Contemporary Clinical Trials 45 (2015) 93–102

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

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

ARTICLE INFO

Article history:
Received 13 April 2015
Received in revised form 5 June 2015
Accepted 10 June 2015
Available online 14 June 2015

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

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.

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 an 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 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, that have evolved from the seminal papers [1] and [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:$\mu_X = \mu_Y$

* Corresponding author at: Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA.
E-mail address: lait@stanford.edu (T.L. Lai).

http://dx.doi.org/10.1016/j.cct.2015.06.007
1551-7144/

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 $X_1, X_2, \ldots \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \ldots \sim N(\mu_Y, \sigma^2)$. Let $t_{\nu,\alpha}$ denotes 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 $|\overline{X}_{n_1} - \overline{Y}_{n_1}| > t_{2n_0-2,\alpha/2}\sqrt{2s_0^2/n_1}$. Stein's two-stage procedure is modified in [2,4] as follows. Viewing $|\overline{X}_{n_1} - \overline{Y}_{n_1}|/\sqrt{2s_1^2/n_1}$ 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 $n(\sigma^2)$ denotes 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 [5–7] 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 $(0 < r < 1)$, $n^{-1/2}S_1 \sim N(r\delta\sqrt{n}, r)$, 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, $(n\gamma)^{-1/2}S_2 \sim N((1-r)\delta\sqrt{\gamma n}, 1-r)$, where $S_2 = \sum_{i=rn+1}^{n^*}(X_i - Y_i)$ and $n^* = rn + \gamma(1-r)n$ are 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 $P_{\delta=0}(S_1 + S_2 > z_\alpha\sqrt{n} \,|\, S_1 = s_1)$. 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 (Eq. 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 $C(\cdot)$; we use ϕ and Φ to denote the standard normal density and distribution function, respectively. Having observed the first-stage z-statistic Z_1, $H_0{:}\delta = 0$ is rejected in favor of $\delta > 0$ if the second stage z-statistic Z_2 satisfies $Z_2 > \Phi^{-1}(1 - C(Z_1))$. 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 hypothesis 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]