# Title

stata.com

**GSD intro —** Introduction to group sequential designs

Description Remarks and examples References Also see

# Description

This entry provides a general introduction to group sequential designs (GSDs) and describes relevant statistical terminology. For an introduction to Stata's commands for GSDs, see [ADAPT] **gs**.

# Remarks and examples

stata.com

Remarks are presented under the following headings:

*Introduction*
    *FSDs*
    *GSDs*
*Components of GSD*
*Origins of GSD*
*Brief overview of GSD*
*Graphing group sequential boundaries*

## Introduction

For a brief introduction to adaptive designs, see [ADAPT] **Intro**. In this section, we describe GSDs for clinical trials in more detail.

Clinical trials are experimental studies in which the investigator assigns treatments to the participants. Each clinical trial begins with a design that determines the number of participants to recruit and how to allocate the participants to the treatments. GSDs are a subset of clinical trial designs that incorporate preplanned analyses of interim data.

In a GSD, the data-collection step is split into multiple predefined stages, and an interim analysis is performed at each stage as the data accumulate [Pocock (1977), O'Brien and Fleming (1979), Lan and DeMets (1983), Jennison and Turnbull (2000), Wassmer and Brannath (2016)]. Each analysis of the data is known as a look. Stopping boundaries are calculated for each look such that analyses from multiple looks are guaranteed not to exceed a predefined overall false-positive error rate, ensuring control of familywise type I error. Unlike fixed-sample designs (FSDs), GSDs can be stopped early in the presence of compelling evidence against or in favor of the null hypothesis.

### FSDs

To understand the process of creating and implementing GSDs, it is helpful to begin by considering FSDs. To plan an FSD, the investigator will begin by calculating the required sample size based on several factors, such as the size of a clinically meaningful effect, the desired power and significance level, results of previous studies, and practical considerations like cost and ability to recruit participants.

The next step is to recruit participants to the study. Depending on the scale of the study, recruitment could take place at a single site or at many sites. Recruitment often continues for months or even years. When a participant is recruited to the study, they are assigned, or randomized, to a treatment group. The gold standard of clinical trial design is a randomized controlled trial, where participants are randomly assigned to control or experimental groups, or arms. Trials without a control group are common in early-phase clinical trials designed to explore the appropriate dosage of a therapeutic agent and to investigate how the treatment affects participants. Uncontrolled trials are less common in late-phase clinical trials designed to demonstrate treatment efficacy, though there are circumstances warranting their use (see *Remarks and examples* in [ADAPT] **gsdesign onemean** and in [ADAPT] **gsdesign oneproportion** for examples).

In a classical two-arm randomized controlled trial, one group receives the experimental treatment, while the other group receives a control treatment. If there are no existing treatments that are comparable with the experimental treatment, the control group will typically receive a placebo. When a standard of care exists, there is often an ethical argument against using a placebo. In this case, an active control is used, wherein participants receive the existing standard of care.

After being assigned to a treatment arm, participants are monitored to collect data on the outcome of interest, which is typically referred to as the endpoint. In studies with multiple endpoints, it is common to designate one primary endpoint or to combine multiple endpoints into a single composite endpoint. Depending on the endpoint, the follow-up period may last for years. This is especially common in trials with survival outcomes (also known as time-to-event endpoints). Some participants might leave the study before their primary endpoint data are collected, a phenomenon known as loss to follow-up.

In a large clinical trial, it is not uncommon for several years to elapse before all the endpoint data are collected. If the trial follows an FSD, no analysis of treatment efficacy is conducted until all endpoint data have been obtained. At the end of an FSD, the data are analyzed and the null hypothesis is either rejected or not. In contrast with some other disciplines, in the context of clinical trials, it is common to describe the failure to reject $H_0$ as "accepting the null hypothesis". The flowchart in figure 1 details the course of an FSD.

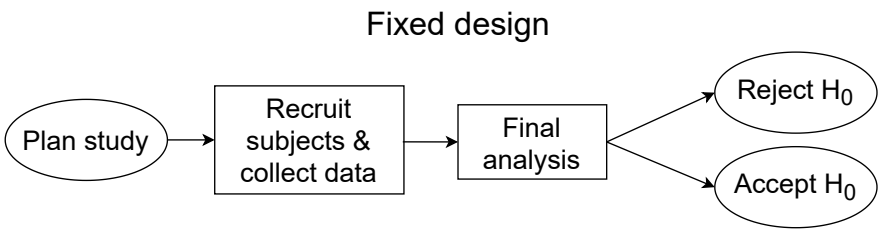


Figure 1. FSD flowchart

## GSDs

In the context of a long clinical trial, there is the potential for substantial benefit to both participants and sponsors if a treatment can be declared effective or ineffective before the trial is scheduled to end. GSDs accomplish this by allowing for multiple preplanned analyses of interim trial data while controlling the familywise error rate.

At each interim look, a statistical test is performed, and the test statistic is compared with sets of critical values called stopping boundaries to determine whether $H_0$ can be rejected (known as efficacy stopping) or accepted (known as futility stopping). If the interim test is inconclusive, the study continues to the next look. At the final look, $H_0$ must be rejected or accepted.

Planning a clinical trial using a GSD is similar to planning a trial using an FSD, but some additional considerations are required. On the logistical side, preparations must be made to ensure that interim data are of high quality and are quickly available to the data analysis team, often an independent group called a Data Monitoring Committee. On the statistical side, investigators must determine the type of stopping rule to apply (efficacy stopping, futility stopping, or both), the number and spacing of interim analyses, and the boundary-calculation procedure to be used.

Stopping boundaries are calculated, and sample-size calculations that account for the planned interim analyses are performed. The recruitment, randomization, treatment, and follow-up of a GSD are akin to those of an FSD. But instead of waiting to collect all endpoint data before analysis, interim analyses are performed, and the study can be terminated early for efficacy (if $H_0$ is rejected) or for futility (if $H_0$ is accepted).

The flowchart in figure 2 details the course of a GSD. Each interim analysis offers the opportunity to terminate the trial to reject $H_0$ (if efficacy bounds are used) and the opportunity to terminate the trial to accept $H_0$ (if futility bounds are used). If the interim test is inconclusive, the trial continues to collect more data until the next look. The process continues until an interim analysis determines that the trial should stop or until all possible data are collected and the final analysis is performed.

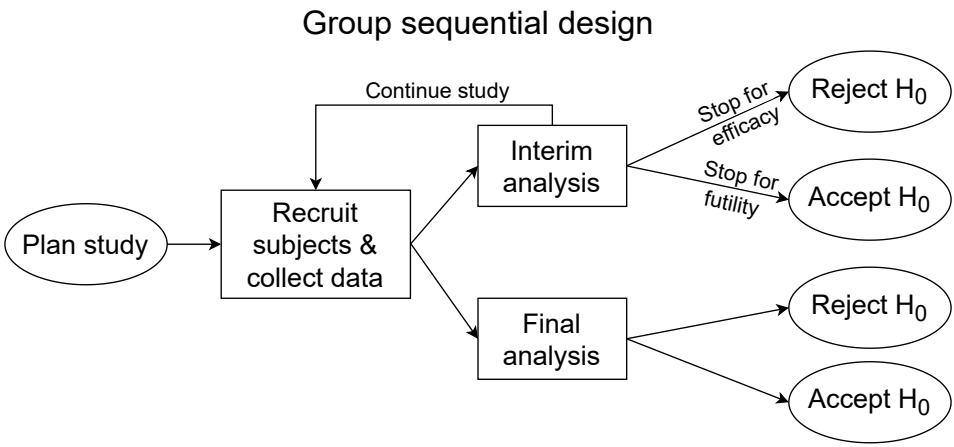


Figure 2. GSD flowchart

## Components of GSD

The components of power and sample-size analysis for FSDs are also relevant to GSDs. Please read *Components of PSS analysis* in [PSS-2] **Intro (power)** before reading this section. The key components of GSDs include variations of components of FSDs as well as components that are specific to GSDs. We describe them below.

- **Statistical analysis method:** A clinical trial following a GSD must identify the intended statistical analysis method during the design stage. The type of statistical test to be used dictates the methodology used in calculating the sample sizes at interim analyses (but not the critical values for the stopping boundaries). A GSD is able to provide strong control of familywise type I error when, under the null hypothesis, the sequence of test statistics from interim analyses follows a multivariate normal distribution with a covariance matrix that depends only on the amount of the data analyzed at each interim look. To use a GSD with a test that does not produce a normally distributed test statistic under the null hypothesis, the

significance level approach, which uses stopping boundaries based on $p$-values instead of critical values, may be used (see *Methods and formulas* in [ADAPT] **gsbounds** for details).

- **Significance level, $\alpha$:** GSDs must account for multiple hypothesis tests being conducted. It is not sufficient to conduct each test at the desired overall significance level $\alpha$ because conducting multiple tests means that the chance of committing a type I error at one or more interim analyses will be greater than the desired $\alpha$. Instead, the familywise error rate is controlled, ensuring that $\Pr(\text{reject } H_0 \text{ at any look} \mid H_0 \text{ is true}) = \alpha$.
- **Power, $1 - \beta$:** The power of a group sequential test is the probability of rejecting a false null hypothesis at any look, or $\Pr(\text{reject } H_0 \text{ at any look} \mid H_0 \text{ is false}) = 1 - \beta$. Power is calculated relative to a prespecified effect size, and the smaller the effect size, the larger the sample required to achieve a given power.
- **Accrual and endpoints:** In a GSD, as in most clinical trial designs, participants are generally recruited, or accrued, over time. The outcome of interest is known as the endpoint. GSDs offer the most benefit when the collection of endpoint data is rapid compared with accrual. If the time between randomization and endpoint-data collection is excessively long, there will be less benefit in terminating a trial early because resources already will have been expended to recruit many participants who are still in follow-up but whose endpoints have not yet been collected.
- **Interim looks:** Interim looks, or interim analyses of the data available to date, are the defining feature of GSDs. To conduct a GSD properly, it is necessary to ensure that endpoint data are collected in a timely and reliable manner and provided to the statistical analysis group or Data Monitoring Committee without unblinding individuals who should remain blinded. In the context of a clinical trial, blinding refers to knowledge of which treatment group a participant was assigned to.
- **Stopping rule:** GSDs can allow for efficacy stopping (early rejection of $H_0$) as well as futility stopping (early acceptance of $H_0$). During the design stage, a set of critical values known as stopping boundaries is calculated. At each interim analysis, the test statistic is compared with the critical values for that look. If the statistic is more extreme than the efficacy critical value, we say that it has crossed the efficacy boundary and the trial is stopped for treatment efficacy. If the statistic is less extreme than the futility critical value, we say that it has crossed the futility boundary and the trial is stopped for futility. Futility bounds can be either binding or nonbinding. If a study with binding futility bounds is not stopped after crossing the futility bound, it risks overrunning the desired type I error. Nonbinding futility bounds are similar to binding futility bounds, but if a nonbinding futility bound is crossed, investigators have the option of stopping for futility or continuing the trial in the hope that more evidence will accumulate in favor of the experimental treatment, and there is no risk of excessive type I error. The cost for nonbinding futility bounds is a slightly larger sample size than required by binding futility bounds with equivalent type I error and power.
- **Expected sample size:** If a group sequential trial, a clinical trial using a GSD, stops early, it can use a substantially smaller sample size than an equivalently powered FSD. But if all the interim tests are inconclusive, the study will continue to the final look and use the maximum possible sample size, which is always larger than that of an equivalent FSD. The expected sample size of a GSD is the average sample size that would be used if the trial were to be repeated many times. Expected sample size is calculated relative to a given effect size; the expected sample size of a GSD with efficacy stopping will decrease when the effect size is large, reflecting the increased probability of early stopping for efficacy. GSDs with futility stopping will often have a smaller expected sample size than an equivalent FSD when the effect size is 0 because of the ability to accept the null hypothesis and stop the trial early for futility.

- **Boundary-calculation procedure:** Frequently, an investigator will decide which stopping rule to employ (efficacy stopping, futility stopping, or both) and how many interim looks to perform before picking the procedure they will use to calculate the bounds. There are several different formulas available to calculate stopping boundaries, but the most popular ones fall into two broad categories: classical bounds and error-spending bounds. Classical boundary-calculation procedures compute the boundary critical values directly, while error-spending procedures define an error-spending function that partitions the type I error (for efficacy bounds) or type II error (for futility bounds) between the planned looks. For designs with both efficacy and futility stopping, it is not necessary to use the same boundary-calculation method for both efficacy and futility bounds, but classical bounds cannot be combined with error-spending bounds.

  Some boundary-calculation procedures are conservative, which means they offer little chance of early stopping unless the true effect size is quite large (in the case of efficacy bounds) or quite small (in the case of futility bounds). A trial employing a conservative bound is more likely to continue to the final look, yielding an expected sample size that is not dramatically smaller than the sample size required by an equivalent fixed-sample trial. However, the maximum sample size (that is, the sample size at the final look) of a trial with a conservative bound is generally not much greater than the sample size required by an equivalent fixed trial. Another direct result of specifying conservative bounds is that the critical value at the final look tends to be close to the critical value employed by an equivalent FSD. In contrast, anticonservative boundaries offer a much better shot at early stopping (often yielding a small expected sample size) at the cost of a larger maximum sample size and final critical values that are considerably larger than the critical value of an equivalent FSD. Other boundary-calculation procedures use a parameter to control their shape; depending on the value of the parameter, these bounds can be conservative, anticonservative, or somewhere in the middle.

- **Information:** The amount of information a dataset contains about an unknown parameter is known as the Fisher information. Generally, information is proportional to the sample size, but not always. For example, with time-to-event data, the amount of information is proportional to the number of events observed. When designing a group sequential trial, the timing of the interim looks is specified in terms of the information fraction, the fraction of the maximum possible information to be collected by the study. For example, in a GSD with four equally spaced looks, analyses will occur when 25%, 50%, 75%, and 100% of the data are collected. If a group sequential trial continues to its final look, the maximum amount of information will be collected, which is always greater than the information of an equivalently powered FSD. The ratio of the maximum information required by a GSD to the information of an equivalent FSD is known as the information ratio, and the maximum sample size (or number of events) of a GSD is the product of the information ratio and the sample size of an equivalent FSD.

## Origins of GSD

Clinical trials are studies investigating the effects of a treatment on human participants. Large clinical trials, such as those designed to determine the efficacy of an experimental treatment, typically enroll participants over months or years, randomizing some participants to the experimental treatment group and others to a control group. In an FSD, no analysis is conducted until all data are collected.

In the context of a clinical trial, there is an ethical imperative not to expose participants to inferior treatments. GSDs address this ethical consideration by providing a protocol for the interim analyses of clinical trial data. If an interim analysis demonstrates that the new treatment is effective, the trial can stop early, hastening regulatory approval and sparing future participants from being assigned to the control group. If an interim analysis demonstrates that the new treatment is ineffective, the trial can stop early and resources can be allocated to testing more promising treatments.

When done naïvely, conducting multiple analyses at a nominal significance level will inflate type I error. Wassmer and Brannath (2016) note that traditional methods of controlling the familywise error rate, such as the Bonferroni correction, are overly conservative because they do not exploit the covariance structure of test statistics from a sequential analysis. Wallis (1980) recounts the origin of modern sequential analysis theory, which arose not in the context of clinical trials, but as a more efficient way to test weaponry during the Second World War. Abraham Wald, a Hungarian Jewish mathematician who immigrated to the United States and participated in the war effort as a member of the Statistical Research Group at Columbia University, developed the sequential probability ratio test (SPRT) in 1943. (Wald is known for several contributions to statistics, including the eponymous Wald test; see the vignette in the [TS] **varwle** entry for more information about his life.)

The SPRT was so useful to the military that access to Wald's report was restricted to prevent it from falling into enemy hands. Two years later, the restriction was lifted and Wald (1945) published the first public account of the SPRT, which uses fixed stopping rules but does not fix the maximum sample size. Applying the Neyman–Pearson theory of hypothesis testing, practitioners of the SPRT begin by formulating two hypotheses, $H_0$ and $H_a$, which are compared using a sequence of likelihood-ratio tests. A continuation interval $(a, b)$ is defined, with critical values $a$ and $b$ chosen so that the probabilities of type I and type II errors are equal to prespecified levels. After each sample is collected, the investigator calculates the likelihood ratio of the two hypotheses; if it falls within the continuation interval, the experiment continues and another sample is taken. If the likelihood ratio lies outside $(a, b)$, the experiment ends and either $H_0$ or $H_a$ is rejected (depending on whether the likelihood ratio is above or below the continuation interval). In contrast with the prevailing modern interpretation of null-hypothesis significance testing, the SPRT provides a mechanism to reject $H_a$ and accept $H_0$.

## Brief overview of GSD

More recent developments in sequential experimental design have introduced classes of sequential tests with different properties, including a fixed maximum sample size. But the appeal of a controlled framework for accepting $H_0$ has endured in sequential experimental designs, in no small part because accepting $H_0$ provides grounds for terminating the experiment due to futility. The term "accept $H_0$" is widely used in literature about sequential clinical trials, and we will use it (without quotes) in the remainder of this manual to refer to the demonstration of futility in a sequential design. The complement of futility stopping is efficacy stopping, where the experiment is terminated because $H_0$ can be rejected, even if the maximum sample size has not been reached.

In most clinical trials, it is not feasible to perform statistical analysis after each sample is collected, so fully sequential designs are rare in practice. GSDs address this logistical challenge by scheduling interim analyses after groups of samples have been collected. A major advance in GSDs came when Pocock (1977) established clear guidelines for calculating efficacy stopping boundaries that attain desired levels of type I and type II errors.

Pocock published critical values for a test statistic that follows a standard normal distribution under $H_0$. For test statistics following other distributions, Pocock recommends using a critical value that has an equivalent significance level to the published $z$ score. For a demonstration, see example 2 in [ADAPT] **gsdesign onemean**.

While Pocock's boundaries use the same critical value at each interim look, O'Brien and Fleming (1979) introduced group sequential boundaries with critical values that are conservative for early looks and less so as more data are collected. O'Brien–Fleming boundaries have proven popular among researchers who are wary of stopping a trial very early for anything less than the strongest evidence, but who appreciate the smaller maximum sample size and final-look critical values compared with those of Pocock's boundaries.

Wang and Tsiatis (1987) developed a one-parameter family of boundaries that includes the Pocock and O'Brien–Fleming boundaries as special cases. Wang–Tsiatis bounds are popular with researchers who want a boundary that is less conservative than O'Brien–Fleming bounds but more conservative than Pocock bounds. However, it is possible to select values of the Wang–Tsiatis parameter that create bounds that are more conservative than the O'Brien–Fleming bound or more anticonservative than the Pocock bound.

Lan and DeMets (1983) introduced the error-spending approach to constructing stopping boundaries. This approach controls the overall probability of type I error by "spending" error probability at interim looks. This allows the number and timing of interim looks to be updated while the trial is in progress.

Lan and DeMets (1983) presented error-spending functions that correspond to boundaries that approximate both Pocock and O'Brien–Fleming bounds. Kim and DeMets (1987) created a useful family of error-spending functions indexed by a power parameter, and Hwang, Shih, and de Cani (1990) introduced another one-parameter family of error-spending functions. While the parameters for Kim–DeMets and Hwang–Shih–de Cani bounds use different scales, both boundary-calculation procedures are quite flexible and can produce bounds that are as conservative or anticonservative as desired.

The process of conducting interim analyses with a GSD is the same regardless of the procedure used to calculate the stopping boundaries. The boundaries comprise a series of critical values, one for each look. At each interim look, the data are analyzed and a test statistic is calculated. If the design includes efficacy bounds, the test statistic is compared with the efficacy critical value, and $H_0$ is rejected if the statistic is more extreme than the efficacy critical value. If the design includes futility bounds, the test statistic is compared with the futility critical value, and $H_0$ is accepted if the statistic is less extreme than the futility critical value.

## Graphing group sequential boundaries

When comparing different GSDs, it is often helpful to visualize the boundaries of different methods. We begin by presenting a simple GSD using O'Brien–Fleming efficacy boundaries for a two-sided test of means in figure 3 below. Here we plan on conducting up to five analyses: four interim looks and one final analysis. At each interim look, if the test statistic calculated from the available data is within the green continuation region, then the study continues accruing more participants, but if the test statistic is outside the efficacy bounds and in the blue rejection region, then $H_0$ is rejected and the experiment is stopped early for efficacy. At the fifth and final look, there is no continuation region; if the final test statistic is not outside the efficacy bounds, it will lie in the red acceptance region and $H_0$ is accepted.

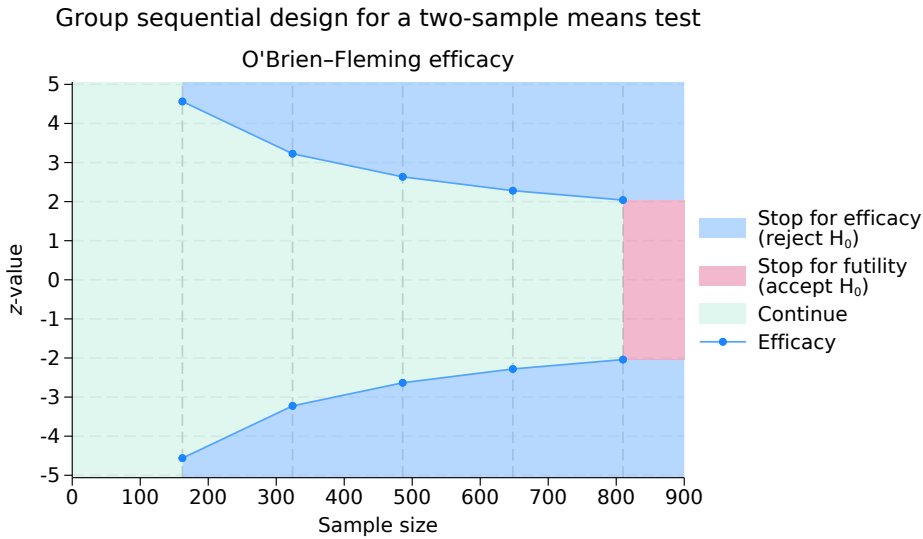


Figure 3. Two-sided O'Brien–Fleming efficacy bounds for a test of the equality of two means

Next we consider a similar scenario that includes futility bounds as well as efficacy bounds. Efficacy bounds separate the rejection region from the continuation region, and futility bounds separate the acceptance region from the continuation region. If the test statistic from an interim analysis falls within the continuation region, then the study proceeds as planned. If it falls within the rejection region, then $H_0$ is rejected and the study is terminated due to treatment efficacy. If the test statistic lies within the acceptance region, then $H_0$ is accepted and the study is terminated due to futility. As in the previous example, at the final look, there is no continuation region and $H_0$ must be accepted or rejected.

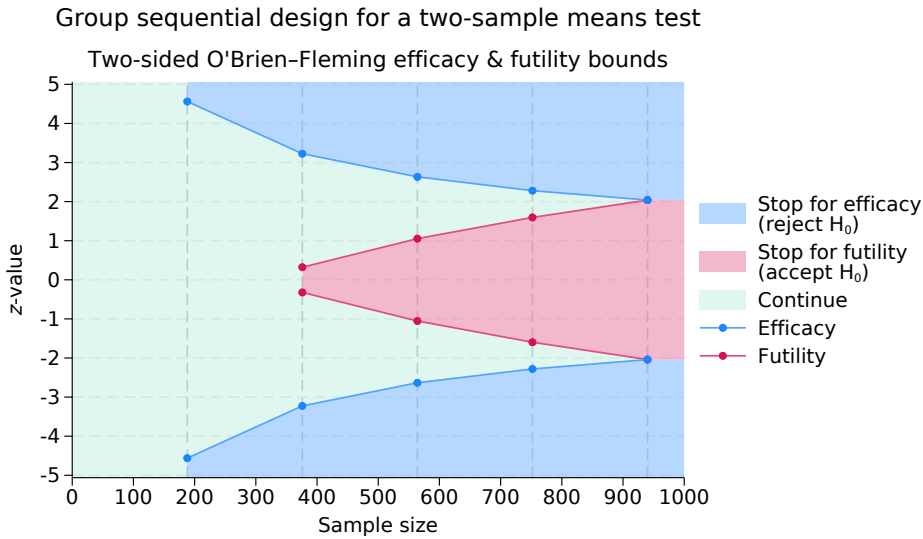


Figure 4. Two-sided O'Brien–Fleming efficacy and futility bounds

In both graphs, the vertical axis is labeled "$z$-value" because the theory underlying a GSD's ability to control familywise type I error is based on a sequence of test statistics whose marginal distribution under the null hypothesis is normal with a mean of 0 and a variance of 1. For details about how to calculate group sequential boundaries in Stata, including how to incorporate test statistics that are not normally distributed, see [ADAPT] **gsbounds**. To additionally calculate sample sizes for interim analyses, see [ADAPT] **gsdesign**.

## References

Hwang, I. K., W. J. Shih, and J. S. de Cani. 1990. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* 9: 1439–1445. https://doi.org/10.1002/sim.4780091207.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC.

Kim, K., and D. L. DeMets. 1987. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 74: 149–154. https://doi.org/10.1093/biomet/74.1.149.

Lan, K. K. G., and D. L. DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70: 659–663. https://doi.org/10.1093/biomet/70.3.659.

O'Brien, P. C., and T. R. Fleming. 1979. A multiple testing procedure for clinical trials. *Biometrics* 35: 549–556. https://doi.org/10.2307/2530245.

Pocock, S. J. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64: 191–199. https://doi.org/10.1093/biomet/64.2.191.

Wald, A. 1945. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 16: 117–186. https://doi.org/10.1214/aoms/1177731118.

Wallis, W. A. 1980. The statistical research group, 1942–1945. *Journal of the American Statistical Association* 75: 320–330. https://doi.org/10.2307/2287451.

Wang, S. K., and A. A. Tsiatis. 1987. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43: 193–199. https://doi.org/10.2307/2531959.

Wassmer, G., and W. Brannath. 2016. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham, Switzerland: Springer.

## Also see

[ADAPT] **Intro** — Introduction to adaptive designs for clinical trials

[ADAPT] **gs** — Introduction to commands for group sequential design

[ADAPT] **Glossary**