Efficiency Perspectives on Adaptive Designs in Stroke Clinical Trials

Ken Cheung, PhD; Petra Kaufmann, MD

Abstract—An adaptive design allows the modifications of various features, such as sample size and treatment assignments, in a clinical study based on the analysis of interim data. The goal is to enhance statistical efficiency by maximizing relevant information obtained from the clinical data. The promise of efficiency, however, comes with a cost, per se, that is seldom made explicit in the literature. This article reviews some commonly used adaptive strategies in early-phase stroke trials and discusses their associated costs. Specifically, we illustrate the trade-offs in several clinical contexts, including dose-finding in the Neuroprotection with Statin Therapy for Acute Recovery Trial (NeuSTART), futility analyses and internal pilot in Phase 2 proof-of-concept trials, and sample size considerations in an imaging-based dose-selection trial. Through these illustrations, we demonstrate the potential tension between the perspectives of an individual investigator and that of the broader community of stakeholders. This understanding is critical to appreciate the limitations, as well as the full promise, of adaptive designs, so that investigators can deploy an appropriate statistical design—be it adaptive or not—in a clinical study. (Stroke. 2011;42:00-00.)

Key Words: continual reassessment method • futility interim analysis • Internal pilot • prospective planning

Clinical development of new therapies for acute ischemic stroke has seen limited success since the approval of tissue-type plasminogen activator\(^1\) by the U.S. Food and Drug Administration. Many factors contribute to the difficulty in developing and testing new therapies, including challenges in consenting patients, variability in the standard of care, inadequate patient recruitment rates, and delays between trial phases as drugs move from early dose-finding to efficacy trials. An adaptive design is a statistical tool that is hoped to accelerate drug development. Recent U.S. Food and Drug Administration draft guidance defines an adaptive design as a “prospectively planned opportunity for modification of one or more specified aspects of the study design” based on interim analysis of a study.\(^2\) The term prospective means that modification is planned before data are examined in an unblinded manner. Behind this overarching definition, the literature of adaptive designs has a long and multifarious history. The concept of adaptive randomization\(^3\) was introduced in the 1930s, sample size recalculation\(^4\) in the 1940s, sequential dose finding\(^5\) in the 1950s, and play-the-winner strategies\(^6\) and group-sequential methods\(^7\) in the 1960s. These concepts have since been studied and refined to suit practical purposes,\(^8–15\) and were recently reviewed by the U.S. Food and Drug Administration\(^16\) and the PhRMA group.\(^17\) In early-phase stroke trials, appropriate use of adaptive designs, possibly in conjunction with advanced biomarkers, has been shown effective in reducing the required number of subjects although still maintaining comparable accuracy.\(^18\) However, there are costs associated with the use of adaptive designs, and such compromise is seldom made explicit in the literature. It is the purpose of this article to review some commonly used adaptive designs and their implied trade-offs, so as to aid in the decision of whether to adopt an adaptive design in a clinical study. Although it is futile to attempt to exhaust all possible adaptive designs, we aim to cover the most common early-phase trial settings, namely, Phase 1 dose-finding, Phase 2 proof-of-concept, and dose-selection trials.

Continual Reassessment Method in Dose-Finding Studies

Phase 1 trials are dose-escalation studies that assess toxicity of a drug. A specific aim is to estimate the maximum tolerated dose (MTD), a dose associated with a target rate of dose-limiting toxicity (DLT). The Neuroprotection with Statin Therapy for Acute Recovery Trial (NeuSTART) drug development program was initiated to test the role of high-dose statins as early therapy in stroke patients. (K.C. was the study statistician, and P.K. served on the Safety Monitoring Board.) In a Phase 1B trial under NeuSTART, high-dose lovastatin was given to patients for 3 days. DLT was defined as clinical or laboratory evidence of hepatic or muscle toxicity, and the objective was to identify the dose associated with a 10% DLT rate.\(^19\) The trial was conducted in 33 subjects...
in a dose-escalation fashion among 5 possible dose-tiers, and the MTD estimate was 8 mg/kg per day. Dose assignments for the subjects enrolled onto the trial were determined by time-to-event continual reassessment method (CRM). The time-to-event CRM was used as an alternative to the 3+3 dose-escalation scheme; the latter, originally motivated by applications in oncology, was previously shown to be inappropriate for stroke trials because it would choose doses at a much higher toxicity level compared with time-to-event CRM.

The CRM is efficient at estimating MTD. Figure A displays the distribution of MTD selection by the NeuSTART design under a scenario where the third dose-tier is the MTD and the toxicity odds ratio (OR) of each subsequent dose tier is 2.5; the MTD was correctly identified with a probability of 0.54. If we use a nonadaptive design by randomizing 33 subjects to the 5 dose-tiers with equal likelihood, the MTD will be selected with a probability of 0.47 (Figure B). Also, the CRM selects an overdose (ie, dose-tiers 4 or 5) less often than does the randomization design. If we increase the sample size and randomize 45 subjects evenly to the doses, we will have comparable accuracy to the CRM with 33 subjects in terms of selecting the MTD; however, the tendency to select an overdose remains (Figure C).

The CRM not only improves accuracy, but also prescribes doses that reduce risks to the study subjects. In the scenario in Figure, the CRM on average enrolls 13 of 33 subjects at the MTD and 6 at an overdose, whereas randomization will place an average of 13 subjects at an overdose and 7 at the MTD (Table 1). This reflects that the CRM adapts to interim observations in an ethically appropriate manner; no escalation

### Table 1. Comparison of the NeuSTART CRM and Even Randomization

<table>
<thead>
<tr>
<th>Design Characteristics</th>
<th>CRM</th>
<th>Randomization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of selecting the MTD*</td>
<td>0.54</td>
<td>0.47</td>
</tr>
<tr>
<td>Probability of selecting an overdose*</td>
<td>0.17</td>
<td>0.26</td>
</tr>
<tr>
<td>Average no. of subjects at MTD</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Average no. of subjects at an overdose</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Median of toxicity odds ratio estimate*</td>
<td>5.2</td>
<td>2.6</td>
</tr>
</tbody>
</table>

*The MTD and the odds ratio were estimated using logistic regression at the end of each simulated trial for the CRM and the randomization design.

Comparison contains 33 subjects under a dose-toxicity curve where dose tier 3 is the MTD and the toxicity rate increases with an odds ratio of 2.5 at each increment.

CRM indicates reassessment method; MTD, maximum tolerated dose.
Futility Interim Analysis in Proof-of-Concept Studies

Phase 2 studies serve as a proof-of-concept by examining pilot efficacy of a new drug. A main consideration is the choice of a biomarker that correlates with stroke outcome. A promising biomarker is magnetic resonance imaging (MRI) response. The simplest Phase 2 trial design is a single-arm study in which patients are given the experimental drug, and the experimental response rate is compared to a historical control rate. Based on the results by MR Stroke Collaborative Group (MRSCG), we may assume 25% MRI response among untreated stroke patients. To have 80% power to detect a 45% response rate at a 5% significance level, we will need to observe at least 14 responses in a fixed sample size of 36 subjects.

For the same power, significance level, and treatment rate, we may alternatively use a 2-stage design with a futility interim analysis. In stage 1, enroll 17 subjects and conclude futility if there are ≤5 responses; if there are ≥6 responses in stage 1, treat an additional 24 subjects in stage 2, and declare the drug efficacious if there are ≥15 responses in the 41 subjects. Because of the provision of early stopping, this 2-stage design will enroll an average sample size of 23 subjects if the experimental response rate is in truth the same as the control rate of 25%. To interpret an average sample size, imagine that 100 single-arm trials of different drugs use this 2-stage design; assuming most of the drugs are no better than are controls, we will expect to enroll about a total of 23×100=2300 subjects onto these trials—although some of the trials will stop after 17 subjects, and some will continue to stage 2 and enroll 41. In contrast, if we use the fixed design with 36 subjects, we will need 3600 subjects for the same 100 trials. Looking at a portfolio of several trials, the 2-stage design is the obvious choice, assuming that most drugs do not work. Conversely, for investigators who hope to show their drug is efficacious, the fixed sample size design is more appealing than is the 2-stage design, because the former will take them 5 fewer subjects (36 versus 41) than will the latter. This numeric comparison demonstrates a potential tension between the perspectives of the individual investigator and the broader community. A statistical theory stipulates that, to achieve the same power at the same significance level, the maximum sample size required by any adaptive designs will always be at least as large as that of a fixed sample size design; that is, the advantage of an adaptive design lies in reduction of the average sample size. This theory thus implies that an adaptive design cannot resolve fundamental difference in perspectives. The individual investigator’s interest resides in keeping the sample size of a single trial small. Given limited resources and finite numbers of stroke patients, the community’s interest resides in keeping the average sample size small so that more trials can be performed.

Internal Pilot in Randomized Studies

For randomized studies comparing event rates of a new treatment to concurrent placebo, sample size calculation requires an assumption of the placebo rate as well as the effect size. The assumed placebo rate should reflect the aggregate experience about the natural history of stroke patients, formed through literature search or the investigators’ clinical experience. However, it is very possible that the assumed placebo rate misses the truth; this is exactly why a randomized study is needed instead of a single-arm study.

Consider an MRI trial where patients are randomized between an experimental treatment and a placebo. With an assumed 10% placebo rate and a target effect size of 20 percentage points, a 1-sided test with 5% type I error and 80% power requires 49 subjects per arm. If the true placebo rate is 25% and the treatment rate is 45%, the power effected by this sample size will reduce to 64%, despite the fact that the effect size remains to be 20 percentage points. If we assume a 25% placebo rate and a 45% treatment rate, the required sample size for 80% power is 70 subjects, which may prove unnecessary if the placebo rate is in truth 10%.

Conducting an internal pilot provides a means to circumvent the dilemma caused by uncertainty in the placebo rate. The idea is to calculate the sample size using blinded estimates of the response rates in an internal pilot. Table 2 shows the properties of a 2-stage design with an internal pilot of n1 = 30 subjects, and that of fixed sample size designs with N=49 and N=70 (Details of 2-stage design are given in online Supplement; http://stroke.ahajournals.org). When the true placebo rate is 10%, the fixed design with N=70 is overpowered, whereas the 2-stage design gives adequate power with an average sample size much smaller than 70 subjects. In contrast, when the true placebo rate is 25%, the

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Fixed Design (N=49)</th>
<th>Fixed Design (N=70)</th>
<th>Internal Pilot (n1 = 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo: 10%/Treatment: 30%</td>
<td>Power, %</td>
<td>80</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Average no. per arm</td>
<td>49</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>IQR (49–49) (70–70)</td>
<td>(50–56)</td>
<td></td>
</tr>
<tr>
<td>Placebo: 25%/Treatment: 45%</td>
<td>Power, %</td>
<td>64</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Average no. per arm</td>
<td>49</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>IQR (49–49) (70–70)</td>
<td>(68–74)</td>
<td></td>
</tr>
</tbody>
</table>

Comparison in a randomized study assuming a 20-percentage-point effect size and 80% power at 5% significance. IQR indicates interquartile range.
The investigators decide to perform a sample size calculation (18). In contrast, a fixed sample size randomization design to achieve 80% power and 5% type I error (Table 2 by Fisher). The total sample size (per arm) is calculated to achieve 80% power under a 20-percentage-point effect size at 5% significance (one-sided).

fixed design with N=49 is underpowered, whereas the 2-stage design achieves 80% power. The adaptable sample size of internal pilot thus offers great flexibility in trial planning.

However, because we use the internal pilot data twice (in the sample size estimation and in the final analysis), adjustments are needed for the final statistical test to preserve the type I error rate (online Supplement). The sample size calculations in Table 3 indicate that the adjusted statistical test (with n1=30) suffers only a slight loss in efficiency when compared with the Z-test in a fixed design. Generally, the efficiency loss depends critically on the ratio of the internal pilot sample size (n1) to the final sample size (N). Suppose the investigators decide to perform a sample size calculation at n1=15 without advance planning. Table 3 shows that the sample size inflation from the fixed design can be substantial; however, because a fixed design with a misspecified placebo rate can be underpowered, an unplanned re-estimation is arguably superior by providing greater flexibility. However, in view of efficiency, planned sample size re-estimation is preferred to unplanned interim calculation.

### Dropping-the-Loser in Dose-Selection Trials

If we believe a dose below the MTD may be efficacious, it is appropriate to conduct a Phase 2B dose-selection study, where primary objectives are to make a "go-or-no-go" decision and to decide with which dose to move forward. Fisher et al (18) describe a drop-the-loser strategy for a 3-arm, placebo-controlled, dose-selection trial using MRI response. The design eliminates at least 1 of 2 doses if the early response rates in the treatment arms are not promising. By assuming a 10% placebo MRI response and a 20-percentage-point effect size, the 2-stage design requires a maximum of 126 subjects to achieve 80% power and 5% type I error (Table 2 by Fisher et al (18)). In contrast, a fixed sample size randomization design requires 216 subjects to achieve the same power and significance level. This comparison stands in contrast with that in single-arm studies, where an adaptive design will always need larger maximum sample size than will a fixed design. In this regard, an adaptive strategy has a universal advantage over the nonadaptive design in multiple-arm dose-selection trials. The intuition is by eliminating inferior doses at an interim time point, resources can be directed toward the promising dose for precise comparison against the placebo.

However, by forcing us to drop at least some doses, we will not be able to estimate the OR of MRI response between 2 doses, which may provide evidence to evaluate the risk-benefit ratio. Therefore, if one holds a local perspective to evaluate fully the dose-response and other clinical parameters of a drug in a study, dropping-the-loser may not provide us with the answer. However, with a global drug development perspective to identify efficiently a good dose for a Phase 3 trial, adaptive designs have much to offer.

To make the matters more complicated, the power of a dose-selection trial depends on the placebo rate. The aforementioned drop-the-loser strategy was designed before the MRSCG analysis became available, and would have had only about 72% power with a 25% placebo response rate. Additional adaptation can be made to accommodate unknown placebo rate. However, the greater the uncertainty at the design stage, the larger the sample size one will need. Therefore, meta-analysis of the natural history of stroke patients will prove to be extremely valuable in terms of reducing the necessary resources.

### Discussion

The importance of prospective planning in an adaptive design is, first of all, the elimination of (the perception of) bias because of unplanned looks at the data. From an efficiency viewpoint, any ad hoc adaptation potentially leads to inefficiency, as illustrated in the discussion of internal pilot. More importantly, the flexibility offered by an adaptive design should not replace careful planning and preliminary investigation. For example, although sample size re-estimation techniques can, in theory, be used to adapt to the uncertainty in effect size, they raise a variety of issues; this includes the necessity of unblinding, a potentially prohibitive final sample size, and inefficiency. Thus, it is crucial for the investigators to consider carefully what constitutes a minimally relevant effect size in the planning stage. In the aforementioned dose-selection trial, a relatively large improvement (ie, a 20-percentage-point effect size) in the MRI response was believed necessary for translation into meaningful clinical benefits. For another example, thorough investigation of the natural history of stroke outcomes in previous clinical data can reduce uncertainty in placebo rate; thus, meta-analysis is particularly useful to build reliable historical controls as were performed by MRSCG. Finally, understanding of the dose-response and drug mechanism in preclinical data can help zero in on appropriate doses and end points in the clinical phase. These conventional good practices may result in more substantial efficiency gain than what adaptive designs can achieve.

Adaptive designs can be associated with logistical challenges, such as forecasting budgets, planning for drug supply, and the potential needs for real-time Data and Safety Monitoring Board decisions. However, these challenges can be addressed, and adaptive designs have much to offer in view of statistical efficiency—from the reduction of average sample size in a futility interim analysis to the adaptive dose assignments in the CRM. These advantages are not new in the literature. In this article, we have emphasized that whether we gain efficiency by using an adaptive design depends on one’s perspective. And, the tension between perspectives is irrefutable by any statistical design, be it adaptive or not.
Rather, the debate about the appropriate perspective for a study should precede and dictate the choice of statistical methods. Indeed, adaptive design is no panacea and should not be mistaken as one.

**Sources of Funding**
This work was supported by the National Institutes of Health.

**Disclosures**
K.C. received funding from the National Institutes of Health to conduct part of this work. This work is related to the authors’ work at Columbia University and should not be considered as the opinion of the National Institutes of Health or its affiliates.

**References**
Efficiency Perspectives on Adaptive Designs in Stroke Clinical Trials
Ken Cheung and Petra Kaufmann

Stroke. published online September 1, 2011;
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/early/2011/09/01/STROKEAHA.111.620765

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2011/09/02/STROKEAHA.111.620765.DC1
ONLINE SUPPLEMENT

Efficiency perspectives on adaptive designs in stroke clinical trials

Ken Cheung, PhD; Petra Kaufmann, MD

Department of Biostatistics (K. C.), Columbia University, Mailman School of Public Health, New York, New York, USA; Office of Clinical Research (K. C.; P. K.), National Institute of Neurological Disorders and Stroke, North Bethesda, Maryland, USA.

Correspondence to: Ken Cheung, PhD, Department of Biostatistics, Columbia University, 722 West 168th Street, Room 641, New York, NY 10032, USA.

Tel.: 212-305-3332
Fax: 212-305-9408
Email: yc632@columbia.edu

Cover title: Adaptive designs in stroke trials
Supplement Methods

This section describes in detail the internal pilot design used in the main article, and provides technical details for power and sample size calculation.

Design approach

The basic design takes a two-stage approach, with an objective to achieve 80% power for an effect size of 20 percentage points at 5% significance (one-sided).

In the internal pilot stage (stage 1), randomize 30 subjects each to the treatment arm and the placebo; thus a total of 60 subjects. Then at an interim, we calculate blinded estimates of the placebo rate and the treatment rate under the assume effect size, as follows:

1. Calculate the pooled response rate in all 60 subjects, denoted by $\bar{p}$
2. Then the placebo rate is estimated by $\bar{p} - 0.10$ and the treatment rate is estimated by $\bar{p} + 0.10$.

The sample size of stage 2 is then calculated based on these estimated rates; see details below.

Statistical test and its efficiency

Because the internal pilot data is used to calculate the stage 2 sample size and used in the final analysis, the standard Z-test (that combines data from both stage 1 and stage 2) may cause inflated type I error rate. While there are many possibilities to perform an adjusted statistical analysis, we consider the following statistical test to facilitate our discussion in the main article: the null hypothesis will be rejected if

$$\frac{Z_1 + Z_2}{\sqrt{2}} \geq 1.645$$

where $Z_1$ and $Z_2$ are respectively the Z-scores based on stage 1 and stage 2 data only. The power of this adjusted test can be shown to be approximately equal to

$$\Phi \left[ \frac{0.1 \left( \sqrt{n_1} + \sqrt{n_2} \right)}{\sqrt{\bar{p}(1-\bar{p})}} - 1.645 \right]$$

where $\Phi$ is the standard normal distribution, and $n_1=30$ is the stage 1 sample size and $n_2$ is the stage 2 sample size. Therefore, sample size calculation at the interim is to choose the smallest $n_2$ that will give an 80% power. To ensure that there is an adequate sample size in stage 2, $n_2$ is set to be at least 19, so that the total sample size is at least 49 (sample size for a 10% placebo rate).

The power of the adjusted test depends on $n_1$ and $n_2$. Generally, and quite intuitively, when a small $n_1$ is used in the internal pilot, a relatively large $n_2$ is needed in stage 2; and vice versa.
Besides, the total sample size (and hence efficiency) of the adjusted test depends on the relative sizes of $n_1$ and $n_2$. Table 3 in the main article shows that if an internal pilot with $n_1 = 15$ is used, the efficiency loss compared to the standard Z-test is more substantial than when $n_1 = 30$ is used. This underlines the importance, in terms of efficiency, of careful planning of internal pilot in advance, and recommends against any ad hoc, unplanned interim analysis.

**Supplement references**