Group Sequential Designs: A Tutorial

Daniël Lakens[1], Friedrich Pahlke[2], & Gernot Wassmer[2]

[1] Eindhoven University of Technology
[2] RPACT GbR

This tutorial illustrates how to design, analyze, and report group sequential designs. In these designs, groups of observations are collected and repeatedly analyzed, while controlling error rates. Compared to a fixed sample size design, where data is analyzed only once, group sequential designs offer the possibility to stop the study at interim looks at the data either for efficacy or futility. Hence, they provide greater flexibility and are more efficient in the sense that due to early stopping the expected sample size is smaller as compared to the sample size in the design with no interim look. In this tutorial we illustrate how to use the R package 'rpact' and the associated Shiny app to design studies that control the Type I error rate when repeatedly analyzing data, even when neither the number of looks at the data, nor the exact timing of looks at the data, is specified. Specifically for *t*-tests, we illustrate how to perform an a-priori power analysis for group sequential designs, and explain how to stop the data collection for futility by rejecting the presence of an effect of interest based on a beta-spending function. Finally, we discuss how to report adjusted effect size estimates and confidence intervals. The recent availability of accessible software such as 'rpact' makes it possible for psychologists to benefit from the efficiency gains provided by group sequential designs.

Keywords: group sequential designs; sequential analyses; hypothesis testing; error control; power analysis
Word count: 8785

Sequential analyses are a more efficient approach to hypothesis testing than analyzing data only once after the maximum sample size has been collected. The reason is that the sample size of the study is expected to be smaller under many circumstances due to the possibility of rejecting the null hypothesis or stopping the study for futility at an interim look. Even though all researchers have resource constraints, and despite the popularity of sequential designs in fields such as medicine, most researchers in other disciplines are not yet benefiting from the efficiency gains that sequential designs provide. One possible reason for this slow adoption of a statistical technique that saves substantial resources might have been a lack of statistical software tools to easily design a study that uses sequential analyses (Albers, 2019). In recent years, however, R packages have been created that make it relatively easy for all researchers to plan and analyze group sequential designs. In this tutorial, we will illustrate how to design and report a sequential analysis with easy to use statistical software.

Sequential analyses have a long history. Dodge and Romig (1929) already realized that analyzing data sequentially was more efficient than doing so only once. Wald popularized the idea of sequential tests of hypotheses during the second world war. He was only allowed to publish his findings after the war had ended (Wald, 1945), as he explains in a historical note (p. 121):

> *Because of the substantial savings in the expected number of observations effected by the sequential probability ratio test, and because of the simplicity of this test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy, at least for a certain period of time. The author was, therefore, requested to submit his findings in a restricted report which was dated September, 1943.*

In other words, sequential analyses provide such an increased efficiency when testing hypotheses that the technique could not be shared with the enemy during the war - yet few researchers outside of medicine currently use it, even though they can. Sequential analyses are established procedures, and have been developed in great detail over the last decades

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research. Gernot Wassmer and Friedrich Pahlke are the owners of the company RPACT GbR, Germany. We would like to thank Isabella R. Ghement for feedback on an earlier version. A step-by-step vignette for all tests reporting in this article, and the computationally reproducible version of this manuscript, are available at https://github.com/Lakens/sequential_tutorial.

Correspondence concerning this article should be addressed to Daniël Lakens, Den Dolech 1, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

(Jennison & Turnbull, 2000; Proschan, Lan, & Wittes, 2006; Wassmer & Brannath, 2016). Just as in designs where the final sample size is fixed in advance (i.e., *fixed designs*), researchers specify their Type I and Type II error rates and design a well-powered study where the error rates are controlled. As opposed to fixed designs, researchers have the freedom to choose how they will spend their error rate across multiple looks at the data.

In a sequential design, researchers plan to collect data, analyze the results, and depending on whether or not the results are statistically significant, they will continue the data collection. If one would analyze the data repeatedly as data comes in without correcting the alpha level (or significance level), the Type I error rate will substantially inflate (Armitage, McPherson, & Rowe, 1969). For example, as Armitage and colleagues show, when analyzing results 5 times (when collecting the same number of observations for each look at the data) the alpha level inflates to 0.142 instead of 0.05. Researchers admit to having used *optional stopping* without adjusting the alpha level (Fiedler & Schwarz, 2016; John, Loewenstein, & Prelec, 2012). The idea that repeatedly analyzing data increases the efficiency of data collection is correct, but the past practice to use optional stopping is flawed. When done right, sequential designs give researchers more flexibility and increase the efficiency of data collection while controlling Type I and Type II error rates.

When moving from a fixed design to a sequential design researchers need to make several decisions. The first decision is how the error rates should be controlled across multiple looks at the data. The second decision is whether the data collection is stopped when there is no support for the alternative hypothesis, and if so, which rule is used to make such a decision. After making both decisions it is possible to perform an a-priori power analysis to determine the sample size required for the sequential design. It is then straightforward to compute the average sample size that one can expect to collect, based on the statistical power at each look, and the probability that the data collection will stop at each look. Finally, at the last look researchers can report an effect size estimate, confidence interval, and p value that takes the sequential nature of the design into account. In this tutorial we explain each of these steps. A detailed step-by-step vignette that illustrates how to perform all calculations presented in this tutorial both in `rpact` as in the online Shiny app is available at https://www.rpact.org/vignettes/step-by-step_tutorial.

Type I Error Control in Sequential Designs

Imagine a researcher who will collect data from at most 188 participants and plans to analyze the data after 66, 132, and 188 participants are recruited into the study. Each analysis is called a look. In this example, the looks at the data are spaced equally, which means they occur after the same number of participants have been collected between each look. There are in total $K = 3$ looks at the data, with two interim analyses, and one final analysis (so tests are performed after 33.3%, 66.7%, and 100% of the data is collected). Not all looks have to occur in practice. If the analysis reveals a statistically significant result after look 1, data collection can be terminated. It is also possible to stop the data collection at look 1 because the observed effect is much smaller than the effect size that the study was designed to detect, and the presence of effects as large or larger than the effect size of interest can be rejected. This is called stopping for futility.

The solution to controlling the Type I error inflation in sequential analyses is conceptually similar to a multiple comparison problem. Because multiple tests are performed, the Type I error rates inflates. By lowering the alpha level at each look, the overall Type I error rate can be controlled. This is much like a Bonferroni correction, and indeed, the Bonferroni correction is a valid (but conservative) approach to control the error rate in sequential analyses (Wassmer & Brannath, 2016). If researchers do not have the time to learn about sequential designs, they can simply preregister studies where they plan to perform a Bonferroni correction for each look. However, a better understanding of group sequential designs will provide researchers with a lot more flexibility and increased efficiency.

In sequential designs the analysis at each look includes all data collected thus far. At look 2 we combine the data collected at look 1 with the new data. This implies that the statistical tests at each looks are not independent. Instead, sequential tests for a continuous endpoint have a well-known multivariate normal distribution with calculable probabilities. This makes it possible to derive adjusted alpha levels that control the Type I error rate more efficiently than a Bonferroni correction. The simplest way to correct the alpha level is the Pocock correction (Pocock, 1977), where an alpha level is chosen that is the same for each look. For example, if we want an overall Type I error rate of 5%[1] for a two-sided test with 1 interim analysis (and one final look), the alpha level for each look would be 0.0294, for three looks in total the alpha level would be 0.0221, for four looks it would be 0.0182, and for five looks it would be 0.0158[2]. We see the correction is slightly more efficient than using a Bonferroni correction (in which case the alpha levels would be 0.025, 0.0167, 0.0125, and 0.01, respectively). Applying the Pocock procedure in this

[1]The alpha level for any test should be carefully justified. In this tutorial we use an alpha level of 0.05 throughout to prevent confusion between the chosen alpha level and adjusted alpha levels at different looks.

[2]Corrected alpha levels can be computed to many digits, but this quickly reaches a level of precision that is meaningless in real life. The observed Type I error rate for all tests you will do in your lifetime is not noticeably different if you set the alpha level at 0.0158 or 0.016.

way requires 1) specifying the number of looks in advance, and 2) equally spaced looks. This second requirement can be weakened by requiring that the number of subjects per look is fixed in advance and cannot be changed during the course of the study.

In this tutorial the R (R Core Team, 2020) package `rpact` will be used (Wassmer & Pahlke, 2020). Many of the options in the package are also available in an online Shiny (Chang, Cheng, Allaire, Xie, & McPherson, 2020) app: https://shiny.rpact.com. As Figure 1 shows researchers can specify the number of looks, whether they plan to perform a one-sided or two-sided test, the overall alpha level, and the type of design (e.g., a design using the Pocock correction). The `rpact` Shiny app computes the significance level for each look (reported as the "Two-sided local significance level" in Figure 2) and provides the R code to reproduce the computations (see Figure 3). In addition, the Shiny app can be used to generate plots, tables, and reports in Rmd, pdf, html, and xlsx formats.

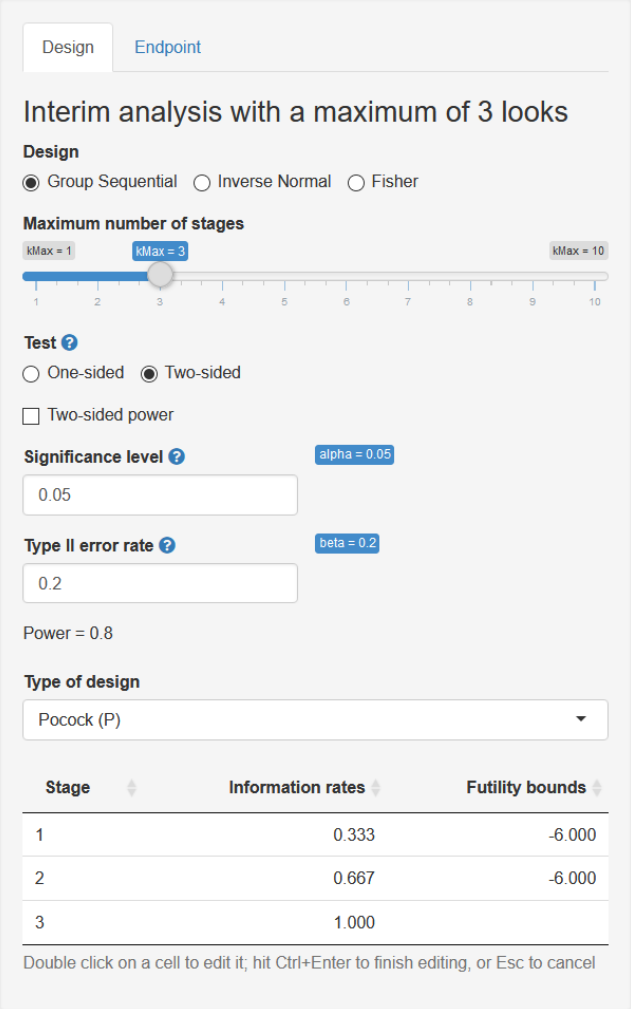

Figure 1. Screenshot of the design tab of the rpact Shiny app.

```
Sequential analysis with a maximum of 3 looks
(group sequential design)

Pocock design, two-sided local significance level 5%,
power 80%, undefined endpoint.

Stage                                    1      2      3
Information rate                     33.3%  66.7%   100%
Efficacy boundary (z-value scale)    2.289  2.289  2.289
Cumulative alpha spent              0.0221 0.0379 0.0500
Two-sided local significance level  0.0221 0.0221 0.0221
```

Figure 2. Summary output of the rpact Shiny app for the design specified in Figure 1.

R Command

```
design <- getDesignGroupSequential(typeOfDesign = "P",
alpha = 0.05, sided = 2)
summary(design)
```

Figure 3. R command output of the rpact Shiny app.

Comparing Corrections

The Pocock correction is one way to control the alpha level across all looks such that the alpha level is identical for each look at the data, resulting in constant critical values (expressed as z values) $u_k = c$ to reject the null hypothesis, H_0, at look k. Other solutions distribute the Type I error rate across looks in different ways. For example, the O'Brien and Fleming correction (O'Brien & Fleming, 1979) uses monotonically decreasing values $u_k = c/\sqrt{k}$, i.e., as c is found (through a numerical algorithm), the critical values u_k can be derived. Here, a higher threshold for early looks is foreseen, but the final look occurs quite close to the uncorrected alpha level (see Figure 4 for a visualization of the critical z values for three equally spaced looks for different ways to spend the Type I error rate across looks). The Haybittle and Peto correction simply suggests to use a critical value of $u = 3$ for each look but the last, u_K, which is corrected to control the overall error rate (and is very close to the uncorrected critical value).

Because the statistical power of a test depends on the alpha level (and the effect size and the sample size), and because the alpha level for an O'Brien and Fleming or a Haybittle and Peto design is close to alpha level for a fixed design with only one look, at the final look the statistical power of these designs is also similar to the power for a fixed design. However, there is a reasonable chance (especially for the O'Brien and Fleming design) to stop the study early and thereby gain efficiency. If the alpha level for the final look is lowered, the sample size of a study needs to be increased to maintain the same statistical power at the last look. Because the Pocock correction leads to a lower alpha level at the last look, this design requires a larger increase in the maximum sample size than the O'Brien and

Fleming or the Haybittle and Peto correction. This increase in sample size when using the Pocock correction is compensated by an increased chance to stop the data collection at an earlier look. We will discuss these issues in more detail in the section on sample size planning.

Another approach to controlling Type I error rates is provided by Wang and Tsiatis (1987), where a power parameter Δ is specified which determines the shape of the critical values over the looks as given by $u_k = k^{\Delta-0.5}$. With $\Delta = 0$ the Wang and Tsiatis correction equals the O'Brien and Fleming correction, and with $\Delta = 0.5$ the Wang and Tsiatis correction equals the Pocock correction. For $0 < \Delta < 0.5$ it is possible to select the shape of decision boundaries somewhere in between these two corrections.

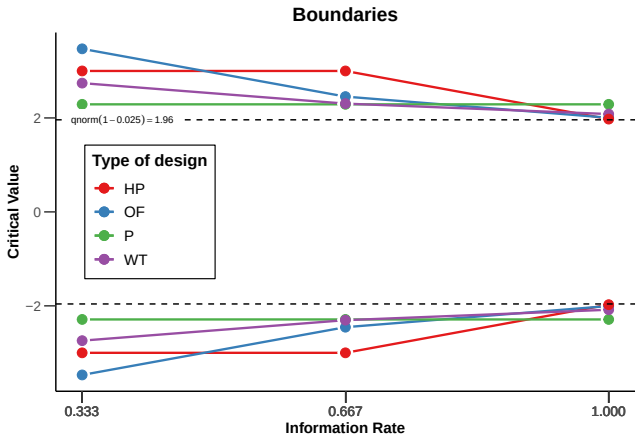

Figure 4. Comparison of critical values (in z values) for the O'Brien and Fleming, Pocock, Wang and Tsiatis with $\Delta = 0.25$, and Haybittle-Peto corrections for 3 looks.

We can see that the O'Brien and Fleming correction is much more conservative at the first look, and close to the uncorrected critical value of 1.96 (the black dashed line - for two-sided tests all critical values are mirrored in the negative direction) at the last look: 3.471, 2.454, and 2.004. The Pocock correction has the same critical value at each look (2.289, 2.289, and 2.289), the Haybittle and Peto correction has the same critical value at each look but the last (3, 3, and 1.975), while the critical values decrease for each look with the Wang and Tsiatis correction (2.741, 2.305, and 2.083).

The Alpha Spending Approach

The approaches to specify the shape of decision boundaries across looks discussed so far have an important limitation (Proschan et al., 2006). They require a pre-specified number of looks (e.g., 4), and the sample size for the interim looks need to be pre-specified as well (e.g., after 25%, 50%, 75%, and 100% of observations). It is logistically not always feasible to stop the data collection exactly at 25% of the planned total sample size. For example, due to inclusion criteria (such as a minimum accuracy level on an experimental task) some observations will excluded from the final analysis, but the number of observations that will be excluded is not known until the data is analyzed. An important contribution to the sequential testing literature was made by Lan and DeMets (1983) who introduced the alpha spending approach. In this approach the cumulative Type I error rate spent across the looks is pre-specified through a function (the *alpha spending function*) in order to yield overall significance level α at the end of the study. For the O'Brien and Fleming approach only a small amount of the Type I error rate is spent for early looks and more is spent on later looks, while for a Pocock design a more uniform spending of α over the looks is anticipated. The idea is of the alpha spending approach is to calculate the critical values (and correspondingly, the alpha levels) at each look according to the prespecified alpha spending function that depends on the fraction of information available at some time point, t, of the study. The monotone increasing alpha spending function is 0 at the beginning of the study ($t = 0$) and it is α at the end ($t = 1$). Lan and Demets showed that the critical values derived from the two functions

$$\alpha_1^*(t) = \begin{cases} 2\,(1 - \Phi(\Phi^{-1}(1 - \alpha/2)/\sqrt{t})) & \text{(one-sided case)} \\ 4\,(1 - \Phi(\Phi^{-1}(1 - \alpha/4)/\sqrt{t})) & \text{(two-sided case)} \end{cases}$$

and

$$\alpha_2^*(t) = \alpha \ln(1 + (e - 1)t)$$

approximate Pocock's and O'Brien and Fleming's group sequential boundaries, respectively, in case of equally sized looks. In Figure 5 the O'Brien and Fleming-like critical values derived from the alpha spending function $\alpha_1^*(t)$ is plotted against the discrete O'Brien and Fleming bounds. We see that the two approaches are not identical, but they are very comparable. Similarly, we see that for Pocock's design, $\alpha_2^*(t)$ yields nearly constant critical values which are very similar to the Pocock correction.

The main benefit of these spending functions is that error rates at interim analyses can be controlled, while neither the number nor the timing of the looks needs to be specified in advance.[3] This makes alpha spending approaches much more flexible than earlier approaches to controlling the Type 1 error in group sequential designs. When using an alpha spending function it is important that the decision to perform an interim analysis is not based on collected data, as this can still increase the Type I error rate. As long as this assumption is met, it is possible to update the alpha levels at each look during a study.

[3]Previous articles on sequential analyses such as Schnuerch and Erdfelder (2020) and Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017) have incorrectly claimed that the number of looks in group sequential designs needs to be specified in advance.

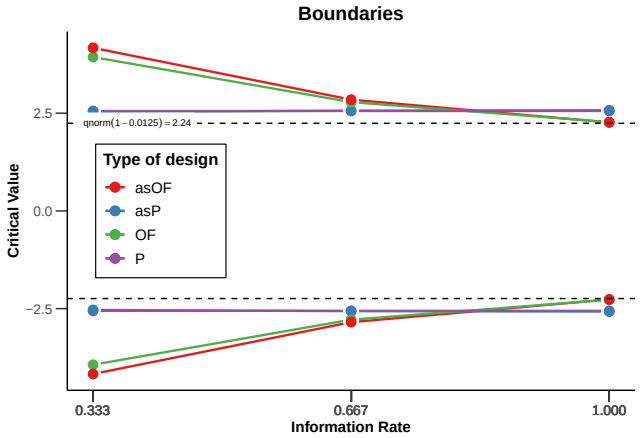

Figure 5. Comparison of the Pocock (P) and O'Brien and Fleming (OF) corrections with the critical values derived from their corresponding alpha spending functions (asP and asOF) for a two-sided test with three looks.

Updating Boundaries During a Study

Although alpha spending functions control the Type I error rate even when there are deviations from the pre-planned number of looks, or their timing, this does require recalculating the boundaries used in the statistical test based on the amount of information that has been observed. Let us assume a researcher designs a study with three equally spaced looks at the data ($K = 3$), using a Pocock-type alpha spending function $\alpha_2^*(t)$, where results will be analyzed in a two-sided t-test with an overall desired Type I error rate of 0.05, and a desired power of 0.9 for a Cohen's d of 0.5. An a-priori power analysis shows that we achieve the desired power in our sequential design if we plan to look after 33, 66, and 99 observations in each condition.

Now imagine that due to logistical issues, we do not have the ability to analyze the data until we have collected data from 38 instead of 33 observations per condition. So our first look at the data does not occur at 33.33% of planned sample, but at 76 /198 = 38.38% of the planned sample. We can recalculate the alpha levels we should use for each look at the data, based on the current look, and planned future looks. Instead of using the alpha levels 0.0226, 0.0217, and 0.0217 at the three respective looks, the updated alpha levels are 0.0253 for the current look, 0.0204 for the second look, and 0.0216 for the final look.

It is also possible to correct the alpha level if the final look at the data changes, for example because a researcher is not able to collect the intended sample size, or because, due to unforeseen circumstances, a researcher collects more data than originally planned. If this happens, the Pocock-like alpha spending function that was intended to be used no longer applies. Instead, a user-defined alpha spending function needs to be provided by updating the timing and alpha spending function to reflect the data collection as it actually occurred up to the final look. Assuming the second look in our earlier example occurred as planned, but the last look occurred at 206 instead of 198, we can compute an updated alpha level for the last look, which is 0.0210 instead of 0.0216. In this case the difference is small, but it demonstrates the flexibility alpha spending functions provide.

Whenever researchers expect there is a realistic possibility that the data collection at each look might deviate from the planned sample size, it is advisable to use an alpha-spending function to control the Type I error rate, as it provides a lot more flexibility. Note that if the number of looks was not pre-planned, we would use similar calculations at each look as in the example above, but in addition to updating the timing, we would update the number of looks in our design (see Proschan et al., 2006, chapter 5).

Stopping for Futility

After performing an interim analysis, it could be impossible for the additional data a researcher planned to collect to lead to a statistically significant result. To illustrate this in a hypothetical scenario, imagine that after collecting 182 out of 192 observations, the observed mean difference between two independent conditions is 0.1, while the study was designed with the idea that the smallest effect deemed worthwhile is a mean difference of 0.5. If the primary dependent variable is measured on a 7 point Likert scale, it might be that even if every of the remaining 5 participants in the control condition answers 1, and every of the remaining participants in the experimental condition answers 7, the effect size after 192 observations will not yield statistical significance. If the goal of your study was to detect whether there was an effect of at least a mean difference of 0.5, at this point a researcher knows for a fact that goal will not be reached. Stopping a study at an interim analysis because the final result cannot yield a significant effect is called *non-stochastic curtailment*.

In less extreme cases, it might still be possible for the study to observe a significant effect, but the probability might be too small. The probability of finding a significant result, given the data that have been observed up to an interim analysis, is called *conditional power*. Imagine that in the previous example, a researcher first looked at the data after they collected 64 observations. At this time, a mean difference of 0.1 is observed. Assume the population standard deviation is 1, and that the researcher was willing to collect 192 observations in total, as this yielded 90% power for the effect size of interest, a mean difference of 0.5. The conditional power, assuming the true mean difference is 0.5, is 0.68. Assuming the true mean difference is similar to the observed effect of 0.1, however, implies that the conditional power is only 0.05. At this point, one might start to worry that an effect of 0.5 is not likely.

One might choose to perform a conditional power analysis, and increase the sample size one is willing to collect, if smaller effects than a mean difference are still deemed interesting. Given that data has already been collected, it seems intuitive to perform the conditional power analysis not based on the effect size that was originally expected, but by updating the prior belief about the true effect size given the observed data. The similar Bayesian alternative is called *predictive power* (Spiegelhalter, Freedman, & Blackburn, 1986). It is possible to increase the final number of observations in sequential designs, but it is important to note that changing the sample size based on observed results at an interim look in general could lead to a substantial inflation of the Type I error rate, even when an alpha spending approach is used. For these situations, group sequential *adaptive* procedures have been developed. The description of these adaptive confirmatory designs, however, is outside the scope of this tutorial (cf., Wassmer & Brannath, 2016).

A better approach than designing a study based on the expected effect size might be to design an experiment by specifying the smallest effect size of interest (Lakens, 2014), and using this as the alternative hypothesis when designing the study. If a study is designed based on a smallest effect size of interest you will not need to increase the sample size if the effect is smaller than hoped for, because you know the study you have design will yield an informative result for the smallest effect size that is practically or theoretically meaningful. If a smallest effect size of interest is specified, it becomes possible to statistically reject the alternative hypothesis, and decide that the effect, if any, is too small to be of interest. In essence, one performs an equivalence or inferiority test (Lakens, Scheel, & Isager, 2018).

As an illustration of a simple stopping rule for futility, let us imagine a researcher who is willing to stop for futility because the observed effect size in a directional test is either zero, or in the opposite direction as was predicted. In Figure 6 the red line indicates critical values to declare a significant effect. If a z value larger than 2.141 is observed at the second interim analysis, we can stop the study and reject H_0. As illustrated by the blue line, if at the second interim analysis we observe a z value smaller than or equal to 0 (i.e., an effect of 0 or in the opposite direction of our prediction) we can stop the study for futility. Stopping for futility does not impact the Type I error rate of a study, as we are not claiming an effect is present. But if we decide to stop the study based on a decision rule of $z \leq 0$, it is possible that we inflate the Type II error rate, because in small samples with large variation an effect in the opposite direction might be observed, even if the directional prediction was correct.

It is therefore more desirable to directly control the Type II error rate across looks, just as we control the Type I error rate. To achieve this, futility bounds based on a beta-spending

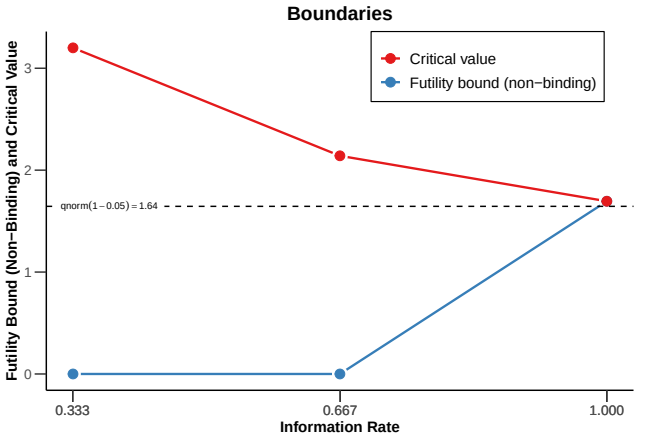

Figure 6. Pocock-type boundaries for 3 looks to stop when rejecting H_0 (red line) or to stop for futility (blue line) when the observed effect is in the opposite direction.

function can be computed. At the last look in our sequential design, which was designed to have 90% power, a researcher is willing to act as if H_0 is true with a 10% error rate. The null and alternative hypothesis can be reversed, which transforms the same decision process into an equivalence test (Lakens et al., 2018). In this view, the goal is to test whether the presence of a meaningful effect can be rejected. For example, if the smallest effect size of interest is a mean difference of 0.5, and a researcher observes a mean difference that is surprisingly far away from 0.5, they might be able to statistically reject the presence of an effect of 0.5 or larger, and reject the presence of an effect large enough to care about.

In essence, in such an equivalence test the Type II error of the original null hypothesis significance test has now become the Type I error rate for an equivalence test (Jennison & Turnbull, 2000). Because the researcher has designed the null hypothesis significance test to have 90% power for a mean difference of 0.5, 10% of the time they would incorrectly decide to act as if an effect of at least 0.5 is absent. This is statistically identical to performing an equivalence test with an alpha level of 10%, and decide to act as if the presence of an effect at least as large as 0.5 can be rejected, which should also happen 10% of the time, in the long run. Although it has become accepted in the social sciences to design studies with strongly unbalanced Type I and Type II error rates due to the work of Cohen (1988), the suggestion to balance error rates, for example through a compromise power analysis, deserves more consideration when designing experiments (Erdfelder, Faul, & Buchner, 1996; Mudge, Baker, Edge, & Houlahan, 2012).

Just as the Type I error rate can be distributed across interim analyses, the Type II error rate can be distributed across looks. A researchers can decide to stop for futility when the presence

of an effect size of interest can be rejected, thereby controlling the probability of making a Type II error in the long run. Such designs are more efficient if there is no effect. One can choose in advance to stop data collection when the presence of the effect the study was designed to detect can be rejected (i.e., binding beta-spending), but it is typically recommended to allow the possibility to continue data collection (i.e., non-binding beta-spending). Adding futility bounds based on beta-spending functions to a sequential design reduces the power of the test, and increases the required sample size for a well-powered design, but this is on average compensated by the fact that studies stop earlier due to futility, which can make designs more efficient. In Figure 7, a Pocock type spending function is used for both the alpha-spending function (with a 5% Type I error rate) and the beta-spending function (with a 10% Type II error rate).

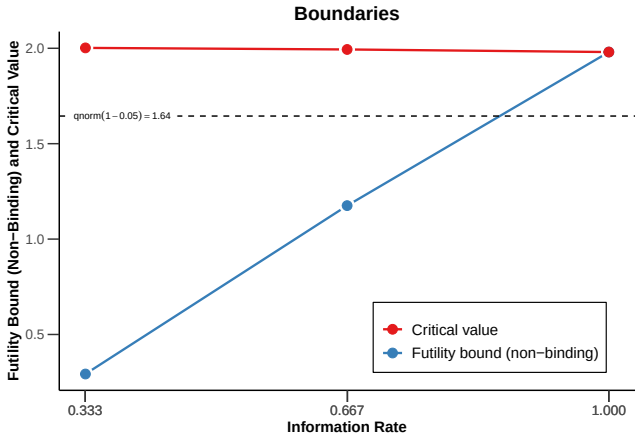

Figure 7. Pocock-type boundaries for 3 looks to stop when rejecting H_0 (red line) or to stop for futility (blue line) based on an beta-spending function.

Imagine a researcher has designed a study with 90% power for a mean difference of 0.5, an alpha level of 5%, and three equally spaced looks. An a-priori power analysis shows 192 observations are needed in total, with a first look after 64 observations. At look 1 a certain portion of the Type I and Type II error rates are spent, in such a way that the overall error rate is controlled across all looks. With a Pocock-type alpha spending function this means a researcher will spend 0.0226 of the 5% Type I error rate at look 1, but also spends 0.0453 of the Type II error rate at this look. Since this is a one-sided test, it means data collection will stop for futility if a $1 - 2 \cdot 0.0453 = 0.9094$ confidence interval around the observed mean excludes a mean difference of 0.5. This means that, at this first look, the presence of an effect of 0.5 can be rejected with a Type I error rate of 0.0453 in an equivalence test.

It is possible to calculate a critical mean difference that, at the first look, is the threshold to decide between stopping for futility or continuing the data collection. Data collection would be stopped for futility if the observed mean difference is smaller than 0.07. Data collection would also be stopped if H_0 can be rejected, based on $p < 0.0226$ (remember this is a one-sided test against H_0), which would happen whenever the observed mean difference is larger than 0.51, given the sample size at the first look. These critical mean differences to reject H_0 or reject the alternative hypothesis are typically expressed as test statistics, and therefore, data collection would stop for futility if the z value is smaller than 0.293, and stop to reject H_0 when the z value is larger than 2.002. These z values can be recalculated as p values. As mentioned previously, data collection would stop to reject H_0 when the p value is smaller than 0.0226, but it would also stop for futility when the p value for a one-sided test is larger than 0.3848.

Figure 8 visualizes the critical z values for futility and rejecting H_0 at the first look. For a one-sided test, we reject H_0 when the observed data is more extreme to the right of the null than the critical value indicated by the solid line, and we stop for futility when the observed data is further to the left than the critical value indicated by the dashed line. At future looks, more data has been collected, and therefore the confidence interval around the observed mean difference will be more narrow. Therefore, an observed a mean difference closer to 0.5 will be sufficient to stop for futility, and a mean difference closer to 0 will be sufficient to stop to reject H_0. At the third look, we always stop data collection based on the critical value that leads us to either reject H_0, or reject the alternative hypothesis, with the desired error rates. We see how specifying both the null and the alternative hypothesis, and designing a study that controls both the Type I and Type II error rates, will always lead to informative results, as long as the required number of observations indicated by the a-priori power analysis can be collected.

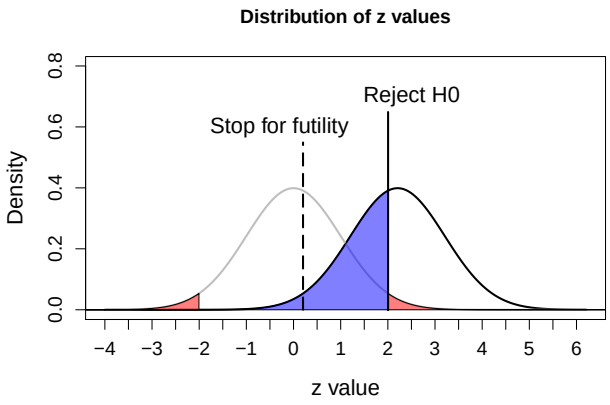

Figure 8. Null and alternative hypothesis (modelled as normal distributions with mean differences of 0 and 0.5) indicating critical Z-values for the first look of the design.

Stopping for futility has benefits and downsides (Schoenfeld & Meade, 2005). Unless you have determined a smallest effect size of interest, and are able to collect sufficient data to have high power to detect (or reject) the effect size of interest, you do not learn whether the effect is too small to be meaningful. Clearly, designing a study is most informative if we specify a smallest effect of interest, and stop data collection when we can either reject the null hypothesis or the alternative hypothesis. When specifying a smallest effect size of interest is not possible, researchers might not want to incorporate stopping for futility into the study design.

Sample Size Justifications for Sequential Designs

When designing a study with the goal to provide statistical support for the claim that a theorized effect is present or absent, it is essential to consider whether sufficient observations are collected to control the Type II error rate. An informative study has a high probability of correctly concluding an effect is present when it is present, and absent when it is absent. An a-priori power analysis is used to choose a sample size to achieve desired Type I and Type II error rates, in the long run, given assumptions about the null and alternative model. As noted previously, sequential designs can be more efficient than fixed designs, and an a-priori power analysis can be used to compute the expected average sample size, taking into account the probability that one can stop at an interim analysis.

In a group sequential design the alpha level at each look is lower than the desired overall Type I error rate. The lower the alpha level, the larger the number of observations that is needed to achieve the same power for the same effect size. Because sequential designs have a lower alpha level at the last look compared to a fixed design, an a-priori power analysis for a sequential design will show a larger number of observations is required at the last look of a sequential design, compared to a fixed design. The expected sample size is smaller for a sequential design, because there is a probability to stop at an interim analysis either for efficacy or futility. It is important to understand that the sample size is not always smaller. Instead, it is *expected to be smaller* for reasonable values of the alternative and H_0.

If the goal is to design a study that can detect a difference of 0.5, with an assumed standard deviation in the population of 1, which means the expected effect is a Cohen's d of 0.5, and a one-sided t test will be performed (given our directional prediction) with an overall alpha level to 0.05, and the desired Type II error probability is 0.1 (or a power of 0.9), 70 observations per condition should be collected, or 140 observations in total.

Instead of analyzing the data only once, a researcher could plan a group sequential design where the data will be analyzed three times, twice at interim analyses, and once at the final look. Compared to a fixed design, where the total number of required observations was 70 per condition, if a Pocock correction is chosen to control the Type I error rate for the sequential design, the maximum number of observations at the last look has increased to 81 per condition. The ratio of the maximum sample size for a sequential design and the sample size for a fixed design is known as the *inflation factor*, which is independent of the effect size. The inflation factor for the Pocock correction is larger than for the O'Brien and Fleming correction, for which the maximum sample size for a sequential design is only 71 per condition.

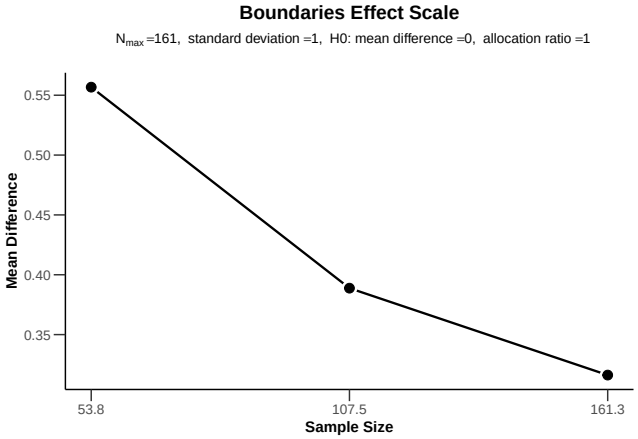

Figure 9. Critical raw effect size boundaries across three looks.

The inflation in the number of observations that are required to achieve the desired power when a sequential design is used is compensated by the fact that there is a probability that the data collection will stop at an interim look. This means that on average less observations will be collected in a sequential design, compared to a fixed design. Continuing the earlier example, if a Pocock correction is used for the three planned analyses, and under the assumption that the population standard deviation is 1, we can stop the data collection and reject H_0 if an effect size larger than 0.56, 0.39, and 0.32 is observed (see Figure 9). The probabilities of observing a mean difference larger than these values at each look is the statistical power, and assuming a true mean difference of 0.5 in the population, the probabilities we will stop at each of the three looks are 0.43, 0.32, and 0.15, respectively, for a total of 0.9 power across the three analyses. By multiplying the probability that we can reject H_0 at each look by the sample size we collect at each look, we can compute the average expected sample size for this sequential design, which is 97.8 if the alternative hypothesis is true. If H_0 is true, we only stop if we observe a Type I error, or at the end of the study, and the average sample size is 158.0. Note that the expected sample size is only dependent of the type of the group sequential design and hence, similar to the inlation factor, can be considered relative

of the fixed sample size.

Because power is a curve, and the true effect size is always unknown, it is useful to plot power across a range of possible effect size. This allows researchers to explore the expected sample size, in the long run, if a sequential design is used, for different true effect sizes (including a null effect). The blue line in Figure 10 indicates the expected number of observations that will be collected. Not surprisingly, when the true effect size is 0, data collection will almost always continue to the last look. Data collection will only stop if a Type I error is observed. The right side of the graph illustrates the scenario for when the true effect size is large, up to a mean difference of 1 (which given the standard deviation of 1, means Cohen's *d* is 1). With such a large effect size, the design will have high power at the first look, and data collection will almost always be able to stop at the first look.

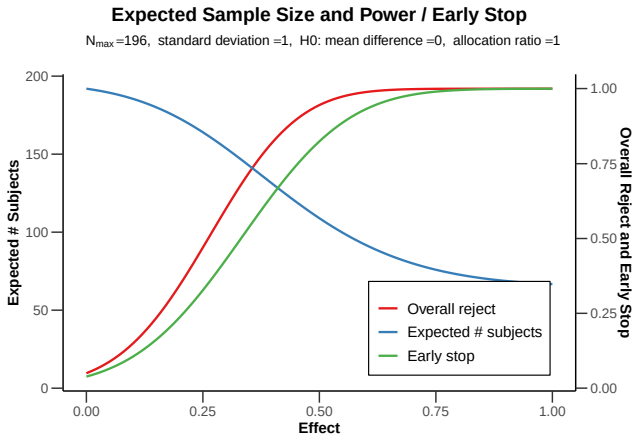

Figure 10. Overall rejection probability, probability of stopping at an interim analysis, and the expected number of observations for mean differences from 0 to 1 using a Pocock alpha-spending function for a one-sided test.

So far, the sequential design would only stop at an interim analysis if H_0 can be rejected. As discussed previously, it is also possible to stop for futility, for example, based on a beta-spending function. We can directly compare the previous design with a design where we stop for futility (see Figure 11).

If a researcher is willing to stop for futility, in addition to stopping when H_0 can be rejected, the probability that data collection will stop at an interim analysis increases (compare the green line in Figures 10 and 11), especially when H_0 is true or effect sizes are very small. This substantially reduces the expected average sample size whenever the effect size is small (see the blue line in Figures 10 and 11), where data collection is now terminated at an interim analysis because the effect size of interest (the alternative hypothesis) can be statistically rejected. On average the largest sample size needs to be collected when the true mean difference falls exactly

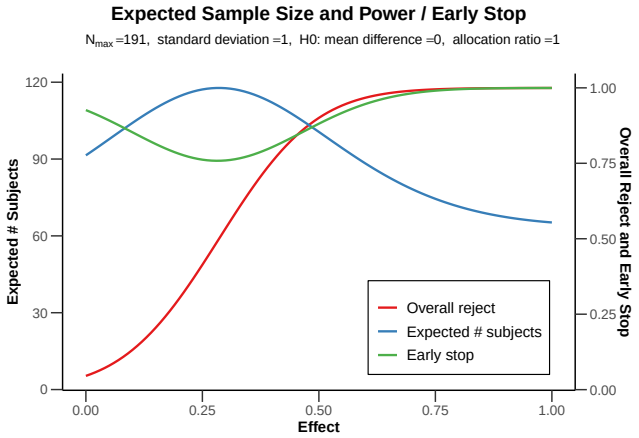

Figure 11. Overall rejection probability, probability of stopping at an interim analysis, and the expected number of observations for mean differences from 0 to 1 using a Pocock alpha-spending and beta-spending function for a one-sided test.

between the null hypothesis (d = 0) and the alternative (d = 0.5), which in this case is a mean difference of 0.25.

When deciding whether to use a group sequential design researchers should examine how the choice about how the Type I error rate is controlled across looks (and if relevant the Type II error rate), and the choice for the number of looks, together determine the maximum sample size one needs to collect, as well as the average expected sample size, depending on the effect size of interest. Although all examples in this tutorial have equally spaced looks, it is worth exploring designs with unequally spaced looks, especially when the statistical power at an early look (e.g., at 25% of the data for four equally spaced looks with an O'Brien-Fleming alpha spending function) would be very low. When deciding upon a specific sequential design, researchers should weigh both the benefits (e.g., a reduction in the average sample size that needs to be collected) and the costs (e.g., the logistics involved in analyzing the results and deciding whether or not to stop the data collection). The final decision about whether a group sequential design is a desirable option, and if so, which type of design fits the research question, depends both on logistics and resources.

Reporting results after a sequential analysis.

Group sequential designs have been developed to efficiently test hypotheses using the Neyman-Pearson approach for statistical inference, where the goal is to decide how to act, while controlling error rates in the long run. Group sequential designs do not have the goal to quantify the strength of evidence, or provide accurate estimates of the effect size (Proschan et al., 2006). Nevertheless, after having reached a conclusion

about whether a hypothesis can be rejected or not, researchers will often want to also interpret the effect size estimate when reporting results.

A challenge when interpreting the observed effect size in sequential designs is that whenever a study is stopped early when H_0 is rejected there is a risk that the data analysis was stopped because, due to random variation, a large effect size was observed at the time of the interim analysis. This means that the observed effect size at these interim analyses over-estimates the true effect size. As Schönbrodt et al. (2017) show, a meta-analysis of studies that used sequential designs will yield an accurate effect size, because studies that stop early have smaller sample sizes, and are weighed less, which is compensated by the smaller effect size estimates in those sequential studies that reach the final look, and are weighed more because of their larger sample size. However, researchers might want to interpret effect sizes from single studies before a meta-analysis can be performed, and in this case, reporting an adjusted effect size estimate can be useful. Although sequential analysis software only allows one to compute adjusted effect size estimates for certain statistical tests, we recommend to report both the adjusted effect size where possible, and to always also report the unadjusted effect size estimate for future meta-analyses.

A similar issue is at play when reporting p values and confidence intervals. When a sequential design is used, the distribution of a p value that does not account for the sequential nature of the design is no longer uniform when H_0 is true. A p value is the probability of observing a result *at least as extreme* as the result that was observed, given that H_0 is true. It is no longer straightforward to determine what 'at least as extreme' means a sequential design (Cook, 2002). The most widely recommended procedure to determine what "at least as extreme" means is to order the outcomes of a series of sequential analyses in terms of the look at which the study was stopped, where earlier stopping is more extreme than later stopping, and where studies with higher z values are more extreme, when different studies are stopped at the same time (Proschan et al., 2006). This is referred to as *stagewise ordering*, which treats rejections at earlier looks as stronger evidence against H_0 than rejections later in the study (Wassmer & Brannath, 2016). Given the direct relationship between a p value and a confidence interval, confidence intervals for sequential designs have also been developed.

Another way to derive confidence intervals and p values is through the concept of *repeated confidence intervals* proposed by Jennison & Turnbull (1989). Here, only the consequence of repeatedly looking at the data is accounted for, and the confidence levels are adjusted accordingly. It can be shown that the null hypothesis value lies within the repeated confidence intervals at a given look if and only if the null hypothesis cannot be rejected at this look. Conversely, all values lying outside the intervals can be rejected if formulated as a hypothesis. It is therefore straightforward to derive test decisions not only for superiority but also for non-inferiority or equivalence tests. *Repeated p-values* are derived by searching for the smallest overall significance level under which the observation would yield significance at a given look. The interpretation is similar: at a given look, the repeated p value is smaller than α if and only if H_0 can be rejected at this look. These confidence intervals and p values can be computed at any stage of the study irrespective of whether the study was stopped at an interim look at the data or not. This is in contrast to the confidence intervals that are based on the stagewise ordering: they can only be computed at the final look of the study. Note that confidence intervals provide additional information about the effect size estimate and are adjusted for the sequential nature of the design.

Adjusted p values and confidence intervals can usually be computed with existing software packages for sequential analyses. Reporting adjusted p values and confidence intervals, however, might be criticized. After a sequential design, a correct interpretation from a Neyman-Pearson framework is to conclude that H_0 is rejected, the alternative hypothesis is rejected, or that the results are inconclusive. The reason that adjusted p values are reported after sequential designs is to allow readers to interpret them as a measure of evidence. Dupont (1983) provides good arguments to doubt that adjusted p values provide a valid measure of the strength of evidence. Furthermore, a strict interpretation of the Neyman-Pearson approach to statistical inferences also provides an argument against interpreting p values as measures of evidence. Therefore, it is recommended, if researchers are interested in communicating the evidence in the data for H_0 relative to the alternative hypothesis, to report likelihoods or Bayes factors, which can always be reported and interpreted after the data collection has been completed. Adjusted confidence intervals, however, are useful tools to evaluate the observed effect estimate relative to its variability at an interim or the final look at the data.

Imagine a researcher observes mean differences between the two conditions of $\Delta M = 0.59$, 95% CI $[0.04, 1.13]$, $t(64) = 2.16$, $p = 0.034$ at look 1, $\Delta M = 0.39$, 95% CI $[0.00, 0.78]$, $t(130) = 1.98$, $p = 0.049$ at look 2, and $\Delta M = 0.47$, 95% CI $[0.15, 0.79]$, $t(196) = 2.92$, $p = 0.004$ at the last look. Based on a Pocock-like alpha spending function with three equally spaced looks, the alpha level at each look for a two-sided t-test is 0.0226, 0.0217, and 0.0217. We can thus reject H_0 after look 3. Where the unadjusted mean difference was 0.47, the adjusted mean difference is 0.40, 95% CI [0.02, 0.74]. The adjusted p value at the final look is 0.0393.

Discussion

This tutorial focused on sequential group designs aimed at testing a difference in means. The `rpact` package also provides options to design studies with survival and binary endpoints. Researchers might wonder how they can plan and analyze a group sequential design for other statistical tests. First, the alpha corrections that can be computed in `rpact` apply for any design where the data is normally distributed, and where each group of observations is independent of the previous group. Since the assumption of normality underlies a wide range of tests (e.g., hierarchical linear models) this means that the corrected alpha levels for each look will be widely applicable.

For the types of designs that are common in clinical trials (one- and two-sample *t*-tests, one- and two-sample tests for binary data, and survival data) `rpact` allows users to perform sample size and power calculations. Researchers in psychology will want to perform an a-priori power analysis for a wider range of tests than is currently implemented in `rpact`. To compute the required total sample size for a sequential design, however, the inflation factor can be used to compute the increased number of observations that is required relative to a fixed design. Researchers can perform an a-priori power analysis for a fixed design with any tool they would normally use, and multiply the total number of observations with the inflation factor to determine the required sample size for a sequential design. Analogously, the expected (relative) reduction in sample size (e.g., under the assumed alternative) can be multiplied with the sample size of the fixed design. Power can also be computed through simulations for any type of design, based on the alpha levels at each look. Creating software to perform power analyses and corrected effect size estimates for a wider range of tests would be useful, and could further facilitate the uptake of sequential designs.

Group sequential designs are one approach to sequentially analyzing data, but many alternative procedures exist. All sequential designs can be more efficient than fixed designs. Alternative approaches include the Sequential Ratio Probability Test (Wald, 1945) which is optimally efficient if researchers are able to analyze the data after every additional observation that is collected (Schnuerch & Erdfelder, 2020). Group sequential designs were developed to accommodate situations where this is not possible, and where data is collected in groups of observations. Building on the group sequential literature, adaptive designs allow researchers to perform sequential analyses where the sample size at later looks is adjusted based on analyses at earlier looks (Proschan et al., 2006; Wassmer & Brannath, 2016). Additional procedures exist that allow researchers to combine independent statistical tests (Westberg, 1985), or sequentially analyze independent sets of observations (Miller & Ulrich, 2021). Some approaches require researchers to collect a minimum number of observations, and continue the data collection until an upper or lower threshold of a test statistic is reached. Error rates are determined based on simulations, and the test statistic thresholds can be either a p value (e.g., data is collected until $p < 0.01$ or $p > 0.36$, see Frick (1998)), or a Bayes factor (Schönbrodt et al., 2017). A more recent development, *safe tests*, allows for optional continuation based on all data in the past, while controlling error rates (Grünwald, Heide, & Koolen, 2019; Schure & Grünwald, 2019). A comparison between these approaches is beyond the scope of this tutorial.

Sequential analyses provide researchers with more flexibility than fixed designs, but this freedom can be abused to inflate the Type 1 error rate when researchers are not honest about how often they performed interim analyses. A survey among editors for journals in psychological science revealed that they are positive about the use of sequential analysis when researchers preregister their statistical analysis plan, but editors are on average negative about the use of sequential analyses when the analysis plan is not preregistered (Lakens, 2017). If researchers choose to preregister their sequential design and statistical analysis plan they can include either the R code, or download a PDF report from the `rpact` shiny app that specifies the planned number of interim analyses and the stopping rule. The Journal Article Reporting Standards of the APA (Appelbaum et al., 2018) require researchers to provide an "explanation of any interim analyses and stopping rules employed". Although no guidelines for reporting sequential designs have been developed for psychologists, researchers should report the planned number of looks, the alpha level at each look, the type of design that is used to control the Type I and/or II error rate (e.g., the Pocock correction, or the Pocock alpha spending function), whether binding or non-binding futility bounds are used, and the planned sample size at each look. This is mandatory in order to guarantee that the Type I error is controlled at the specified level and a valid confirmatory statistical conclusion can be reached.

In clinical trials a data monitoring committee takes the responsibility of data analysis and the decision about whether the study should be continued out of the hands of the researcher who performs the study. The use of such a data monitoring committee is not common in psychology. For important studies, researchers could organize a data monitoring committee by enlisting the help of an independent research team that performs the interim analyses. Without an independent data monitoring committee, a researcher can maintain a publicly available log of any deviations from the planned analysis (e.g., looks at the data after a different number of participants than planned), the content of which should allow peers to transparently evaluate the severity of the reported test (Lakens, 2019; Mayo, 2008).

Whenever it is logistically possible to perform interim analyses in a study, the use of a group sequential design has the

potential to increase the efficiency of data collection. Resources in science are limited, and given that group sequential designs can save both time and money required to compensate participants, it is surprising that group sequential designs are not used more widely. We hope this tutorial increases the uptake of group sequential designs in psychological science, thereby increasing the efficiency of data collection. As an increasing number of journals are starting to expect a sample size justification for each study that is published, sequential designs offer a useful extension to current research practices, especially when there is large uncertainty about the effect size one expects (Lakens, 2021). Instead of an a-priori power analysis for an expected effect size, researchers can determine the smallest effect size they would be interested in detecting, determine the maximum sample size they will collect, and perform interim analyses that allow them to stop early when the null hypothesis can be rejected, or the smallest effect size of interest can be rejected.

References

Albers, C. (2019). The problem with unadjusted multiple and sequential statistical testing. *Nature Communications*, *10*(1), 1921. https://doi.org/10.1038/s41467-019-09941-0

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*(2), 235–244. https://doi.org/https://doi.org/10.2307/2343787

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.

Cook, T. D. (2002). P-value adjustment in sequential clinical trials. *Biometrics*, *58*(4), 1005–1011. https://doi.org/https://doi.org/10.1111/j.0006-341X.2002.01005.x

Dodge, H. F., & Romig, H. G. (1929). A Method of Sampling Inspection. *Bell System Technical Journal*, *8*(4), 613–631. https://doi.org/10.1002/j.1538-7305.1929.tb01240.x

Dupont, W. D. (1983). Sequential stopping rules and sequentially adjusted P values: Does one require the other? *Controlled Clinical Trials*, *4*(1), 3–10. https://doi.org/10.1016/S0197-2456(83)80003-8

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 690–697. https://doi.org/https://doi.org/10.3758/BF03209488

Grünwald, P., Heide, R. de, & Koolen, W. (2019). Safe Testing. *arXiv:1906.07801 [Cs, Math, Stat]*. Retrieved from http://arxiv.org/abs/1906.07801

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2017). *Will knowledge about more efficient study designs increase the willingness to pre-register?* MetaArXiv. https://doi.org/10.31222/osf.io/svzyc

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, (62(3)), 221–230. https://doi.org/https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2021). *Sample Size Justification*. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lan, K. K. G., & DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika*, *70*(3), 659–663. https://doi.org/10.2307/2336502

Mayo, D. G. (2008). How to Discount Double-Counting When It Counts: Some Clarifications. *The British Journal for the Philosophy of Science*, *59*(4), 857–879. https://doi.org/10.1093/bjps/axn034

Miller, J., & Ulrich, R. (2021). A simple, general, and efficient procedure for sequential hypothesis testing: The independent segments procedure. *Psychological Methods*.

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. https://doi.org/https://doi.org/10.2307/2530245

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York, NY: Springer.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, *25*(2), 206.

Schoenfeld, D. A., & Meade, M. O. (2005). Pro/con clinical debate: It is acceptable to stop large multicentre randomized controlled trials at interim analysis for futility. *Critical Care*, *9*(1), 34–36. https://doi.org/10.1186/cc3013

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/MET0000061

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32562-0

Schure, J. ter, & Grünwald, P. (2019). Accumulation Bias in meta-analysis: The need to consider time in error control. *F1000Research*, *8*, 962. https://doi.org/10.12688/f1000research.19375.1

Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, *7*(1), 8–17. https://doi.org/https://doi.org/10.1016/0197-2456(86)90003-6

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, *43*(1), 193–199. https://doi.org/https://doi.org/10.2307/2531959

Wassmer, G., & Pahlke, F. (2020). *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. Retrieved from https://www.rpact.org

Westberg, M. (1985). Combining Independent Statistical Tests. *Journal of the Royal Statistical Society. Series D (the Statistician)*, *34*(3), 287–296. https://doi.org/10.2307/2987655