Group sequential designs for in vivo studies

Blotwijk, Susanne; Hernot, Sophie; Barbé, Kurt

Published in:
Research in Veterinary Science

DOI:
10.1016/j.rvsc.2022.03.003

Publication date:
2022

License:
CC BY-NC-ND

Document Version:
Accepted author manuscript

Link to publication

Citation for published version (APA):
Blotwijk, S., Hernot, S., & Barbé, K. (2022). Group sequential designs for in vivo studies: Minimizing animal numbers and handling uncertainty in power analysis. *Research in Veterinary Science*, *145*, 248-254. https://doi.org/10.1016/j.rvsc.2022.03.003

Group sequential designs for in vivo studies: Minimizing animal numbers and handling uncertainty in power analysis

Susanne Blotwijk[a*], Sophie Hernot[b], and Kurt Barbé[a]

[a] Biostatistics and Medical Informatics Research Group (BISI), Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium. [b] Laboratory for In vivo Cellular and Molecular Imaging, (ICMI-BEFY/MIMA), Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium.

*Corresponding author, e-mail: susanne.blotwijk@vub.be

Interim analysis is the practice of performing a statistical analysis when the data have only been partially collected, for example, to save resources or to handle the uncertainty of the true effect size. Most statistical designs featuring interim analysis have been developed either in a general statistical setting or for application in clinical trials. As a result, most of them make assumptions and have conditions that in a preclinical setting are usually not met. In this paper, we present necessary changes to the most common forms of interim analysis enhanced for animal experiments, specifically for the t-test and the one-way ANOVA. Finally, we present software that allows freeware use to serve the research community to facilitate the design of experiments featuring interim analyses.

The app can be found at icds.be/gsdesigner. It is in the public domain and its code can be found on github.com/ICDS-vubUZ/gsd-designer. In this GitHub folder, one can also find a tutorial for the app.

The use of interim analyses is common in clinical trials, due to its potential benefits. An appropriate statistical design featuring an interim analysis can reduce the sample size for an experiment by 20% (Neumann et al., 2017; Wassmer and Brannath, 2016), which can bring significant practical, financial, and ethical benefits. Such a design can also be used to help balance concerns in power analysis caused by the uncertainty of the effect size. This is especially applicable in preclinical studies involving animals, where generally very little information is available in advance, making it hard to estimate an appropriate sample size.

Given the potential benefits, it should be no surprise that several papers (Fitts, 2011, 2010; Ludbrook, 2003; Maïofiss-Dullin et al., 2007; Neumann et al., 2017; Steward and Balice-Gordon, 2014; van Wilgenburg et al., 2003) have been written to investigate or encourage the use of interim analyses in preclinical studies. The papers by van Wilgenburg et al. (2003), and Steward and Balice-Gordon (2014) have a much wider scope and do not discuss any particular models which should be used. Others (Ludbrook, 2003; Maïofiss-Dullin et al., 2007; Neumann et al., 2017), despite being explicitly written for animal experiments, describe methods which are unsuitable for this context, or at the very least are severely suboptimal. This is either because they use bounds that are only suitable at large sample sizes or because they lose a considerable amount of statistical power in ways that could easily have been avoided by enhancing the design mathematically. To the best of our knowledge, only the bounds proposed by Fitts (2011, 2010) are truly suitable for the preclinical context for which they were intended. However, they are inflexible both for handling data loss and for error spending, thereby usually requiring a higher maximum total sample size.

In this paper, we discuss the use of interim analyses in the context of the null hypothesis significance testing (NHST) framework. While the use of p-values to draw conclusions is flawed and often misinterpreted (Tong, 2019; Ziliak and McCloskey, 2008), it remains the dominant form of statistical analysis in scientific literature. In order to counter some of the problems created by the NHST, it is becoming more common to encourage or even require reporting of the magnitude effect and its uncertainty, rather than overly focusing on statistical significance (Betensky, 2019; Sullivan and Feinn, 2012). As such, the impact of using interim analyses on the estimate of the effect size and its confidence interval are also discussed in this paper.

Problem statement and objectives

Problem statement

Consider a study with a few experimental treatments and a control group. In a classical experimental design, we would wait until all measurements are made, all the data have been collected, and only then do we perform statistical analysis. However, it is also possible to perform an analysis when only part of the data was collected, obtain a significant result, and finish the study. If the result is not significant, but still sufficiently promising, we can continue collecting more data and re-evaluate later. This practice is referred to as performing an interim analysis and when performed correctly, this can have significant benefits. Obtaining a significant result early will save time, effort, and resources required to collect the remaining measurements, as well as minimize the number of animals to be used and prevent associated animal suffering. Performing interim analyses can be done solely with those aspects in mind, but it can also solve more problematic issues rendered by classical designs

The gold standard for sample size calculation is through power analysis (Silverman et al., 2014; van Wilgenburg et al., 2003), where the resulting sample size will depend on the assumed effect size. However, the true effect size is uncertain in advance; otherwise, there would be little value in performing the experiment. When we expect the effect size to be larger than the minimal scientifically relevant difference, it can be difficult to determine an appropriate sample size. We do not want to end up with non-significant results merely because we were too optimistic about our effect size, nor do we want to overspend and cause unnecessary suffering just because we were too cautious. Adding interim analyses balances those considerations.

Researchers have also reported issues in power analysis sample size determination due to practical limitations in terms of personnel and equipment (Fitzpatrick et al., 2018). The required sample size may be larger than what can be processed at once, e.g. due to labor-intensive animal procedures and data collection processes, or limitations in housing capacity. Such constraints create an extra burden on researchers and while a sequential design cannot completely remove this problem, it can certainly make it generally less burdensome.

Another dilemma resolved through interim analysis occurs in case of larger than expected data loss. In this case, the researcher can either collect a second batch of data, to compensate for the data loss, or perform the data analysis with the limited data available, knowing that the design is underpowered. The latter option contains a significant risk that even if a meaningful effect is present, it will not be significant. On the other hand, the former option might significantly prolong the duration of the experiment. In such circumstances, performing an interim analysis can prevent this in case the results are significant, but without the need to discard the collected data if the interim result was not significant. Either way, the design will be sufficiently powered. Some extra precautions need to be taken when implementing an interim analysis for these reasons. These are discussed in appendix B.

Regardless of the reason for performing an interim analysis, there are some consequences. When we set a significance level, it is meant to limit the probability of a false positive, the type I error. If we perform multiple analyses, we have multiple opportunities to obtain a significant result, so our total probability of a false positive increases. Similarly, if we decide to stop early because the data seems insufficiently promising, this decreases the total probability of obtaining a significant result. However, it also increases the probability of a false negative, the type II error. Both types of errors can be controlled by adapting each analysis to that p-value at which our result is significant and from which p-value our treatment is insufficiently promising to continue our experiment.

If we want to increase the probability of getting a significant result early, then we can increase the allowed probability of a false positive at an earlier analysis. To control the type I error, the total probability of a false positive under the null hypothesis needs to stay the same. In order to compensate for the increase at the earlier analysis, we need to decrease the probability of getting a significant result at a later analysis. However, at the later analysis, we have a larger total sample size, so more power. If the loss of power is too severe, we can compensate by slightly increasing the sample size at the last analysis. These levels of freedom are studied and adapted to enhance and optimize animal studies in this paper.

Experimental set-up

The statistical designs we discuss in this paper are Group Sequential Designs (GSD). In this type of design, interim analyses provide the opportunity to determine if the results are (in)sufficiently significant and to end the experiment early.

In this article, we discuss GSDs for the t-test and the one-way ANOVA only. Just as in a fixed sample size experiment, i.e. a design without interim analysis, we assume the data to be identically and independently distributed. This means the experimental design is not changed once the experiment has started, the same procedures, dose, mouse type, etc. are used in the first set of collected data points as in all proceeding measurements.

Similarly, the statistical design and the rules for the GSD should not be changed once the experiment has started. The most important reason is that once one has knowledge of the data, any change to the model almost certainly introduces a bias rendering conclusions unreliable. The second reason is practical, namely that the choice of sequential design will influence the sample size calculation. Therefore, determining the appropriate statistical design should be done simultaneously with the power analysis.

Nowadays, GSDs are considered to be a special case of adaptive designs. Other types of adaptive designs may or may not have this same ability to stop early, but mainly they allow to change key features of the design at the time of the interim analysis, e.g. doses or number of experimental branches. These extra adaptive features are often unsuitable for hypothesis testing at small sample sizes, or they reduce the power of the test, requiring a larger sample size to compensate

(Jennison and Turnbull, 2005; Kelly et al., 2005; Tsiatis and Mehta, 2003; Wassmer and Brannath, 2016).

Such adaptive designs might certainly be of interest in explorative preclinical experiments or to merge experiments that are currently performed separately. In this paper, however, we focus on improving on, and dealing with issues in, hypothesis testing experiments as they are currently performed in preclinical settings. As such, the GSDs are the most powerful and most suitable designs for this confirmatory context. Additionally, GSDs are more similar to traditional statistical designs and hence easier to learn and use for most researchers.

Existing methodology

The main difference between various GSDs is generally the choice of critical values, i.e. the values that the test statistics need to exceed or not in order to be considered significant or to be insufficiently promising to continue the experiment. One of the older and better-known GSDs are the Pocock bounds (Pocock, 1977). These keep the critical values the same over all analyses, which has the advantage that they are easy to use. A significant downside is that this method is not very statistically powerful. They can also lead to the awkward situation where an effect is not found to be statistically significant despite the test statistic being much larger than it would have to be for a fixed sample design. The O’Brien-Fleming bounds (O’Brien and Fleming, 1979) reduce these problems by having stricter bounds at early analyses and less strict as more data is collected. The alpha spending approach developed by Lan and Demets (1983) allows the user to specify exactly how strict or flexible they wish to be early on.

Both the Pocock and the O’Brien-Fleming bounds are fixed bounds designs, which require the number of interim analyses and the amount of data collected at each analysis to be determined in advance. The alpha spending approach is more flexible and can easily be adapted in case the data collection does not go as planned, e.g. in case of data loss. In theory, the alpha spending approach does not even require the number of analyses to be fixed in advance, although doing so is not advised in practice.

Originally, all these methods were only developed to stop early for significance. Since then, natural extensions of each of these methods have been published to stop early for futility, i.e. for insufficiently promising data. While the above methods are in theory not restricted to any specific test, applying the theory is easier in some cases than in others. The bounds or software packages one will find in practice are often calculated for normally distributed test statistics. At the time of writing, this is the case in the original papers themselves in the SEQDESIGN procedure for SAS and the gsDesign R-package. The reason for this is that many test statistics asymptotically approach a normal distribution if the sample size is sufficiently large. This asymptotic approximation works well if the sample size is large, as is common in clinical trials, but becomes inaccurate at the smaller sample sizes generally used in preclinical studies.

For preclinical studies, Fitts (2011, 2010) obtained Pocock-style bounds through simulation for several different tests commonly used in preclinical research. In the context of clinical trials with small sample sizes, Shao and Feng (2007) did the same for Pocock-style bounds of the t-test. The reason Fitts' and Shao and Feng's bounds differ, is that the former provides significance bounds for the p-values, whereas the latter provides them for the test statistics. For normally distributed test statistics both approaches have the same result, therefore in the original Pocock paper this distinction was not relevant and as such not discussed.

As for the alpha spending approach, techniques for small sample sizes have only been discussed in the clinical context and only for the t-test. Rom and McTague (2020) have described a numerical technique to calculate the exact significance bounds for designs with only one interim analysis and no futility bounds. For designs with beta spending and/or with more analyses, Nikolakopoulos et al. (2018) discuss an approximate analytical correction to improve the significance bounds of the normal asymptotic approximation.

In this paper, we extended the formulas for the exact approach of Rom and McTague to calculate exact futility bounds as well. We improved the analytical approximation of Nikolakopoulos et al. Consequently, we also provide several recommendations on how to simulate and evaluate the critical bounds, the nominal error level, and the power quickly and with the desired level of accuracy.

Objectives

The main objective of this paper is to propose efficient group sequential designs for the preclinical setting. This includes providing methods to approximate the corresponding critical values such that the correct significance level and power level are achieved at the small sample sizes common in these types of experiments. Additional properties in the designs we discuss, are the flexibility to handle data loss efficiently and a minimization of the expected costs, sample size, and/or duration of the experiment.

A secondary objective is to facilitate the design of such experiments by providing open-source software and by providing technical details useful for design purposes in a preclinical context.

Toy example

To illustrate the concepts in this paper, we apply them to a toy example. This toy example is an experiment on mice where the researchers wish to investigate the difference between a treatment group and a control group. This same control group has been used for other experiments in the past, so the mean and standard deviation we expect there are estimated with values of 1 and 0.1 respectively.

The treatment group, on the other hand, is completely new. From similar experiments, the researchers think it is likely that the treatment group can outperform the control group with a mean that is 20% higher. However, if we are sufficiently confident that the improvement is less than 14%, this is a strong enough claim to publish and justify not pursuing follow-up experiments. Here, sufficiently confident is 1 – β = 80%, the desired power of the design. The significance level in this experiment is the usual α = 5%. If we are 95% confident that the improvement is larger than 0%, this is a strong enough claim to publish and justify follow-up experiments.

The researchers will compare these two groups using a one-sided t-test, for which the effect size is called Cohen's d (Cohen, 2013). By combining all the above information, one obtains a likely effect size of 2 and a minimally relevant effect size of 1.4. Under a normal fixed sample design, the minimum sample size to obtain sufficient power for the minimally relevant effect size is 8 mice per group or 16 mice in total.

The process of collecting the data from these mice is very labor-intensive and as a result, only 6 mice can be processed per day. This means that the total data collection process will take 3 days. In this toy example, the researchers choose to perform a statistical test at the end of each day.

Revisiting alpha and beta spending

Alpha spending

The alpha spending approach is a type of group sequential design developed by Lan and Demets (1983). Unlike earlier designs, such as the Pocock (1977) and O'Brian-Fleming (1979) bounds, this approach allows considerably more flexibility in choosing when and how often to perform interim analyses. This is done by defining how large the type I error is allowed to be at any point in time during the experiment. A larger type I error allowed at an earlier analysis increases the probability of stopping early and thereby saving more time and resources. Since the increase in power at the earlier analysis is smaller than the loss of power at later analyses, the price paid is that the total power of the experimental design decreases.

Based on the allowed type I error probabilities, we can calculate critical values determining the threshold for significance. This can be done either for the test statistics, in which case they are called significance bounds, or for their corresponding p-values.

These test statistics or p-values are calculated the same way as without interim analysis. Most, if not all, commonly available statistical software return these values for a normal t-test or one-way ANOVA. We conclude the result is significant if the test statistic is larger than the significance bound or if the obtained p-value is smaller than the critical p-value. Mathematically speaking, these two approaches are completely equivalent. From a researcher's perspective, however, they might not be.

One reason is that in traditional designs, the p-value is the probability that the null-hypothesis is rejected, in the case that the null-hypothesis is true.. After the first interim analysis, that is no longer the case. Since the data from our first interim analysis is also used in the second analysis, there is a correlation between their test statistics and hence the traditional probability distributions no longer apply. An example of the difference between the critical values for the p-values and the actual probability of a type I error is illustrated for a specific design for the toy example in table 1. Because of this difference, our intuitive understanding of what these p-values mean, tends to be wrong. Hence it is generally preferable to work with significance bounds instead.

The distribution of the type I error over the different analyses can be quantified with an alpha spending function $\alpha(t)$, which is defined as the total allowed probability that we have made a type I error before or at time t. When we have collected no data yet, this probability should be zero. At the other extreme, when we have collected all our data, this probability should be equal to the desired significance level α. Other than that, the only restriction on our spending function is that it should be non-decreasing, as we cannot retroactively reduce the probability of what we did earlier on.

In the above paragraph, t has been stated to represent time, but it does not have to. It is usually more meaningful to let the alpha spending function depend on the amount of data collected, where the time of the data collection does not matter. In this case, t would represent the sample size. Both of these interpretations are used in clinical studies and in both cases it is common to rescale t such that is not going from begin time to end time or from zero to maximum total sample size, but rather from 0 to 1. This way t can be interpreted as the information fraction, the ratio of the information gathered at interim relative to the total information gathered in case the experiment does not terminate at any of the interim analyses. It is this information fraction that allows us to handle data loss flexibly and efficiently. This is discussed more in Appendix B. In general, a meaningful choice for the information fraction is the ratio of the sample size at each analysis and the maximum total sample size. This choice was used in the example in table 1 with t equal to 6/16, 12/16, and 16/16 at the respective analyses.

Once it has been determined which information fraction to enter into the alpha spending function, we should mention the choice of the alpha spending function itself. There are infinite possibilities for choosing an alpha spending function, none of which are uniformly optimal. The best choice will depend on several factors, but this discussion is out of scope for this paper. Functions that are steeper early on and flatter towards the end have a higher probability of stopping early but are less powerful and require a higher maximum sample size to compensate. Conversely, functions that stay low in the beginning and only start rising near the end have higher power, but a lower probability of stopping early. The effect on the toy example of several different spending functions is illustrated in table 2.

We can define an expected sample size by weighing the used sample size at each analysis by the probability to stop at that analysis. In an optimistic scenario where the effect size is larger than the minimal relevant effect size, we are more likely to obtain a significant result early on, and therefore have a lower expected sample size N. Both the power and the odds of stopping early depend on the underlying effect size as well as the design. This is illustrated on the toy example in table 2. Due to the discrete nature of small sample sizes, it is hard to predict the exact effects of each choice. It is therefore probably wise to look at several options during the planning phase of the experiment.

When reporting the results of an experiment, it is good practice to report the magnitude of the observed effect, as statistical significance (or non-significance) by itself is not particularly meaningful. For the first analysis, one can simply use the regular effect size estimate and confidence interval as one would without GSD. However, at later analyses, the classical formula leads to an overestimation of the effect size and its confidence interval.

Assume in the toy example a design with O'Brien-Fleming spending function rendering a significant result at the second analysis, with a T statistic of 2.311. The naive, uncorrected estimate of Cohen's d would be 1.33 with 90%-confidence interval [0.283, 2.62]. However, applying the correction, the Cohen's d drops to 1.29 and [0.231, 2.37] respectively. This correction can be calculated in the app. For a more in-depth discussion of correction methods, we refer to Appendix A.

Beta spending

The concept of beta spending is entirely analogous to that of alpha spending, but rather than stopping early because we have reached significance, we can now stop early because the data is insufficiently significant. Instead of controlling the false positive rate under the null hypothesis, with beta spending, we are controlling the false-negative rate under the alternative hypothesis. This requires defining the alternative hypothesis, which in this case is the minimal scientifically relevant effect size or most pessimistic scenario for which we require sufficient power.

Based on the allowed type II error probabilities, we can once again calculate critical values either for the test statistics, now referred to as futility bounds, or for their corresponding p-values under the traditional probability. Due to the same reasoning as in the previous section, we prefer to work with the futility bounds rather than the p-values.

While it is possible to perform beta spending by itself, it is most commonly applied in combination with alpha spending. In this case, the result is still significant if the test statistic exceeds the significance bound, but insufficiently promising if the test statistic is lower than the futility bound. One only continues collecting data if the test statistic lies somewhere in between the significance and futility bound.

Since the test needs to achieve the required significance level, the last futility bound is determined by the allowed type I error, rather than the type II error. In case we apply alpha spending as well, this means we set the futility bound to be equal to the significance bound of the final analysis.

As with the type I error in the previous section, it is possible to quantify the type II error spending through a beta spending function, $\beta(t)$. The significance and futility bounds of the toy example for several error spending functions can be found in table 3. Note that the significance bounds in the later analyses are lower for a design with beta spending than that of the corresponding design without beta spending in table 2. Similarly, the futility bounds are higher in a design with alpha spending than in its equivalent without alpha spending. Since alpha and beta spending partially negate each other's downsides, they are often applied in a balanced way using identical spending functions.

Unlike in the situation where we only use alpha spending or beta spending, it is no longer the case that the effect size is exclusively over- resp. underestimated. Nevertheless, even when applying both alpha and beta spending, a correction of the effect size estimate and its confidence interval is still needed.

Application

The first part of any study is planning. For the largest part, this remains the same as one would do without interim analysis, save for two additional steps and one significantly affected step. The first new step is determining rules for when to perform an interim analysis. The second new step is to determine the allowed type I and II errors at these analyses, i.e. choosing the error spending functions. The step that is affected by adding interim analysis, is the power analysis for the sample size calculation.

Number and timing of the interim analyses

In the toy example, the implementation of the interim analyses is straightforward as the data is gathered in batches and therefore it is natural to update the analysis periodically. The only real choice one needs to make is if one chooses to perform an interim analysis at the end of every single batch or if some will be skipped. In other types of experiments, the researchers may not have the same restrictions and can choose the size of each batch, hence having complete freedom over the number and timing of the analyses. In yet other experiments, adding interim analyses might bring its own costs. Since group sequential designs work on the principle that the next batch is only started after the previous one has been processed, this might significantly prolong the duration of certain experiments in such a way that the added costs outweigh the benefits. The practical restrictions and possibilities will differ per experiment and need to be looked into on a case-by-case basis, but some general recommendations can be made on a statistical basis.

While there is no theoretical limit to the number of interim analyses, it was mentioned earlier that each interim analysis 'uses' some of our allowed probability of making type I and/or type II errors, we spend our alpha and beta. By implementing too many analyses, the probability of drawing a conclusion becomes so small that it undercuts the benefits of the group sequential design. Generally, having a total of 2 or 3 analyses works out well. Having more than 5 analyses usually becomes inefficient. The exact ideal amount and timing depend, among other things, on the difference between the optimistic and pessimist effect sizes. A larger optimistic effect size will benefit from more and earlier testing.

Power analysis and sample size calculation

In regular, fixed sample designs the achieved power can be substantially higher than the required power since we can only have whole numbers as sample size. This is the case in our toy example, where a sample size of 8 mice per group leads to a power of 0.845 or 4.5% above our required power, but having only 7 mice per group will leave the model underpowered.

Adding interim analyses with no other design changes will generally reduce the statistical power of the design. However, unlike in the situation where we have a large sample size, this does not need to imply that the power drops below the required level. This can be seen for the toy example in tables 2 and 3 where one of the choices of error spending functions still has sufficient power, even though it has the same total sample size as the fixed sample design.

The natural way of fixing the other designs is by increasing the maximum total sample size until the desired power has been reached. Other ways to make a design more powerful are to decrease the number of analyses, change their timing or choose a different error spending function.

In table 4 we have adapted the sample sizes of the examples from tables 2 and 3 such that the minimum required power for the toy example, 0.8, has been achieved. In this case, adding one or two mice per group sufficed, giving a maximum total sample of 18 or 20 mice per design. This is the sample size of the worst-case scenario where we cannot draw any conclusions in the interim analyses and continue to the final analysis. In contrast to the fixed sample size designs, we do not know what the final sample size will be until after the experiment. However, we can calculate the expected value of the sample size, i.e. the average obtained sample size if we were to repeat the experiment often enough.

This expected sample size will also depend on the effect size since the probability of obtaining a significant result is larger if the effect size is large. The expected sample sizes for the toy example are shown in case there is no effect (d = 0), for the pessimistic effect size (d = 1.4), and for the optimistic effect size (d = 2).

While it might seem tempting to choose alternative solutions to increase the power that do not require raising the maximum total sample size, it is worth pointing out that a lower maximum sample size does not necessarily lead to a lower expected sample size. In the designs covered for our toy example, the designs to achieve the best average sample sizes are the designs featuring the Pocock-type and compromise error spending functions. This is despite having a larger maximum sample size than other competing designs.

From the above discussion, it should be clear the potential gains of GSDs depend on the properties of the design, but also the true effect size. Even so, in our toy example, the expected sample size remains below 13 for all three designs featuring alpha and beta spending, regardless of the effect size. This shows that substantial gains can be made, even without researchers actively putting effort into optimizing the GSD.

Calculating the critical values

Either during the planning or at the interim analysis itself, the critical value to determine significance needs to be calculated. Unfortunately, the critical values can only be approximated. This can be done in a few different ways.

The most common analytical approach in clinical trials is through asymptotic approximation. The more data get collected, the more the t-distribution resembles a normal distribution, so the critical values are based on Z-tests rather than t-tests. This is fine for the large sample sizes common in clinical trials, but problematic at the much smaller sample sizes common in preclinical contexts. Most existing software uses this approach without mentioning this restriction. So if researchers choose to work with software other than our free web application, it is important they verify the applicability of their software package.

To obtain boundaries suitable for the t-test in the preclinical context there are currently three options: use simulation as we do in our application, use iterative numerical integration as proposed by Rom and McTague (2020), or improve the analytical approximation through a formula as was done by Nikolakopoulos et al (2018). For the one-way ANOVA, to the best of our knowledge only simulation is available. For a technical discussion of these techniques, their advantages, disadvantages, and our extensions and improvements, we refer to Appendix A.

Rules of thumb for preclinical studies

Planning a group sequential design involves more choices than planning a traditional fixed sample design. Here are some guidelines that should facilitate those choices and help avoid the most common pitfalls.(Kelly et al., 2005)

- **Check if the chosen software is suitable for small sample sizes.** If not, apply a t-transformation as described in appendix A under the section "Analytical approximation".
- **Keep the number of analyses limited.** A total of two or three analyses usually works well, more than five is generally inefficient.
- **Compare several spending functions before making a decision.** The best choice differs per experimental set-up.
- **Determine rules** on how to handle data loss or other required flexibility of the design **before the start of the experiment**.
- **Do not make ad hoc changes to the design after the experiment has started.**
- **Performing both alpha and beta usually has a better trade-off** between expected sample size and power. If it is deemed highly unlikely that there is no relevant effect, then it is better to only apply alpha spending.

Conclusion

Implementing group sequential designs can reduce the average cost, duration, and sample size in preclinical experiments. This type of design can aid in navigating the uncertainty of the true effect size as well as providing a flexible and efficient way of dealing with data loss.

Due to the small sample sizes common in this setting, specialized techniques need to be applied. In this paper, we discussed and improved such techniques for the t-test and the one-way ANOVA. Furthermore, a free simulation tool is presented specifically designed for preclinical applications. This tool circumvents the typical limitations of other methods wherein large sample approximations are used.

References

Betensky, R.A., 2019. The p-Value Requires Context, Not a Threshold. Am. Stat. 73, 115–117.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Routledge.

Fitts, D.A., 2010. Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. Behav. Res. Methods 42, 3–22.

Fitts, D.A., 2011. Minimizing Animal Numbers: The Variable-Criteria Sequential Stopping Rule. Comp. Med. 61, 206–218.

Fitzpatrick, B.G., Koustova, E., Wang, Y., 2018. Getting personal with the "reproducibility crisis": interviews in the animal research community. Lab Anim. 47, 175–177.

Jennison, C., Turnbull, B.W., 2005. Meta-Analyses and Adaptive Group Sequential Designs in the Clinical Development Process. J. Biopharm. Stat. 15, 537–558.

Kelly, P.J., Sooriyarachchi, M.R., Stallard, N., Todd, S., 2005. A Practical Comparison of Group-Sequential and Adaptive Designs. J. Biopharm. Stat. 15, 719–738.

Lan, K.K.G., Demets, D.L., 1983. Discrete sequential boundaries for clinical trials. Biometrika 70, 659–663.

Ludbrook, J., 2003. Interim analyses of data as they accumulate in laboratory experimentation. BMC Med. Res. Methodol. 3, 15.

Maïofiss-Dullin, L., Boussac-Marlière, N., Geffray, B., Haimez, C., Harriong, S., Hitier, S., Onado, V., 2007. On the Efficiency of Interim Analyses Applied to Nonclinical Studies. Drug Inf. J. 41, 517–526.

Neumann, K., Grittner, U., Piper, S.K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J.P.A., Kimmelman, J., Dirnagl, U., 2017. Increasing efficiency of preclinical research by group sequential designs. PLOS Biol. 15, e2001307.

Nikolakopoulos, S., Roes, K.C., van der Tweel, I., 2018. Sequential designs with small samples: Evaluation and recommendations for normal responses. Stat. Methods Med. Res. 27, 1115–1127.

O'Brien, P.C., Fleming, T.R., 1979. A Multiple Testing Procedure for Clinical Trials. Biometrics 35, 549–556.

Pocock, S.J., 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika 64, 191–199.

Rom, D.M., McTague, J.A., 2020. Exact critical values for group sequential designs with small sample sizes. J. Biopharm. Stat. 30, 752–764.

Shao, J., Feng, H., 2007. Group sequential t-test for clinical trials with small sample sizes across stages. Contemp. Clin. Trials 28, 563–571.

Silverman, J., Suckow, M.A., Murthy, S., 2014. The IACUC Handbook, Third Edition. CRC Press.

Steward, O., Balice-Gordon, R., 2014. Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience. Neuron 84, 572–581.

Sullivan, G.M., Feinn, R., 2012. Using Effect Size—or Why the P Value Is Not Enough. J. Grad. Med. Educ. 4, 279–282.

Tong, C., 2019. Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. Am. Stat. 73, 246–261.

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

van Wilgenburg, H., van Schaick Zillesen, P.G., Krulichova, I., 2003. Sample Power and ExpDesign: tools for improving design of animal experiments. Lab Anim. 32, 39–43.

Wassmer, G., Brannath, W., 2016. Group Sequential and Confirmatory Adaptive Designs in Clinical Trials, 1st ed, Springer Series in Pharmaceutical Statistics. Springer International Publishing, Cham.

Ziliak, S.T., McCloskey, D.N., 2008. The cult of statistical significance: how the standard error costs us jobs, justice, and lives, Economics, cognition, and society. University of Michigan Press, Ann Arbor.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.