Journal of Biopharmaceutical Statistics, 20: 1171–1177, 2010

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2010.514460

Taylor & Francis
Taylor & Francis Group

ON FLEXIBILITY OF ADAPTIVE DESIGNS AND CRITERIA FOR CHOOSING A GOOD ONE—A DISCUSSION OF FDA DRAFT GUIDANCE

Bin Cheng[1] and Shein-Chung Chow[2]
[1]*Department of Biostatistics, Columbia University, New York, New York, USA*
[2]*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA*

In recent years, the use of adaptive design methods in clinical trials has attracted much attention due to its flexibility in identifying the best clinical benefit of the test treatment under investigation. The flexibility, however, comes at the price of decreasing the accuracy and reliability of the statistical inference drawn. In addition, it is susceptible to abuse. The Food and Drug Administration (FDA) draft guidance justifiably distinguishes between well-understood and less well-understood adaptive designs and suggests the use of the latter with caution. In this discussion paper, we further classify the less well-understood adaptive designs into the categories of flexible and wildly flexible ones and recommend the latter not be used. In addition, we suggest a set of performance characteristics as criteria for choosing a good design from a pool of flexible adaptive designs and group sequential designs.

Key Words: Countability; Decision rules; Efficiency; Flexibility; Parsimony; Validity.

1. INTRODUCTION

The use of adaptive design methods in clinical trials has attracted a lot of attention since the early 2000s. An adaptive design means to offer the investigator the flexibility in identifying any potential (preferably the best) clinical benefit of the test treatment under investigation. The motivation and intention of the use of adaptive design methods in clinical trials are good, but many clinical scientists conceptually misuse or abuse the adaptive design methods in clinical trials. Although the use of adaptive designs (e.g., adaptive randomization) can be traced back to early 1970s, there is no universally agreed definition for adaptive design. The PhRMA Working Group on Adaptive Design published a white paper that gives a formal definition of adaptive design (Gallo et al., 2006). PhRMA defines an adaptive design as a clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial. Similarly, the Food and Drug Administration (FDA)

Address correspondence to Bin Cheng, Department of Biostatistics, Columbia University, New York, NY 10032, USA; E-mail: bc2159@columbia.edu

defines an adaptive design clinical study as a study that includes a *prospectively* planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study. As indicated in the FDA draft guidance, adaptation is a prospectively planned opportunity and changes are made based on analysis of data (usually interim data). However, in real practice, many ad hoc changes are made concurrently during the conduct of the study and/or retrospectively prior to database lock and/or unblinding (Chow and Chang, 2006).

The ramification of definitions reflects the fact that there is a wide spectrum of flexibility in adaptive designs. On one hand, such great flexibility generates excitement in the exploration and usage of adaptive designs. On the other hand, great flexibility may open up the opportunity of substantial statistical and operational bias and serious impairment of trial integrity, which will ultimately undermine the value of adaptive designs. The FDA draft guidance makes an important distinction between well-understood and less well-understood adaptive designs. Such an effort is well justified and welcome. However, for those sponsors who would like to utilize adaptive designs in their adequate and well-controlled (A&WC) studies, such a distinction will not be able to provide further guidance on what kinds of adaptive designs should be considered. In this discussion paper, we complement FDA's effort by making (1) a further distinction among the less well-understood adaptive designs according to their degrees of flexibility and (2) recommendations regarding the avoidance of adaptive designs that are too flexible.

Another practically important issue is to compare an adaptive design and a (conventional) nonadaptive design in terms of their relative advantages and disadvantages. A comprehensive comparison would allow the investigator to choose a *good* design (regardless of whether it is adaptive or not) for achieving the study objectives in a timely and more efficient way. A sensible comparison requires a good set of criteria. The FDA draft guidance provides a detailed and pertinent discussion on statistical issues that should be considered in justifying an adaptive design but it does not provide a clearly stated set of criteria for trial sponsors. In this discussion paper, we propose a set of criteria to assist the trial sponsors in choosing a good design. While by no means claiming our proposed criteria are complete or superior, such criteria would be most helpful to the trial sponsors if agreed and endorsed by the FDA.

The remaining of this discussion paper is organized as follows. In the next section, a classification of the less well-understood adaptive designs based on their flexibility is described and recommendations are made on what types of adaptive designs should be avoided. Section 3 proposes a set of criteria for choosing a good adaptive design for an intended clinical trial. A brief conclusion is given in section 4.

2. FLEXIBILITY OF AN ADAPTIVE DESIGN

The draft guidance distinguishes between well-understood designs and less well-understood designs. Although a formal definition is not given, the draft guidance seems to classify the designs with treatment blinded adaptations as "well-understood designs". According to the draft guidance, these designs will enhance efficiency while limiting risk of introducing bias or impairing interpretability. It seems that any design with treatment unblinded adaptations

should be classified as "less well-understood designs" according to the guidance. While encouraging use of less well-understood adaptive designs for exploratory purposes, the draft guidance has clearly expressed reservation toward their use in A&WC trials.

We agree to the draft guidance that it is important to make such a distinction. However, we notice that the collection of less well-understood adaptive designs is very broad and some designs are much more flexible than others, and intuitively too much flexibility may cause problems in design validity and justification. Therefore, it is suggested that further distinction is necessarily made among the less well-understood designs. To this end, a less well-understood adaptive design is said to be wildly flexible if it allows for uncountably many possible adaptations; otherwise, it is said to be flexible. As an illustration, consider an adaptation design involving sample size modification, treatment deletion and/or addition, endpoint switch, and population modification. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ be a collection of distinct parameters measuring treatment effects that are of potential interest. The parameters in Θ could be based on the same type of outcomes but different treatment arms, as in the adaptive design with first-stage treatment selection considered by Thall et al. (1988), or they could be based on the same outcomes but different targeted populations, as in an adaptive design with second-stage population modification; or could be based on different types of outcomes, as in the adaptive design with endpoint switch at the second stage considered by Liu and Pledger (2005), Bretz et al. (2006), and Jennison and Turnbull (2007). The total number of parameters of potential interest K could be either finite or infinite. Consider an adaptive design with up to $m \geq 2$ stages. Let Θ_j be the subset of parameters that are estimable based on the jth stage data, $j = 1, \ldots, m$. Note that Θ_j is a random set for $j \geq 2$, taking values from 2^Θ, the power set of Θ. Under this setting, an adaptive design would be wildly flexible if with a nonzero probability there exists a $j \geq 2$ such that the possible value of Θ_j forms an uncountable subset of 2^Θ. Otherwise, it is flexible. To further illustrate the concept, consider the two-stage Thall et al. (1988) design with k treatment arms at the first stage and assume that the total number of potential treatment effect parameters are infinitely many, i.e., $K = \infty$, $m = 2$, and $\Theta_1 = \{\theta_1, \ldots, \theta_k\}$. If the adaptation is to promote some of the θ_is in Θ_1, or to drop some of the first-stage treatment arms and add one new arm at the same time, then the design would be a flexible design. However, if there is no restriction on the number of new arms to be added at the second stage, then the design is considered a wildly flexible design.

In practice, a wildly flexible adaptive design should be avoided whenever possible. Besides, in addition to the concern of implementation, such designs are often difficult, if not impossible, to justify (Liu et al., 2002). The preceding argument justifies our purpose to distinguish such designs from flexible adaptive designs.

Flexible adaptive designs are much better understood. For example, sample size modification methods based on observed treatment effect were proposed by Proschan and Hunsberger (1995), Cui et al. (1999), and Denne (2001), among others; and adaptive testing with treatment selection was considered by Thall et al. (1988), and Stallard and Todd (2003), among others. For the justification of a general flexible design, i.e., one that is not limited to sample size modification, important work has been done by Bauer and Köhne (1994), Fisher (1998), Brannath et al. (2002), Liu et al. (2002), and Bretz et al. (2006), among others.

There are debates on the relative merits of a flexible adaptive design comparing with its nonadaptive counterpart, such as a group sequential design, see, for example, Tsiatis and Mehta (2003), Jennison and Turnbull (2006), and Burman and Sonesson (2006). However, neither group sequential designs nor adaptive designs have a definitive advantage in efficiency in the sense that for any given design in one family, there is a design in the other family that dominates it in power (Proschan et al., 2006). Therefore, in our opinion, it is not crucial to advocate one type over the other, particularly when there does not exist a unique optimal design. Rather, it is more sensible to choose the best one from a collection of feasible candidate designs, either adaptive or nonadaptive, based on some important criteria. Allowing for flexible adaptive designs is a convenient way to enhance such a collection of candidate designs and it may increase the chance of identifying a good design. After all, the claimed appealing properties of adaptive designs, if any, are better manifested in those adaptive designs that allows for considerable amount of flexibility as long as they are not too flexible.

3. CRITERIA FOR CHOOSING A GOOD ADAPTIVE DESIGN

In the previous section, we suggest that candidate designs for A&WC trials should include flexible adaptive designs, in addition to group sequential designs. Intuitively, adaptive designs are very appealing and their advocators believe that such flexible designs maximize the chance of success and expedite the drug development process. It seems that adaptive designs are natural candidates for good designs. However, Jennison and Turnbull (2006) showed that the gain in efficiency by using adaptive designs, if any, is marginal. The criteria adopted in their paper consist of power and expected information at trial termination. This raises an important question: Is it possible to come up with a set of criteria for good designs that are more consistent with investigator's intuition so that under these criteria adaptive designs have a better chance of dominating their nonadaptive counterparts?

Before setting up our criteria, we share some thoughts on fair comparison between adaptive designs and group sequential designs. Consider the specific clinical setting of $\Theta = \{\theta_1, \ldots, \theta_K\}$ that we discussed in the previous section. We emphasize the random nature of Θ_j, the set of estimable treatment effects at stage j, when $j \geq 2$. In reality, at the outset of a trial, the only set of parameters that people are aware of is Θ_1, and without first-stage data, some parameters in Θ_2 may never be explored. For a flexible design where Θ_j is not a subset of Θ_1 for some $j \geq 2$, this indicates accumulation of new knowledge, an important feature of a flexible design. Based on this observation, we argue that a comparison is only fair when all designs start with the same Θ_1. By the same token, we believe that the design selection criteria should reflect the overall benefit on knowledge gained during the whole drug development process. The criteria we propose consists of three rules:

- The rule of validity.
  - V1: A valid design must have a controlled study-wise type I error rate.
  - V2: A valid design should also yield good point or interval estimation of treatment effects during and upon the completion of the trial.

- The rule of efficiency.
  - E1: The smaller the type II error rate, the better.
  - E2: The smaller the type III error rate (defined as the probability of recommending a suboptimal treatment), the better.
- The rule of parsimony.
  - P1: The smaller the total expected sample size under null and alternative, the better.
  - P2: The fewer the expected number of stages under null and alternative, the better.

As acknowledged in the guidance, it is important to provide good estimation of treatment effect upon completion of a multistage trial. We include good estimation as part of the validity rule because we feel that an adaptive design with controlled study-wise error rate does not necessarily yield a good estimation of the treatment effect. A good estimation of the treatment effect is important for future trial designing and final drug labeling. Study-end treatment effect estimation is a challenging topic. Methods have been proposed by Posch et al. (2005) and Stallard and Todd (2005), among others. The efficiency rule E2 is relevant to the comparison between a flexible adaptive design and a group sequential design. To see this, consider a two-stage setting with one treatment arm and one control arm at the first stage, that is, $m = 2$ and $\Theta_1 = \{\theta_1\}$. Suppose at the completion of the first stage, the group sequential design recommends continuation of the treatment while an adaptive design recommends, switching to a new treatment arm with treatment effect θ_2. If $\theta_2 > \theta_1 > 0$, assuming positive values of θ_is indicating treatment effect and the larger the value the more pronounced the treatment effect, then a type III error has been committed under the group sequential design but not under the adaptive design.

The preceding proposed set of criteria is an effort to provide a better understanding of the relative merits and disadvantages of different designs. In clinical trials, flexible adaptive designs aim at obtaining maximal knowledge from the existing data and putting it into best use for subsequent stages, and the underlying philosophy of this aim is compatible with Bayesian thinking. In this sense, a decision-theoretic formulation to incorporate the preceding criteria seems appropriate. See Stallard (2003) and Jennison and Turnbull (2006) for works along this line. We fully agree with the FDA that extensive clinical trial simulation plays an important role in planning and evaluating flexible adaptive designs. It, however, should be noted that clinical trial simulation is "a" solution but not "the" solution for evaluation of the flexible adaptive designs.

We echo the FDA's viewpoint that adaptive designs are susceptible to bias, both statistically and managerially (see also Hung et al., 2006). This means that not all the concern in bias can be addressed statistically even if a composite criterion is used. To address such issues, we suggest a modified "two-trial rule": First run a trial under a good (and possibly flexible adaptive) design to gain maximal information on multiple treatments and/or endpoints within a relatively short period of time, and then run the second trial under a classical single-stage design or a group

sequential design for further confirmation. The drug effect is claimed only when both trials result in success.

4. CONCLUSION

The current draft guidance makes a distinction between well-understood adaptive designs and less well-understood adaptive designs. We comment that it would be helpful to further classify the less well-understood adaptive designs according to their degrees of flexibility. In this case, the sponsors who are interested in conducting adaptive clinical trials will have a better idea regarding the type of adaptive designs they should focus on. In addition, we propose a set of criteria for choosing a good design in clinical trials. The proposed set of criteria reflects more closely the real clinical practice and provides a fair evaluation of different adaptive designs. We emphasize that choosing a good design from a pool of candidates is of more importance and significance than choosing between adaptive designs and group sequential designs.

REFERENCES

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

Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* 97:236–244.

Bretz, F., Schmidli, H., König, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* 48:623–634.

Burman, C., Sonesson, C. (2006). Are flexible designs sound? *Biometrics* 62:664–683.

Chow, S. C., Chang, M. (2006). *Adaptive Design Methods in Clinical Trials*. New York: Chapman and Hall/CRC Press, Taylor & Francis.

Cui, L., Hung, H. M. J., Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* 55:853–857.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* 20:2645–2660.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* 17:1551–1562.

Food and Drug Administration. (2010). *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*. Rockville, MD: Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, Food and Drug Administration.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J. (2006). Adaptive design in clinical drug development—An executive summary of the PhRMA Working Group (with discussions). *Journal of Biopharmaceutical Statistics* 16:275–283.

Hung, H. M. J., O'Neill, R. T., Wang, S. J., Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* 48:565–573.

Jennison, C., Turnbull, B. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika* 93:1–17.

Jennison, C., Turnbull, B. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics* 17:1135–1161.

Liu, Q., Proschan, M. A., Pledger, G. W., (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 97:1034–1041.

Liu, Q., Pledger, G. W. (2005). Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association* 100:493–502.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 24:3697–3714.

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

Proschan, M. A., Lan, K. K. G., Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trails: A Unified Approach*. New York: Springer.

Stallard, N. (2003). Decision-theoretic designs for phase II clinical trials allowing for competing studies. *Biometrics* 59:402–409.

Stallard, N., Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 22:689–703.

Stallard, N., Todd, S. (2005). Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference* 135:402–419.

Thall, P. F., Simon, R., Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 75:303–310.

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