drug applications

because of potential regulatory difficulties and needed guidance from regulatory agencies on what would be acceptable.

3.1. The 2010 FDA Draft Guidance for Industry on Adaptive Design

In February 2010, this much awaited FDA Draft Guidance [34] appeared, $2\frac{1}{2}$ years after the EMA (European Medicines Agency) reflection paper on adaptive designs [35]. Both documents discuss newly developed statistical methods that allow confirmatory research with data-driven changes of the design. The FDA Draft Guidance focuses mainly on "adequate and well controlled" study settings and defines an adaptive clinical study as one that includes a prospectively planned opportunity for modification of specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study. Analyses of the accumulating data are performed at prospectively planned timepoints within the study, in which "prospective" means that the adaptation was planned (and details specified) before examining the data. Avoiding increased rates of false positive study results (increased Type I error rate) is emphasized in the Draft Guidance, which also highlights the importance of minimizing statistical and operational biases introduced by adaptation. In particular, the Draft Guidance advises shielding the investigators as much as possible from knowledge of the chosen adaptive changes and from the unblinded data used in the interim analyses, because knowledge of the specific adaptation decisions can lead investigators to treat and evaluate patients differently, leading to operational bias.

Despite the warnings about potential biases and possible inflation of Type I error probability, the Draft Guidance acknowledges the value of adaptive designs to innovate clinical trials in the development of new drugs and biologics. It points out that "well-understood" methods represent, in many cases, well-established and relatively low-risk means of enhancing study efficiency and informativeness that may deserve wider use. It describes potential benefits of using Bayesian methods in clinical trials for medical devices, particularly with respect to their flexibility, use of all available evidence and predictive probability for decision making, and efficient data-dependent sample sizes and randomization schemes. On the other hand, it also points out potential challenges in using the Bayesian approach, which requires extensive pre-planning and model building besides specific computational and statistical expertise, and which may have substantial Type I error inflation. Its Section VII.D says, "Using simulations to demonstrate control of the Type I error rate, however, is controversial and not fully understood." It also points

out "simulation bias" that can arise when simulations are set up by the modeling group who can decide on the choice of simulation scenarios to mask the advantages of competing designs, and on the parameter configurations in the null hypothesis to mask Type I error inflation.

Two years after the FDA Draft Guidance on Adaptive Design, the President's Council of Advisors on Science and Technology (PCAST) issued a report on "Propelling Innovation in Drug Discovery, Development, and Evaluation" in September 2012. The report points out that clinical trials constitute the largest single component (representing nearly 40%) of the R&D budget of major biopharmaceutical companies and yet are inefficient in terms of cost, time, organization, and delivery of evidence supporting or against the new medical product. It says that time is ripe for improving the efficiency of clinical trials because "it is increasingly possible to obtain clear answers with many fewer patients and with less time" by focusing studies on "specific subsets of patients most likely to benefit, identified based on validated biomarkers." It also says:

> Another approach is to use innovative new approaches for trial design that can provide more information more quickly. Bayesian statistical designs potentially allow for smaller trials with patients receiving, on average, better treatments. These and other modern statistical designs can improve on current protocols, which have only a very limited ability to explore multiple factors simultaneously. Such factors importantly include individual patient responses to a drug, the effects of simultaneous multiple treatment interventions, and the diversity of biomarkers and disease sub-types.

It refers to [27] and cites the I-SPY trials as examples of "exciting innovative models for clinical trials." It also recommends that the FDA "run pilot projects to explore adaptive approval mechanisms to generate evidence across the life cycle of a drug from pre-market through the post-market phase."

3.2. Response to the FDA Draft Guidance and industry's perspectives

After the 3-month public comment period following the FDA Draft Guidance, a special issue on adaptive designs of clinical trials appeared in the *Journal of Biopharmaceutical Statistics*. The papers [36, 37, 38, 39, 40] in this special issue are viewpoints from biostatisticians from the pharmaceutical industry on the Draft Guidance, while [41, 42, 43] represent those from academia. In addition, [44, 45] summarize the highlights and a panel discussion in the

Basel Conference on Perspectives on the Use of Adaptive Designs in Clinical Trials (March 12, 2010). The editorial [46] to this special issue says:

> The overwhelming interests in adaptive designs in lieu of group sequential designs will generate more challenges and invite more controversies. Different from the fixed design and the group sequential design, the adaptive design not only provides the opportunities to change or select from the initial null hypotheses, but also gives the flexibility to modify the maximum statistical information prospectively planned, which can alter the alternative hypothesis of ultimate interest. The main challenge of unblinding in group sequential designs has been extended to adaptive designs due to the potential of inviting changes in the future course of the remaining trial or the related trials that are either underway or ongoing. The critical concerns, which need to be scrutinized, are the likely implications due to the unblinded data-dependent changes, which are prone to operationally and statistically induce the bias. Use of the DMC or DSMB solely for interim adaptive monitoring and adaptive recommendation in addition to safety monitoring with an adaptive design, though it has been proposed, is under scrutiny for its ability to be completely objective. Meanwhile, other trial logistics models, e.g., combination of the DMC and an independent statistical analysis center, have also been proposed.

A follow-up note [46] by the Editor of the journal says:

> In practice, while we enjoy the flexibility of the adaptive trial designs, the quality, integrity, and validity of the trial may be at a greater risk, especially for those less-well-understood designs as described in the FDA draft guidance. From a regulatory perspective, there is always concern about whether the p-value or confidence interval regarding the treatment effect under an adaptive trial design is reliable or correct. In addition, the misuse or abuse of adaptive design methods in a clinical trial may lead to a totally different trial that is unable to address scientific/medical questions that the trial is intended to answer.

The paper [36] also summaries the work of a PhRMA (Pharmaceutical Research and Manufacturers of America) Working Group on adaptive clinical trial designs prior to the FDA Draft Guidance. It agrees with the Draft Guidance that "the number of aspects of a trial that are subject to adaptation should be quite limited." On the other hand, it argues that seamless Phase II-III designs, for which the Draft Guidance says that "these terms provide

no additional meaning beyond the term *adaptive*", have great advantages in certain studies and "deserve further emphasis" in the Guidance. Concerning unblinding issues in an adaptive trial, it says:

> The PhRMA group had proposed in the Executive Summary the possibility of limited and carefully controlled sponsor involvement in the decision and adaptation processes, in situations where that perspective was felt to be needed for the decision, but always totally insulated from trial personnel. This required clear justification of the need and purpose of the sponsor involvement; "minimal" access to information, in terms of the number of individuals involved, the times they would receive information, and the amount of information they would receive; total separation of these personnel from trial operations; and strong firewalls and documentation of processes. This would clearly not be a "one-size-fits-all" model, and would be highly dependent on case-by-case details. ... The following points are made clear to sponsors: Interim results must remain inaccessible to personnel involved in trial conduct; standard operating procedures and charters will need to be more detailed and specific than in familiar monitoring settings; detailed descriptions will be required as to how the analysis will be performed, who will have access to the results, and under what conditions; it will be prospectively described and retrospectively documented how compliance with the specified procedures will be monitored.

The papers [37], [38] and [39] provide further discussions on blinding and operational bias, together with other aspects of the Draft Guidance. Liu and Chi [40] give an insightful discussion of the history of regulatory research, legal basis, bias and blinding in connection with the Draft Guidance. In addition, they also raise a number of statistical issues with widely accepted frequentist methods in group sequential and adaptive clinical trials. In particular, they cite the critiques of Cox [48] and Armitage [49] on error spending, conditional power, and stochastic curtailment.

Cook and DeMets [41] commented that the Draft Guidelines "are extremely useful and should provide both industry and academia valuable information concerning regulatory perspective on the design, conduct, and analyses of adaptive clinical trials." They also note that "in exchange for potential efficiencies in resource utilization, adaptive trials suffer from limitations in scientific conclusions, complications and inefficiencies in the statistical analysis, and logistical difficulties relative to fixed sample or fixed

duration trials." Emerson and Fleming [42] discuss "the extent to which the adaptive designs do not meet the goals of having greater efficiency, being more likely to identify truly effective treatments, being more informative, and providing greater flexibility," and support the FDA's requirement of "adequate and well-controlled confirmatory studies, complete with prospective, detailed specification of the entire randomized clinical trial design in a way that allows accurate and precise estimation of treatment effectiveness." Cheng and Chow [43] highlight the Draft Guidance's distinction between well-understood and less well-understood adaptive designs, and further classify the less well-understood adaptive designs into "flexible" and "wildly flexible" ones and recommend the latter not be used.

4. Towards flexible and efficient adaptive designs satisfying regulatory requirements

4.1. Efficiency, flexibility, and validity via mainstream statistical methods

Mainstream statistical methods form the core of graduate programs in Statistics and have well-established theories, efficiency properties, and implementation details/software. Bayesian methods are clearly part of this core, but so are parametric, semiparametric, and nonparametric (empirical) likelihood methods. While Bayesian inference is based on the posterior distribution, likelihood inference is based on the likelihood function. For large sample sizes and under mild regularity conditions, the Bayesian and likelihood approaches yield asymptotically equivalent estimates and tests. The argument of the Bayesian camp that the Bayesian approach is the only efficient way to predict outcomes at the end of the trial given the data at interim analysis ignores the fact that adaptive predictors which replace the unknown parameters by sequential maximum likelihood estimates have also been found to be asymptotically optimal in time series and control systems [50, 51, 52]. Note that time series analysis and forecasting is another core in mainstream Statistics, and likelihood methods are workhorses of this core, as is Bayesian methodology.

Closely related to time series analysis is sequential analysis and dynamic stochastic optimization, which form another core in mainstream Statistics. Applying efficient test statistics is only using half of the toolbox for building an efficient adaptive design, and the other half consists of dynamic stochastic optimization. In fact, the three-stage design proposed in [17, 18] for sample size re-estimation as described in Section 2.1 was derived as an approximate

solution to the associated optimal stopping problem. It turns out that the complicated Bayesian designs proposed in the literature are sub-optimal because they are myopic in the sense of minimizing the current posterior loss rather than the sum of current and future posterior losses, whereas the optimal solutions can be well approximated by relatively simple frequentist designs such as those in [17, 18]. Another example can be found in the classical multi-armed bandit problem, which addresses the dilemma between "exploration" (to generate information about unknown system parameters) and "exploitation" (to set system inputs in order to maximize expected rewards from the outputs).

Suppose there are K treatments of unknown efficacy to be chosen sequentially to treat a large class of n patients. How should we allocate the treatment to maximize the mean treatment effect? Lai and Robbins [53] and Lai [54] consider the problem in the setting in which the treatment effect has a density function $f(x;\theta_k)$ for the kth treatment, where θ_k are unknown parameters. There is an apparent dilemma between the need to learn the unknown parameters and the objective of allocating patients to the best treatment to maximize the total treatment effect $S_n = X_1 + + X_n$ for the n patients. If θ_k were known, then the optimal rule would use the treatment with parameter $\theta^* = \arg\max_{1\le k\le K} \mu(\theta_k)$, where $\mu(\theta) = E_\theta(X)$. In ignorance of θ_k, Lai and Robbins [53] define the *regret* of an allocation rule by

$$R_n(\theta) = n\mu(\theta^*) - E_\theta(S_n) = \sum_{k\mu(\theta_k)<\mu(\theta^*)} (\mu(\theta^*) - \mu(\theta_k))E_\theta T_n(k),$$

where $T_n(k)$ is the number of patients receiving treatment k. They show that adaptive allocation rules can be constructed to attain the asymptotically minimal order of $\log n$ for the regret, in contrast to the regret of order n for the traditional equal randomization rule that assigns patients to each treatment with equal probability $1/K$. A subsequent refinement by Lai [54] shows the relatively simple rule that chooses the treatment with the largest upper confidence bound $U_k^{(n)}$ for θ_k to be asymptotically optimal if the upper confidence bound at stage n, with $n > k$, is defined by

$$U_k^{(n)} = \inf\left\{\theta \in A:\ \theta \ge \widehat{\theta}_k \text{ and } 2T_n(k)I(\widehat{\theta}_k,\theta) \ge h^2(T_n(k)/n)\right\}, \quad \inf\emptyset = \infty,$$

where A is some open interval known to contain θ, $\widehat{\theta}_k$ is the maximum likelihood estimate of θ_k, $I(\theta,\lambda)$ is the KullbackLeibler information number, and

the function h has a closed-form approximation. It is noted in [55] that the upper confidence bound $U_k^{(n)}$ corresponds to inverting a generalized likelihood ratio (GLR) test based on the GLR statistic $T_n(k)I(\widehat{\theta}_k, \theta)$ for testing $\theta_k = \theta$.

The multi-arm bandit problem has the same "learn-as-we-go" spirit of the BATTLE trial described in Section 2.3. Making use of these ideas, Lai, Liao and Kim [55] have recently introduced a frequentist alternative to the Bayesian adaptive design of the BATTLE trial. While the spirit of the BATTLE trial focuses on attaining the best response rate for patients in the trial, it does not establish which treatment is the best for future patients, with a guaranteed probability of correct selection. A group sequential design is introduced in [55] for jointly developing and testing treatment recommendations for biomarker classes, while using multi-armed bandit ideas to provide sequentially optimizing treatments to patients in the trial. Thus, the design has to fulfill multiple objectives, which include (a) treating accrued patients with the best (yet unknown) available treatment, (b) developing a treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has better treatment effect than the historical mean effect of standard of care plus a predetermined threshold. Because of the need for informed consent, the treatment allocation that uses the upper confidence bound rule for multi-arm bandits is no longer appropriate. It is unlikely for patients to consent to being assigned to a seemingly inferior treatment for the sake of collecting more information to ensure that it is significantly inferior (as measured by the upper confidence bounds). Instead, randomization in a double blind setting is required, and the randomization probability $\pi_{jk}^{(i)}$, determined at the ith interim analysis, of assigning a patient in group j to treatment k cannot be too small to suggest obvious inferiority of the treatments being tried, that is, $\pi_{jk}^{(i)} \geq \epsilon$ for some $0 < \epsilon < 1/K$. The unknown mean treatment effect μ_{jk} of treatment k in biomarker class j can be estimated by the sample mean $\hat{\mu}_{ijk}$ at interim analysis i. Let $k_j = \arg\max_k \mu_{jk}$, which can be estimated by $\hat{k}_{ij} = \arg\max_k \hat{\mu}_{ijk}$ at the ith interim analysis. Multi-arm bandit theory suggests assigning the highest randomization probability to treatment $\hat{k}_{ij}$ and randomizing to the other available treatments in biomarker class j with probability ϵ. This adaptive randomization scheme is called "ϵ-greedy" in the machine learning literature and is much simpler than the Thompson-type adaptive randomization scheme based on posterior probabilities. This frequentist design is shown to be asymptotically opti-

mal in [55], which also carries out simulation studies that show its superior finite-sample performance.

Biomarker-guided personalized therapies studied in the BATTLE trial are an example of personalized medicine. Personalization in medical treatments and in web-based recommender systems and electronic marketing belongs to the emerging area of contextual bandit theory in sequential analysis and experimentation of "big data". Contextual multi-armed bandits, also called multi-armed bandits with side (covariate) information, provide a mathematical framework for this stochastic dynamic optimization problem. Whereas the BATTLE trial assumes that the cut-points defining the biomarker classes are available and thereby reduce treatment allocation for each class to a multi-armed bandit problem, contextual bandit theory allows learning these cut-points (or more general regression functions of treatment response on the covariates). To illustrate with an example, the stochastic optimization problem of web-based personalization in showing online ads for each user, with the goal of maximizing its effectiveness, measured in terms of click-through rate or total revenue, has been formulated as a contextual multi-armed bandit problem with the page request of each user as side (covariate) information and layouts of ads available for the requested page as arms. We have recently solved the dynamic stochastic optimization problem associated with contextual bandits, and adaptive randomization of the type in [55] together with arm elimination again plays a key role in our solution which is presented elsewhere.

The comments of Liu and Chi [40, Sections 7.4, 7.5, 8.1, 8.4] on the plethora and controversies of *ad hoc* adaptation methods in the burgeoning literature on adaptive clinical trials demonstrate the importance of combining the principles and methods from different cores of mainstream Statistics to address the inherently complex and difficult problems in adaptive design of clinical trials. In particular, consider their comments on Tsiatis and Mehta's claim [11] on the inefficiency of the sample size re-estimation designs in the second paragraph of Section 2.1. They say that the claim is "logically flawed at the fundamental level" because [11] does not consider expected sample size properties and only focuses on power at a pre-specified alternative for all tests with the same type I error and error-spending function at a simple null. In the terminology of this section, [11] only dwells on the first half of the toolbox for building an efficient adaptive design but does not consider the second half of the toolbox, opening up the possibility of a flawed overall design that [40] discusses. The second half of the toolbox on dynamic stochastic optimiza-

tion is arguably much more difficult than the first half. In principle, this can be formulated as a Bayesian sequential decision problem that can be solved by dynamic programming. Specifically, given a prior distribution of the unknown parameters, one can formulate the dynamic stochastic optimization as a dynamic programming problem in which the state at time t is the posterior distribution of the parameters given the observations up to t. However, the dynamic programming equations are often prohibitively difficult to handle, both computationally and analytically. Moreover, it may also be difficult to specify a reasonable prior distribution. Chapter 3 of [56] gives an overview of stochastic optimization over time, dynamic programming and approximate dynamic programming, together with their applications to Phase I cancer trial designs and sequential tests of composite hypotheses. In particular, for the sequential testing application, analytic approximations to the Bayes rules are developed by using Laplace's asymptotic formula to relate the posterior distributions to GLR statistics and large or moderate deviations approximations to boundary crossing probabilities. A review of the multi-arm bandit problem is given in [57], which explains the suboptimal nature of the myopic rule and demonstrates that the rules in [53, 54] achieve asymptotic efficiency by introducing relatively simple uncertainty adjustments to the myopic rule and thereby attaining certain lower bounds on the expected sample size from the inferior arms for "uniformly good" rules.

Developing information-theoretic asymptotic lower bounds such as those in [53, 54] and finding procedures to attain them bypass the difficulties of dynamic programming but require deep insights and synthesis of several mainstream cores of Statistics. This approach has proved useful in more complicated design problems than those reviewed in Section 2. In particular, it was recently used in [58] for the problem of adaptive choice of patient subgroup for comparing two treatments, motivated by a clinical trial design problem posed to us by colleagues of the Stanford Stroke Center that will be explained in Section 4.3. Adaptive (data-dependent) choice of the patient subgroup to compare the new and control treatments is a natural compromise between ignoring patient heterogeneity and using stringent inclusion-exclusion criteria in the trial design and analysis. Section 2 of [58] first provides an asymptotic theory for trials with fixed sample size, in which n patients are randomized to the new and control treatments and the responses are normally distributed, with mean μ_j for the new treatment and μ_{0j} for the control treatment if the patient falls in a pre-defined subgroup Π_j for $j = 1, \ldots, J$, and with common known variance σ^2. Let Π_J denote the entire patient population for a tra-

ditional randomized controlled trial (RCT) comparing the two treatments. Since there is typically little information from previous studies about the subgroup effect size $\mu_j - \mu_{0j}$ for $j \neq J$, [58] begins with a standard RCT to compare the new treatment with the control over the entire population, but allows adaptive choice of the patient subgroup $\hat{I}$, in the event H_J is not rejected, to continue testing $H_i : \mu_i \leq \mu_{0i}$ with $i = \hat{I}$ so that the new treatment can be claimed to be better than control for the patient subgroup $\hat{I}$ if $H_{\hat{I}}$ is rejected.

Letting $\theta_j = \mu_j - \mu_{0j}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$, the probability of a false claim is the type I error

$$\alpha(\boldsymbol{\theta}) = \begin{cases} P_{\boldsymbol{\theta}}(\text{reject } H_J) + P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \leq 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}) & \text{if } \theta_J \leq 0 \\ P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \leq 0, \text{ accept } H_J \text{ and Reject } H_{\hat{I}}) & \text{if } \theta_J > 0, \end{cases}$$

for $\boldsymbol{\theta} \in \Theta_0$. Subject to the constraint $\alpha(\boldsymbol{\theta}) \leq \alpha$, [58, Appendix A] establishes the asymptotic efficiency of the procedure that randomly assigns n patients to the experimental treatment and the control, rejects H_J if $\text{GLR}_i \geq c_\alpha$ for $i = J$, and otherwise chooses the patient subgroup $\hat{I} \neq J$ with the largest value of the generalized likelihood ratio statistic

$$\text{GLR}_i = \{n_i n_{0i}/(n_i + n_{0i})\}(\hat{\mu}_i - \hat{\mu}_{0i})_+^2/\sigma^2$$

among all subgroups $i \neq J$ and rejects $H_{\hat{I}}$ if $\text{GLR}_{\hat{I}} \geq c_\alpha$, where $\hat{\mu}_i(\hat{\mu}_{0i})$ is the mean response of patients in Π_i from the treatment (control) arm and $n_i(n_{0i})$ is the corresponding sample size. The test statistic GLR_i is the sample estimate of the Kullback-Leibler information $(np_i/4)(\mu_i - \mu_{0i})_+^2/\sigma^2$, noting that $n_i n_{0i}/(n_i + n_{0i}) \approx np_i$ as study subjects are equally likely to receive the new treatment or control. After establishing the asymptotic efficiency of the procedure in the fixed sample size case, [58] proceeds to extend it to a 3-stage sequential design by making use of the theory of Bartroff and Lai [17, 18] reviewed in Section 2.1. It then extends the theory from the normal setting to asymptotically normal test statistics, such as the Wilcoxon rank sum statistics that are commonly used for analyzing the clinical endpoints (Rankin scores) of stroke patients. A modification of this argument, details of which are given elsewhere, can be used to derive asymptotically efficient seamless Phase II-III designs reviewed in Section 2.2, for which the hypothesis H_j now corresponds to the jth dose or treatment regimen of the new drug.

The discussion in this section up to now has focused on the efficiency and flexibility aspects in the title, and we have highlighted how different cores

of mainstream Statistics have provided concepts and techniques to address these aspects. We now move to "validity" in the title, which is related to satisfying regulatory requirements for the trial design and analysis. The "frequentist twist" [27, p.33] to modify a Bayesian adaptive design, as described in the last paragraph of Section 2.3, still falls short of FDA's requirement of valid frequentist false positive rate, for all parameter configurations in the null hypothesis, that we have referred to the second paragraph of Section 3.1. The approximation of dynamic Bayes rules by sequential procedures based on GLRs, which is called "pseudo-maximization" in [59, p.240], has an important bonus that makes it particularly suited to frequentist inference, namely that GLR is an approximate pivot [59, pp.207, 216] under a general null hypothesis, which ensures that the distribution of the GLR statistic is approximately the same irrespective of the parameter configuration in the null hypothesis. As explained in [59, Chapter 16], using GLR or general Studentized statistics in bootstrapping is important for the second-order accuracy of the bootstrap method because they are asymptotically pivotal. Note that the bootstrap and other resampling methods form another core in mainstream Statistics. For group sequential or adaptive designs, the approximate pivotal property of Efron's bootstrap [60, 61] breaks down, but a more general resampling scheme called *hybrid resampling* can be used as the sequential GLR or Studentized statistics are still approximately pivotal under hybrid resampling; see [62, 63, 64].

4.2. Hybrid resampling, survival outcomes and more complicated settings

The results of the BATTLE trial, whose Bayesian adaptive design has been reviewed in Section 2.3, are reported by Kim et al. [65]. Despite applying Bayesian adaptive randomization and arm elimination, "standard statistical methods included Fisher's exact test for contingency tables and log-rank test for survival data", and standard confidence intervals and p-values based on normal approximations without adjustments for Bayesian adaptive randomization and possible treatment suspension. This paper shows the dilemma faced by the clinical investigators who bought into Bayesian design but had to carry out frequentist inference such as p-values and confidence intervals to publish their results in medical journals. What previous research to attain frequentist validity for regulatory approval seemed to have missed is that hybrid resampling already provides a versatile and accurate method for valid frequentist inference in Bayesian and other adaptive designs. One may argue, however, that the MCMC computations of posterior probabilities may

be too computationally intensive to carry out hybrid resampling. On the other hand, the code for performing the frequentist operating characteristics in the frequentist twist [27] can be readily modified to carry out hybrid resampling. Furthermore, as noted in the preceding section, the Thompson-type adaptive randomization scheme based on posterior probabilities in the BATTLE design is suboptimal, and a much simpler and yet asymptotically optimal adaptive sampling scheme based on the truncated MLE $\hat{p}_{jk}^{(t)}$ can be used instead. Hybrid resampling is especially suited for primary and secondary analysis, in a regulatory environment, of complex data in adaptive clinical trials. A detailed discussion is given in [66] for the case of censored survival data, the complexity of which has led to inefficient adaptive designs, as reviewed in the second paragraph of Section 2.2. A new approach is also developed in [66] that can adapt the choice of the test statistics to the observed survival pattern.

4.3. An illustrative example on adaptive subgroup selection

In 2014 our colleagues at the Stanford Stroke Center came to us with a request to help them design a clinical trial evaluating a new method for augmenting usual medical care with endovascular removal of the clot after a stroke, resulting in reperfusion of the area of the brain under threat, with the intent of salvaging tissue and improving outcomes, compared to standard medical care with intravenous tissue plasminogen activator (tPA) alone. The investigators were also interested in tailoring the reference population to optimize the power to detect a significant effect. In the course of discussions it emerged that there were a nested sequence of six subsets of patients, defined by a combination of elapsed time from stroke to start of tPA and an imaging-based estimate of the size of the unsalvageable core region of the lesion. The sequence was defined by cumulating the cells in a two-way (3 volumes $\times$ 2 times) cross-tabulation as described in [58, p.195]. In the upper left cell, c_{11}, which consisted of the patients with a shorter time to treatment and smallest core volume, the investigators were most confident of a positive effect, while in the lower right cell c_{23} with the longer time and largest core area, there was less confidence in the effect. The six cumulated groups, $\Pi_1, \ldots, \Pi_6$ give rise to corresponding one-sided null hypotheses, $H_1, \ldots, H_6$ for the treatment effects in the cumulated groups. This is the trial that motivated the problem of adaptive choice of patient subgroup for comparing two treatments discussed in the penultimate paragraph of Section 4.1.

Shortly before the final reviews of the protocol for funding were completed, four randomized controlled trials of endovascular reperfusion therapy administered to stroke patients within 6 hours after symptom onset demonstrated decisive clinical benefits [67, 68, 69]. As a result, the equipoise of the investigators shifted, making it necessary to adjust the intake criteria to exclude patients for whom the new therapy had been proven to work better than the standard treatment. The subset selection strategy became even more central to the design, since the primary question was no longer whether the treatment was effective at all, but for which patients should it be adopted as the new standard of care. Moreover, besides adapting the intake criteria to the new findings, another constraint was imposed by the NIH sponsor, which effectively limited the total randomization to 476 patients. Recall that in the original design, if H_J was accepted at an interim stage, the study would go on to recruit to the maximum sample size *in the selected subgroup*. The limit on total randomization is an example of a constraint on adaptive design that reflects more conservative budgeting at the NIH after positive results are reported by other studies with more restrictive inclusion criteria.

We redefine the design as in [58] using the maximum randomization limit, but with the futility stopping rule based on GLR testing of the one-sided null at the alternative implied by the maximum sample size, which is now a random variable since it depends on whether H_J is rejected at an interim analysis. We estimate the expected value of the maximum sample size by simulation under the null, then recalculate the implied alternative (which becomes larger since the expected maximum is smaller) and replace it in the design. The operating characteristics of the adjusted design are simulated under various scenarios shown in Table 1, in which each result is based on 5000 simulations. The projected overall effect of endovascular therapy is based on (a) the observed 90-day outcomes in an earlier study of similar patients treated > 6 hrs after symptom onset and (b) the assumption that early reperfusion will be achieved in 75% of the endovascular arm vs. 20% of the medical therapy arm. Using these projections, we calculate an expected standardized effect size of 0.36 (normal approximation to the Wilcoxon statistic comparing the modified Rankin scores across treatments). If this effect is uniform in all 6 cells, the fixed sample size, non-adaptive design requires 376 patients per group (overall), to have 90% power at a two-sided alpha of 5% (Wilcoxon test);100 additional patients are added for the adaptive design, for a maximum randomization of 476.

Table 1 compares the performance of a traditional fixed sample-size design

Table 1: Operating characteristics of adaptive design and fixed-sample conventional design.

Scenario	Effect (standardized) in cells						Avg.	Adaptive		Conventional	
	c_{11}	c_{12}	c_{21}	c_{22}	c_{31}	c_{32}	effect	Avg. #	Power	#	Power
#0	0	0	0	0	0	0	0	361	2.2%	476	2.5%
#1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	354	80%	476	89%
#2	0.5	0.4	0.3	0	0	0	0.2	400	86%	476	55%
#3	0.5	0.5	0	0	0	0	0.17	403	87%	476	41%

(fixed $n = 476$) for testing H_J to the adaptive design (random sample size $n \leq 476$) that allows choice of H_j ($j \neq J$) if H_J is accepted. The effect sizes in the table are standardized. Under the null hypothesis of no treatment effect (Scenario #0), the adaptive design controls the total Type 1 error below 2.5%, stops early for futility 63% of the time, and the mean number of patients randomized (Avg. #) is 361. If the effect is uniform across cells (scenario #1), the fixed-sample design is optimal, but the adaptive design results in only a small loss of power (from 89 to 80%). The adaptive design performs much better (higher power and smaller expected sample size) than the fixed-sample, conventional trial when the effect size distribution across the subgroups is in accord with the biological assumptions (scenarios #2 and 3). If the effect is concentrated in two cells with small core volumes (scenario #3), the adaptive design maintains 87% power while the conventional design collapses (41% power). The adaptive design also performs well compared to a non-adaptive, fixed-sample design in [58] that allows adaptive subgroup choice at the end of the study.

5. Discussion

The FDA document [70] on the critical path to new medical products points out that today's advances in the biomedical sciences offer new possibilities to treat many diseases but that the number of new drug and biologic applications submitted to the FDA has been declining. The PCAST report described in the last paragraph of Section 3.1 argues that innovations in clinical trial design and improvements in efficiency and cost effectiveness of clinical trials are needed for biomedical advances to reach their full potential in treating diseases. Although it is widely recognized that adaptive designs represent promising innovations that can reverse the "stagnation"

trend pointed out in [70], the overview of the state of the art in adaptive designs in Sections 2 and 3 shows that despite the advances reviewed therein, major hurdles still remain and need to be resolved before adaptive designs can meet the desiderata of efficiency, flexibility, and validity from a regulatory perspective. Section 4 describes our recent work at the Stanford Center for Innovative Study Design to address these challenges. In particular, we show how the methodological problems underlying these hurdles are related to other branches of mainstream Statistics, from which we can synthesize and further develop the methods and results to yield (a) efficient and flexible adaptive designs of confirmatory clinical trials and (b) valid statistical analyses of the data that can meet regulatory requirements. Berry [71] has given a cogent summary of the advantages and disadvantages ("promise" and "caution") of adaptive designs. He points out that their potential advantages are greatest in complicated settings and mentions I-SPY2 as an "example of a complicated trial made possible by adaptive design," in which "the adaptive randomization provides information about which drugs benefit which patients" and "trial participants receive better treatment." Since he advocates using the Bayesian approach that analyzes the data via posterior distributions, the increased complexity of a trial only increases the complexity of the MCMC simulations. On the other hand, this increased complexity would make it even harder for the "frequentist twist" mentioned in Section 4.1 to be able to satisfy the regulatory requirements. However, it should be noted that Section 4.1 and 4.2 have developed methods which have frequentist validity while sharing the efficiency and flexibility of Bayesian adaptive designs. We are therefore in agreement with Berry that adaptive designs have greatest advantages over the traditional designs in ambitious and complicated trials that translate today's advances in the biomedical sciences into effective treatments of complex diseases.

An important concern with adaptive designs mentioned by Berry [71] is "the issue of trends in patient populations during a period of time." We have illustrated in Section 4.3 how this concern can be handled at interim analyses. One can clearly adapt to these changes in an adaptive design, as sequential change-point detection and estimation methods form another core of mainstream Statistics. This issue and the related statistical methodologies will be treated elsewhere. Another major concern that has also been mentioned in the 2010 FDA Draft Guidance and in the review in Section 3.2 is about blinding and operational bias. Berry [71] says, "Modifications (of the trial) that occur during the trial may convey information outside the sphere of con-

fidentiality of the DSMB and affect the types of patients who are accrued to the trial." On the other hand, "all DSMB reports of interim analyses convey some information about the relative performance of the arms of the trials." We propose to keep whatever is disclosed to the DSMB within its sphere of confidentiality. Blinding should be maintained outside the DSMB throughout the trial, and the protocol's adaptation rule should also be kept confidential by the DSMB. We can regard the adaptation that is built into the protocol as an algorithm that is executed by a computer analogous to algorithmic trading in financial markets with electronic platforms. Since the computer technology is already available to carry out automated high-frequency trading, the same should be true for implementing adaptive clinical trials. If the design has been well thought through in advance and its performance characteristics are well understood, there is no need for the sponsor of the trial to participate in the adaptive decisions at interim analyses, similarly to how trades are executed by computers without human intervention in algorithmic trading.

The principles and methods discussed herein for adaptive design of Phase III trials can be extended to adaptation in clinical trial plans (CDPs). As noted in [72], in the development of a new drug in the pharmaceutical industry, 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, and there is a core team that makes decisions guided by a 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 or termination of further development. 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. At the planning stage of a CDP, 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 standard clinical trials designs, which are well understood and relatively simple to present to senior management, are good ways to start. Because of the uncertainties and the strong assumptions underlying the "simplified" CDP [72] proposes to modify it by using adaptive designs in place of standard designs in the simplified CDP, which is the essence of the adaptive CDP developed therein.

References

[1] Bauer P. Multistage testing with adaptive designs (with Discussion). Biom Inform Med Bio 1989; 20:130–148.

[2] Wittes J, Brittain E. The role of internal pilots in increasing the efficiency of clinical trials. Stat Med 1990; 9:65–72.

[3] Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat 1945; 16:243–258.

[4] Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Comm Stat Ser A 1992; 21:2833–2853.

[5] Herson J, Wittes J. The use of interim analysis in sample size adjustment. Drug Inform J 1993; 27:753–760.

[6] Birkett M, Day S. Internal pilot studies for estimating sample size. Stat Med 1994: 13:2455–2463.

[7] Denne JS, Jennison C. A group sequential t-test with updating of sample size. Biometrika 2000; 87:125–134.

[8] Fisher L. Self-designing clinical trials. Stat Med 1998; 17:1551–1562.

[9] Denne S. Sample size recalculation using conditional power. Stat Med 2001; 20:2645–2660.

[10] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Stat Med 2003; 22:971–993.

[11] Tsiatis AA, Mehta C. On the efficiency of the adaptive design for monitoring clinical trials. Biometrika 2003; 90:367–378.

[12] Proschan M, Hunsberger S. Designed extension studies based on conditional power. Biometrics 1995; 51:1315–1324.

[13] Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses Biometrics 1994; 50:1029-1041

[14] Posch M, Bauer P. Adaptive two stage designs and the conditional error function. Biom J 1999; 41:689-696.

[15] Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. Biometrika 2006; 93:1-21.

[16] Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. Stat Med 2006; 25:917-932.

[17] Bartroff J, Lai TL. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Stat Med 2008; 27:1593–1611.

[18] Bartroff J, Lai TL. Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. Sequent Anal 2008; 27:254-276.

[19] Lorden G. Asymptotic efficiency of three-stage hypothesis tests. Ann Stat 1983; 11:129-140.

[20] Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biom J 2006; 4:623–634.

[21] Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. Biom J 2006; 4:635–643.

[22] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73:751–754.

[23] Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. Stat Med 2003; 22:689–703.

[24] Wang Y, Lan KKG, Ouyang SP. A group sequential procedure for interim treatment selection. Stat Biopharm Res 2011: 3:1–13

[25] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Stat Med 2009; 28:1445–1463.

[26] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm Stat 2011; 10:347-356.

[27] Berry DA. Bayesian clinical trials. Nature Rev Drug Disc 2006; 5:27–36.

[28] Berry SM, Carlin BP, Lee JJ, Müller P. Bayesian Adaptive Methods for Clinical Trials. Boca Raton FL: Chapman & Hall/CRC; 2011.

[29] Thompson W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 1933; 25:285–294.

[30] Meuer WJ, Lewis RJ, Berry DA. Adaptive clinical trials: A partial remedy for the therapeutic misconception? J Am Med Assoc 2012; 307:2377–2378.

[31] Zhou X, Liu S, Kim ES, Herbst RS, Lee JL. Bayesian adaptive design for targeted therapy development in lung cancer–A step toward personalized medicine. Clin Trials 2008; 5:181–193.

[32] Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther 2009; 86:97–100.

[33] Berry DA et al. Adjuvant chemotherapy in older women with early-stage breast cancer. N Engl J Med 2009; 360:2055–2065.

[34] Food and Drug Administration Center for Drug Evaluation and Re-search. Guidelines for Industry: Adaptive Design Clinical Tri-als for Drugs and Biologics, 2010. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf`.

[35] Committee for Medicinal Products for Human Use Reection paper on methodological issues in conrmatory clinical trials planned with an adaptive design. Eur Med Agency, 2007. http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf.

[36] Gallo P, Anderson K, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Viewpoints on the FDA draft adaptive designs guidance from the PhRMA working group. J Biopharm Stat 2010; 20: 1115–1124.

[37] Brannath W, Burger HU, Glimm E, Stallard N, Vandemeulebroecke M, Wassmer G. Comments on the Draft Guidance on "Adaptive design clinical trials for drugs and biologics" of the U.S. Food and Drug Administration. J Biopharm Stat 2010; 20: 1125–1131.

[38] Chuang-Stein C, Beltangady M. FDA Draft Guidance on adaptive design clinical trials: Pfizer's perspective. J Biopharm Stat 2010;20: 1143–1149.

[39] Wittes J. Comments on the FDA Draft Guidance on adaptive designs. J Biopharm Stat 2010; 20: 1166–1170.

[40] Liu Q, Chi YH. Understanding the FDA Guidance on adaptive designs: historical, legal, and statistical perspectives. J Biopharm Stat 2010; 20: 1178–1219.

[41] Cook T, DeMets DL. Review of draft FDA adaptive design guidance. J Biopharm Stat 2010; 20: 1132–1142.

[42] Emerson SS, Fleming TR. Adaptive methods: telling "the rest of the story". J Biopharm Stat 2010; 20: 1150–1165.

[43] Cheng B, Chow SC. On flexibility of adaptive designs and criteria for choosing a good one—a discussion of FDA Draft Guidance. J Biopharm Stat 2010; 20: 1171–1177.

[44] Wang SJ. Perspectives on the use of adaptive designs in clinical trials. Part I. Statistical considerations and issues. J Biopharm Stat 2010; 20: 1090–1097.

[45] Benda,N, Brannath W, Bretz F, Burger HU, Friede T, Maurer W, Wang SU. Perspectives on the use of adaptive designs in clinical trials. Part II. Panel discussion. J Biopharm Stat 2010; 20: 1098–1112.

[46] Wang SJ. Editorial: Adaptive designs: Appealing in development of therapeutics, and where do controversies lie? J Biopharm Stat 2010; 20: 1083–1087.

[47] Chow S-C. A note on special articles on adaptive clinical trial designs. J Biopharm Stat 2010; 20:1088–1089

[48] Cox DR. Commentary: The likelihood paradigm for statistical evidence by Richard Royall. In: The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations (Tapper ML, Lele SR, eds), 138–140. Univ Chicago Press, 2004.

[49] Armitage P. Discussion of the paper by Jennison and Turnbull. J Roy Stat Soc Ser B 1989; 51:334–335.

[50] Goodwin GC, Sin KS. Adaptive Filtering, Prediction and Control. Mineola NY: Dover; 2009.

[51] Lai TL, Ying Z. Recursive identification and adaptive prediction in linear stochastic systems. SIAM J Contr Optim 1991; 29:1061–1090.

[52] Lai TL, Zhu G. Adaptive prediction in nonlinear autoregressive models and control systems. Stat Sinica 1991; 1:309–334.

[53] Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. Adv Appl Math 1985; 6: 4–22.

[54] Lai TL. Adaptive treatment allocation and the multi-armed bandit problem. Ann Stat 1987; 15: 1091–1114.

[55] Lai TL, Liao OY-W, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. Cont Clin Trials 2013; 36:651–663.

[56] Bartroff J, Lai TL, Shih MC. Sequential Experimentation in Clinical Trials: Design and Analysis. New York: Springer; 2013.

[57] Lai TL. Information bounds, certainty equivalence and learning in asymptotically efficient adaptive control of time-invariant stochastic systems. In Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control (Gerencsèr L, Caines PE, eds), 335–368. New York: Springer; 1991.

[58] Lai TL, Lavori PW, Liao OY. Adaptive choice of patient subgroup for comparing two treatments. Con Clin Trials 2014; 39:191–200.

[59] dela Peña V, Lai TL, Shao QM. Self-normalized process: Limit theory and statistical applications. New York: Springer; 2009.

[60] Efron B. Bootstrap methods: Another look at the jackknife. Ann Stat 1979; 7:1–26.

[61] Efron B. Better bootstrap confidence intervals. J Amer Stat Assoc 1987; 82:171–185.

[62] Chuang CS, Lai TL. Hybrid resampling methods for confidence intervals (with Discussion). Stat Sinica 2000; 10:1–50.

[63] Lai TL, Li W. Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. Biometrika 2006; 93:641–654.

[64] Lai TL, Shih MC, Su Z. Tests and confidence intervals for secondary endpoints in sequential clinical trials. Biometrika 2009; 96:903–915.

[65] Kim et al. The BATTLE trial: Personalizing therapy for lung cancer. Cancer Discovery 2011; 1:44–53.

[66] He P, Lai TL, Zheng S. Design of clinical trials with failure-time endpoints and interim analyses: An update after 15 years. Con Clin Trials 2015; this issue.

[67] Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ et al. A randomized trial of intraarterial treatment for acute ischemic stroke. N Engl J Med 2015; 372: 11–20

[68] Campbell BC, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. N Engl J Med 2015; 372:1009–18.

[69] Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. N Engl J Med 2015; 372:1019-1030.

[70] Food and Drug Administration. Innovative/stagnation: Challenge and opportunity in the critical path to new medical products. FDA report 2004 http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

[71] Berry DA. Adaptive clinical trials: The promise and the caution. J Clin Oncol 2011; 606–609.

[72] Lai TL, Liao OY-W, Zhu RZ. Adaptation in clinical development plans and adaptive clinical trials. Stat. & Its Interface 2012; 5:431–442.