Considerations on the Use of Adaptive Designs in Clinical Trials

By George DeMuth, MS
gdemuth@statonellc.com
Stat One LLC, Wilmington NC
April 2019

Abstract

This paper discusses factors to consider when considering the use of an adaptive design in the clinical trial setting with a focus on a pivotal study. The objective is to provide a framework for non-statisticians to consider when adaptive designs are proposed. First the paper introduces some definitions and the type of adaptive designs that will be discussed for this paper. Next, it discusses reasons that one might choose or choose not to use an adaptive design. Finally, some metrics for assessing competing designs are shown with an example.

It is assumed that the reader has some familiarity with the concepts of statistical testing and the interpretation of p-values. This paper is based on a Frequentist approach, but many of the concepts are applicable for any adaptive design setting.

Definition of an Adaptive Design

A very broad definition of an adaptive clinical trial could include any kind of change in the conduction of the study after it has started. A partial list of designs approaches is shown below:

Table 1, Types of Adaptive Designs

Design	Description
Adaptive Randomization	Seeks to adapt the randomization balance to increase overall probability of success
Group Sequential	Interim analyses performed to stop studies early for effectiveness or futility
Sample size adjustment	Provides sample size re-estimation
Drop-to-loser / Pick-the-Winner	Drops study groups, usually doses, to focus on the best performers.
Adaptive dose-finding	Studies designed to pick the best or maximum tolerated dose
Adaptive treatment-switching	Designs that allow non-responders to switch treatments
Adaptive hypothesis	Modification of the study primary hypothesis
Seamless design	Designs that start in one study phase that, based on interim results, continue to a second (regulatory) study phase (e.g. Phase I/II or Phase II/III)
Multiple-adaptive	A combination of design features

Abstracted from Pong and Chow (2011) Handbook of Adaptive Designs in Pharmaceutical and Clinical Development.

There is a tremendous volume of literature on adaptive designs (a Google search on “adaptive clinical trials” returned 44.9 million results at the time of this writing) and it has been an area of research for many years. This paper will only consider group sequential and sample size re-estimation study designs as they are among the more commonly considered adaptive designs. Further, we will only consider several specific examples.

Clinical Trial Setting and Types of Statistical Errors

This section provides a discussion of controlling error in a study, statistical power, and several actions that can occur in a group sequential study.

For the purposes of this review, we will assume we are only talking about the clinical trial setting. Further, the clinical trial has a primary objective and a primary analysis based on an endpoint representative of the primary objective. The goal of the study is to get a statistically significant result in the primary analysis where statistically significant is defined as getting a p-value from a statistical test being less than or equal to a cut-off. That is a standard Frequentist approach, but if we were consider a Bayesian approach it might be that we want a posterior probability about some hypothesis being greater than or equal to some cut-off.

For example, suppose we are comparing two treatments for a condition and we want to evaluate the response rate at the end of the study where the endpoint is a success/fail outcome (e.g. binary endpoint). There are many statistical tests for this setting: Chi-square tests, logistic model, Fisher's exact test, etc. We will use the example to discuss various study outcomes and, later, consider a test that evaluates the null hypothesis that experimental treatment response is the less than or equal to the control response rate (e.g. a one-sided hypothesis).

In the context of statistical testing of the primary endpoint, there are two types of errors that can occur in the study. The errors are:

Type I	This is when the study gives a statistically significant result when there is no true difference in the treatments. This is a false positive.
Type II	This is when the study does not give a statistically significant result but there is a true difference in the treatments. This is a false negative.

Statistical power is the probability of a study getting a statistically significant result when there is a true difference (e.g. 1 minus the probability of a Type II error). In the regulatory setting, we typically want to have studies that have at least 0.8 power (80% chance of success) and, better yet, 0.9 power (90% chance of success). From a pragmatic basis, we only want to run a trial that has a good chance to succeed because trials are typically expensive and time consuming. For an ethical consideration, it may not be reasonable to ask subjects to receive an experimental or potentially ineffective treatment in a study with little chance to get a statistically significant result.

An important point is that statistical power in a study is not fixed. For any given assumption about response rates, adding more subjects will increase the study's power. The number of subjects in a study is also referred to as the subject's sample size.

In the regulatory setting, we are almost always interested in fixing the Type I error rate for a study (aka, the alpha level or α level). This is because Type I involves a risk saying a product that truly does not work is effective. This Type I error rate applies to the entire study, including any changes in an adaptive design. Hence, adaptive designs have rules about how they operate in order to avoid making a Type I error. Frequently the Type I error rate is 0.05 (α) for a two-sided test or 0.025 for a one-sided test. In the setting of an adaptive design, it is most common to consider a one-sided test with the null hypothesis that the experimental treatment is worse than or equal to the control and the alternative hypothesis that the experimental treatment is better than control (e.g. a one-sided 0.025 test is used).

In summary, we must perform the statistical analysis in a way that maintains the Type I error for a study and we want to maximize power at the same time. The objective of an adaptive design is to give us better power or allow us to stop early.

When to Consider an Adaptive Design (and Not)

Adaptive trials should be part of everyone's design tool box as a way to keep an updated dataset for the study for a more accurate outcome. This section outlines when you choose an adaptive design. Although adaptive trials should be a part of everyone's design there are some cons associated with this type of design:

- Adaptive designs involve additional analyses
- Additional outside experts or statistical groups may need to be involved so the company stays appropriately blinded
- The database will have to be suitably cleaned multiple times
- The timing of the completion of a study and total cost of the study will be variable, especially if sample size re-estimation allows for a wide range of total subjects
- The maximum sample size may be higher than a fixed sample size study even if the expected subjects is lower.

This paper only focuses on sample size re-estimation and group sequential (interim analyses). The reasons to use these approaches are:

- The treatment effect is not well understood and there are a range of commercially and clinically interesting differences
- The treatment effect may be bigger than expected and a treatment effect may be clear in a more limited sample size
- The study enrollment will be slow or expensive, so early stopping may provide considerable savings in time and expense
- There is a chance that treatment effect is so small it may not be worth running the trial because it will have very low statistical power

Sample size re-estimation can adjust a sample size in case the assumptions about a treatment effect are incorrect. Group sequential designs allow us to stop for either good results (i.e. for effectiveness) or for bad results (i.e. futility).

There are some important, but non-statistical reasons, that effects the consideration of an adaptive design. A primary consideration is whether a company can afford to run a larger study. It may not be feasible for the study to pursue smaller, though possibly clinically interesting, treatment effects. The flip situation may be where a company cannot achieve initial funding, but contingent on it not being futile may qualify for additional funding. Adaptive designs should be black boxes that keep the study results hidden from the company.

In order to make the decision about using an adaptive design, one really must understand the performance of a design based on assumed treatment effects and any other assumptions needed to understand the primary endpoint. Things we consider are the expected (or average) number of subjects given some assumptions; the chance of stopping early, and the power. These factors can be balanced against the cost inherent in the maximum sample size, the plausible range of treatment effects, and any cons associated with the adaptive design. In short, the decision to use an adaptive design must fit the knowledge about the setting and the company must have suitable resources to support the study.

Example Adaptive Designs

This section outlines a sample size re-estimation method and a standard group-sequential design. For this discussion, we will consider a one-sided effectiveness hypothesis that we will test at the 0.025 α-level. For our binary endpoint study example above, the null hypothesis is that the experimental response rate is less than or equal to the control response rate versus the alternative that the experimental response rate is higher than the control rate.

A common group sequential method is based on the concept of alpha-spending. These approaches adjust the alpha-level for interim and final analyses in order to make sure that the Type I error rate is maintained across all interim analyses. One starts with a maximum sample size and we perform interim analysis with the possibility of stopping early. The Lan-DeMet method (1983) is a commonly used approach and readily available using the ldbounds package in R. P-value cut-offs can be varied to allow for bigger p-values at earlier time points with lower values at later time points or, being conservative, using smaller cut-offs early to keep bigger p-value cut-offs for the final analysis. The table below provides some example alpha-spending with different parameters obtained using R and the bounds function.

Table 2, Example Alpha-Spending Rule P-value Cut-offs

	Timing of Interim Analyses				
Label	33%	50%	67%	75%	100%
No interim example	-	-	-	-	0.0250
2 Interim Looks equally space, Rule 1 (O'Brien-Fleming)	0.0001	-	0.0060	-	0.0231
2 Interim Looks equally space, Rule 2 (Pocock)	0.0113	-	0.0109	-	0.0108
1 Interim Look (O'Brien-Fleming)	-	0.0015	-	-	0.0245
1 Interim Look (Pocock)	-	0.0155	-		0.0139
2 Interim looks at 50% and 75% (O'Brien-Fleming)	-	0.0015	-	0.0092	0.0220

As an example, we will take the last row with the cut-offs of 0.0015 at 50%, 0.0092 at 75%, and 0.022 for the final analysis. If we have a p-value of 0.054 at the first analysis, the study would continue to next look. Suppose, the p-value is 0.008 at the second analysis, then study would stop with success.

Without simulation or other calculations, it is tough to compare these designs. However, one can see that lower p-value cut-offs in the interim analyses give larger p-values at the end. Note that each interim analysis uses the accumulated data available for each interim analysis.

Also, note that one can perform a futility analysis where the study would stop for inadequate response at any point in the study without affecting the final p-value cut-off.

One approach to sample size re-estimation is using the principle of independent p-values. These designs split a study into phases and get a p-value for each phase of the study using just the data in each phase of the study. Because the data in one phase of the study are independent of the data from the other phases, one can adjust the study in the later phases without changing the results from the early phase of the study. P-values from multiple phases of the study can be combined in multiple ways, so long as the Type I error of the overall study is preserved. A common application of the combination of p-values is do a single interim analysis with the following actions: stop for futility, stop for effectiveness, or adjust the sample size for the rest of the study. Hence, these methods combine the attributes of a group sequential method and incorporate sample size re-estimation.

For the example in the next section, we will use the product of the p-values approach for a 2-phase study. The looks at the multiplied values of the p-values from the two phases to determine statistical significance if the study goes to a second phase. We must specify these things in the rule:

- The number of subjects in the first phase of the study
- A p-value used to denote futility (e.g. the study stops for futility if the p-value is above this value)
- The minimum number of subjects in the second phase of the study
- The maximum number of subjects in the second phase of the study
- The value the product of the p-values must be less than or equal to for the study to be considered statistically significant.

The table shows several cut-off values that conserve an overall Type I error rate of 0.025 on a one-sided test.

Table 3, Product of P-values Significance Cut-offs based on Futility and Interim Analysis Cut-offs

	Interim Analysis P-value Cut-off for Significance		
Futility P-value	0.001	0.005	0.010
0.2	0.0054	0.0053	0.0050
0.3	0.0049	0.0048	0.0044
0.4	0.0045	0.0045	0.0040
0.5	0.0043	0.0043	0.0038

Suppose we use the 0.4 futility cut-off and 0.005 for the significant cut-off for the first analysis. Suppose the first p-value is 0.054, so the study continues to phase 2. We only need a p-value of less than or equal to 0.0045 / 0.054 or 0.0833 to get a successful study. E.g. 0.054*0.0833 = 0.0045 matches the cut-off for the product of the p-values from the two phases.

Sample size re-estimation uses the observed results in the first part of the study to calculate the number of subjects required to achieve the target p-value needed in the second part of the study to claim success.

Another common form of combining p-values is using weighted combinations of the p-values from the two phases. There are special cases when the sample size can be increased without altering the final p-value when the p-value is low at the time of assessment.

Operating Characteristics for Assessing Adaptive Designs

The final step in considering the utility of an adaptive design is how it will perform based on differing study assumptions. There are several measures that are common operating characteristics of adaptive designs including:

Table 4 Operating Characteristic in Adaptive Designs

Factor	Definition
Overall power	The probability of the study achieving a statistically significant result
Expected subjects	The average number of subjects we can expect if the trial were run many times
Probability of stopping early for success	The probability a study stops before the maximum sample size
Probability of stopping early for futility	The probability a study stops for lack of treatment effect before the maximum sample size is enrolled

These operating characteristics are frequently studied using simulation methods.

We will look at some situations based on hypothetical comparisons of the success rate of an experimental treatment to a control treatment. We want to design a study with 90% power. We will assume the control success rate is established to be 60% and the experimental treatment response rate is hypothesized to be 75%. However, the company has some interest in success rates as low as 70% and it is easily plausible the experimental control rate may be 80%. We are considering a 1:1 randomized study and we want to perform a one-sided 0.025 hypothesis test.

First, we might consider a fixed sample size without a group sequential analysis. The sample sizes required for 90% power for the 3 assumptions are follows:

Table 5 Sample sizes for 90% Power for a Fixed Sample Comparison for Binary Rates (1:1 randomization, one-sided 0.025 test, no Adjustment for Drop-out)

Control Response Rate	Experimental Response Rate	Total Sample Size
60%	70%	952
	75%	406
	80%	218

As seen above, this is a dramatic range of sample sizes based on the treatment differences. We might consider a group sequential design because we may have a good chance of stopping early if the response rate turns out to be 80%. We also might consider a sample size re-estimation if we want to have some power for a result where the treatment difference is slightly less than 75%. We will propose 3 analysis options and consider the operating characteristics across a range of assumed response rates for the experimental treatment.

Table 6, Design Alternatives for a Pivotal Study Design

Rule	Sample Size and Cut-offs
Fixed Sample Size (#1)	We will use the 90% power sample size associated with 75% response rate (N=406) without adaptive design elements
Group Sequential Design (#2)	We will use the 90% power sample size associated with 75% response rate, but we will perform interim analyses at 50% (N=202) and 75% (N=306) of the total sample. We will use the p-value cut-offs 0.0015, 0.0095, and 0.0220. Further if the one-sided p-value is 0.5 or greater at the first interim analysis, we will stop for futility. In short, this design adds group sequential looks to the fixed sample size analysis.
Sample Size Re-estimation Design (#3)	We will use the product of p-values approach and perform an interim analysis when 202 subjects have enrolled. We will stop for success if the one-sided p-value is less than or equal to 0.01. We will stop for futility if p-value at the interim analysis is greater than 0.4. If the study does not stop after phase 1, we will enroll a minimum of 98 more subjects and up to 398 subjects for a maximum study size of 600 subjects. The second phase sample size will attempt to achieve 90% power based on the observed results in phase 1. If the study does not stop at the first analysis, the product of the p-values must be less than or equal to 0.0040.

Table 7 shows the operating characteristics of these designs for experimental response rates of 60%, 70%, 72.5%, 75%, 77.5%, and 80%.

Table 7 Operating Characteristics for Design Alternatives in a Binary Response Study

Control Group Response Rate	Experimental Group Response Rate	Design	Overall Power	Expected Sample Size	Stopped for Futility	Stopped for Success at First Interim Analysis	Stopped for Success at Second Interim Analysis
60%	60%	#1	0.025	406	NA	NA	NA
		#2	0.026	297	0.530	0.002	0.008
		#3	0.026	359	0.585	0.011	NA
	70%	#1	0.561	406	NA	NA	NA
		#2	0.551	351	0.077	0.076	0.235
		#3	0.654	446	0.101	0.213	NA
	72.5%	#1	0.761	406	NA	NA	NA
		#2	0.752	333	0.035	0.149	0.346
		#3	0.843	411	0.048	0.342	NA
	75.0%	#1	0.900	406	NA	NA	NA
		#2	0.896	306	0.013	0.267	0.419
		#3	0.946	359	0.019	0.498	NA
	77.5%	#1	0.970	406	NA	NA	NA
		#2	0.967	277	0.004	0.417	0.429
		#3	0.985	305	0.006	0.660	NA
	80%	#1	0.994	406	NA	NA	NA
		#2	0.993	251	<0.001	0.586	0.351
		#3	0.997	259	0.002	0.802	NA

The figures below show the expected sample size and power for the 3 designs as a function of the experimental group treatment response.

Figure 1: Expected Sample Size for Study Designs 1 to 3 as a Function of Experimental Response

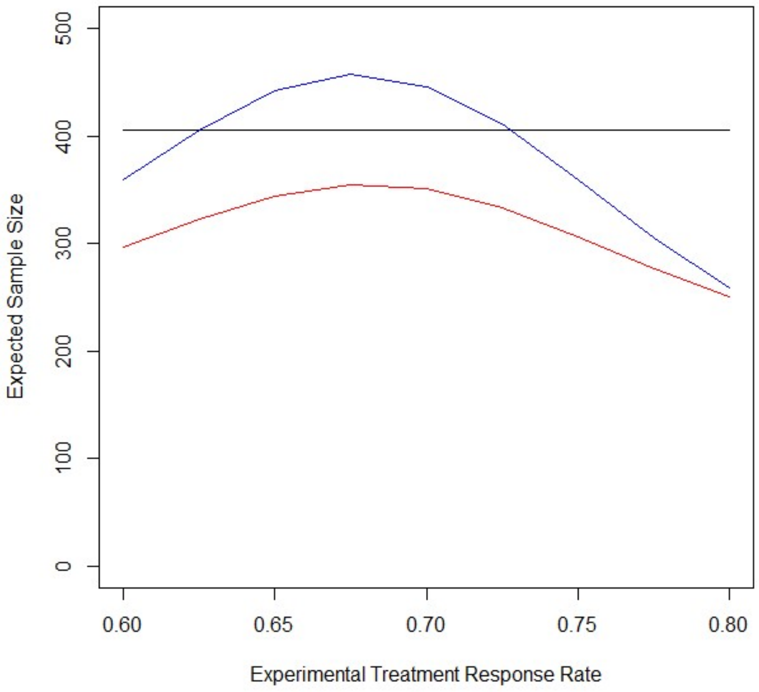
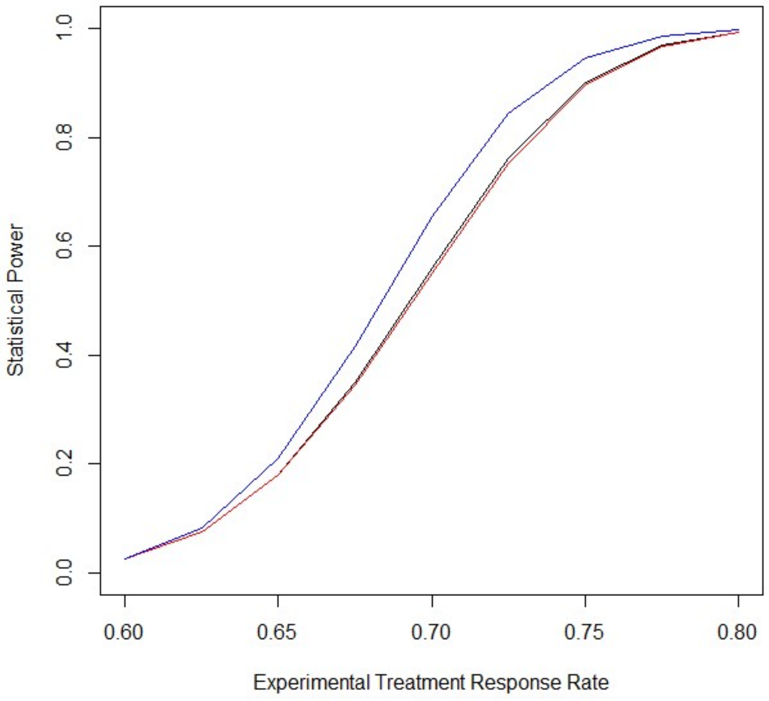


Figure 2: Statistical Power for Study Designs 1 to 3 as a Function of Experimental Response



Black = fixed sample site, Red = Group sequential, and Blue = Sample-size re-estimation

What are some observations about the relative performance of these designs?

Design #1

This design offers slightly better power than the group sequential design, but it may require substantially more subjects as it cannot stop at an interim analysis for futility or efficacy.

Design #2

At a slight cost of power relative to Design #1, Design #2 had the lowest expected sample size. This would likely be a good choice over Design #1 if there was little overhead to the interim analyses and rate of enrollment was such that stopping at an interim analysis would give a useful time savings to the company.

Design #3

This design had the best power in all response settings, but it requires the company to commit to a bigger possible study. The expected sample size was better than Design #1 for no treatment difference because of the futility assessment and for experimental response rates of 75% and higher in part because of the more aggressive p-value at the first interim analysis. Factors that would encourage the use of this design would be the company's desire to get a statistically significant study result even if the experimental response rate is less than 75% and having the highest chance of success at the first interim analysis.

This example demonstrates how adaptive design choices can be used to serve different objectives and balance the need for resources differently. The operating characteristics of the study designs are the necessary tool to compare the performance of designs.

Concluding Remarks

The goal of this paper was to expose readers to types of adaptive study designs and point out that the decision to use adaptive designs should account for the level of knowledge about the treatment effects, company constraints on the ability to complete a study, and the range of clinically and commercially meaningful treatment differences. Once we answer the questions about these questions, we can then use the operating characteristics to evaluate competing designs to determine which best fits a company's situation. Further, the operating characteristics gives us support for our decision and should allow us to have faith in the study design chosen.

Final comments:

- The two most important things to compare between designs are 1) the overall statistical power of the designs and 2) the average number of subjects. In certain cases, individuals may focus on maintaining the highest possible p-value for the final comparison even though there is no difference in overall power between two designs. If two designs have equal power, then the best design has the lowest average sample size requirement.

- In certain cases, adaptive designs can offer some insight into the likely status of the study, especially when a sample size re-estimation step is included. However, to the extent possible, companies should let the study run their course. Second guessing the study and trying to terminate early only increases the chances of a study failing and complicates the analysis and interpretation. The power obtained through calculations or simulation have already told us how the study will perform.
- Adaptive designs should be one of the things you consider for any study where the potential sample sizes are large or the studies are of long durations. In practice, budgets, knowledge about the treatment effect, clinically meaningful effects, and commercially viable treatment effects frequently create a narrow range of samples sizes that can be used in a study. In that case, a fixed sample size is often the best choice.

References:

Chang, M (2014) Adaptive Design Theory and Implementation Using SAS and R. CRC Press, Boca Raton, FL

Lan, K.K.G. and DeMets, D.L. (1983) Discrete sequential boundaries for clinical trials. Biometrika, 70:659-63.

Pong, A. and Chow, S-C (2011) Handbook of Adaptive Designs in Pharmaceutical and Clinical Development. CRC Press, Boca Raton, FL