Charité – Universitätsmedizin Berlin
Institut für Biometrie und Klinische Epidemiologie

User Guide for Biometric Planning of Animal Trials

Sophie K. Piper, Dario Zocholl, Ulf Toelch, Robert Roehle, Frank Konietschke

This is a living document. Please check whether you have the latest release: https://doi.org/10.5281/zenodo.7038608

Date: November 25, 2022

Contents

1 Introduction

1.1 About this booklet

This booklet is a user guide for experimenters who intend to carry out an animal trial and seek to obtain ethics approval from regulatory authorities. It is also meant to aid scientists and biometricians who have to judge statistical soundness and efficiency of a trial in order to give a proper ethical evaluation.

Parts of this booklet have been published in Piper et al. "Statistical review of animal trials - A guideline" (Piper et al. (2022)).

In order to accompany ethics approval for animal trials, we developed a biometric form to be filled and handed in with the proposal at the local authority of animal welfare. The form is shown on the next page in English language and it has already been in use in German language by the local authority of animal welfare in Berlin, Germany, since beginning of 2020. The German version is available online at
https://www.berlin.de/lageso/gesundheit/veterinaerwesen/tierschutz/versuchsvorhaben/.

The booklet is composed of 8 chapters subsequently addressing each section in the biometric planning form. In the overview of this form sheet on the next page, the reader may click directly on each referenced number to be directly forwarded to the corresponding chapter within this booklet. An example template was published in Piper et al. (2022).

As it is crucial to distinguish between exploratory and confirmatory research already in the planning phase, we will briefly explain the two concepts in the following subsection 1.4. For better understanding and illustration we provide example text blocks for two typical scenarios of animal trials: i) an exploratory setting and ii) a confirmatory setting which are introduced at the end of the introduction in subsections 1.4.1 and 1.4.2. In the appendix (see chapter 9) we provide a brief glossary of statistical terms.

1.2 Recommended resources

An overview of very useful links is hosted under
https://charite3r.charite.de/3r_service/charite_3r_toolbox/.

1.3 Biometric Planning Form

Experiment number:
Animal species:
Number of animals per investigated group in this experiment:
Total number of animals (including dropouts) in this experiment:

1. Goal of the (sub)-trial (including research question or hypothesis and indicating whether this is an exploratory or confirmatory experiment or a technical pilot study): →see chapter 2
2. Primary endpoint of the experiment, with unit of measurement, measurement method and time of measurement:→see chapter 3
3. Description of the study design, e.g. which groups are compared, which interventions are performed, when is the outcome measured? →see chapter 4
   (a) Design (if possible with flowchart):
   (b) Blinding (e.g. double-blind or evaluation blinded; if no blinding is done, please explain why):
   (c) Randomization (which type; if no randomization is used, please explain why):
4. Sample size calculation: →see chapter 5
   (a) For a confirmatory trial: →see chapter 5.2
      (i) Statistical test used for sample size calculation:
      (ii) Significance level (α) and power (1-β), one-sided or two-sided test:
      (iii) Biologically relevant effect size (please do not only provide the effect size, but also which data(value) would lead to this effect size, e.g. desired/expected mean value and variance/standard deviation per group, including a reference)
      (iv) Calculated sample size per group (including explanation with respect to i-iii):
      (v) The required number of reserved animals/dropouts due to premature death, faulty interventions, etc. (Specify a dropout rate and the required absolute number of animals):

(vi) Software used for sample size calculation (including version number):

(b) Exploratory trial/pilot study or orientation test/technical preliminary test: →see chapter 5.1

(i) Sample size calculation explanation (e.g. feasibility, precision of estimation that can be achieved with given sample size)

(ii) The required number of reserved animals/dropouts due to premature death, faulty interventions, etc. (Specify a dropout rate and the required absolute number of animals):

(iii) Software used for sample size calculation (including version number):

5. Statistical analyses: (e.g.: type of statistical modelling, adjustment for potential bias, adjustment for multiple testing, secondary analysis, handling of missing values?) →see chapter 6

6. Is there a logical or sequential order of the experiments planned ? (e.g. prerequisites that have to be fulfilled and consequences on any of the following experiments that can arise) →see chapter 7

7. Summary table for the sample size planning: Table 1→see chapter 8

Table 1: Tabulated overview of groups, effect sizes and sample size calculations

Group	Primary Endpoint	Expected Effect (e.g., means with standard deviations)	Reference for expected effect	Effect Size	Drop-Out Rate	Sample Size (including Dropouts)
A: control	weight [g] after 3 weeks	A: mean 100g (SD: 20g)	Name et al. (2020)	Cohen's d=1	25%	$17/0.75 \approx 23$
B: treatment		B: mean 120g (SD: 20g)			25%	$17/0.75 \approx 23$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total:						46

1.4 Exploratory versus confirmatory research

There are two modes of research, exploration and confirmation, which have to be distinguished conceptually and practically (Kimmelman et al. (2014); Dirnagl (2019)). They differ substantially in their methodology and in the interpretation of their findings, and therefore also require different approaches to the statistical planning of the experiment. "While exploration may start without any hypothesis ('unbiased'), a proper hypothesis is the obligatory starting point of any confirmation." (Dirnagl (2019)). The vast majority of animal trials is in the area of exploratory experiments, but the number of confirmatory trials has been increasing, recently.
An exploratory experiment aims at investigating physiological or pathophysiological mechanisms or potential drug development (Dirnagl (2019)). There does not need to be a pre-specified effect or hypothesis before observing the experiment. Instead, the goal in exploratory research is to generate new hypotheses and estimate effect sizes, which have to be tested and confirmed later. Thus, if no prior data and reliable knowledge about the desired effect sizes exist, experiments usually fall into the category of exploratory research. Typical study examples from exploratory research aim estimation (possibly descriptive) estimation of effects ("pilot study"), exploration of potential biomarkers, assessing feasibility or dose-finding, respectively. There might already be a pre-defined hypothesis though not enough prior knowledge for an adequate sample size planning. Previous evidence might exist but it is potentially associated with high uncertainty due to low number of experimental units.
A confirmatory experiment is based on previously found differences, in order to confirm an expected effect, and tests a specific, pre-defined hypothesis. The statistical properties of the primary hypothesis test need to be clearly described, so that the experiment is adequately powered. In this way, the planning is analogue to a clinical trial. Confirmatory experiments can be replications of previous experiments if there is a clear rationale what additional evidence will be generated. Confirmatory experiments are further strengthened by describing the converging and discriminant evidence that is generated. That is, what additional measurements will strengthen the hypothesised causal relationship? What measurement instruments yield similar findings (mRNA upregulated $\rightarrow$ higher protein concentration) and what experiments are needed to rule out viable alternatives.

1.4.1 Specific exploratory example

A (fictive) example for an exploratory experiment in fundamental research is to investigate the influence of a growth factor GX on fracture healing. No specific

preliminary data are available in the specific type of mice used, but there are hints from other models or exvivo experiments that increased GX levels are associated with reduced bone formation and GX regulation might be used to influence fracture healing. In order to capture dynamics of the healing process, bone fraction of the callus volume shall be measured by µCT after 3, 14 and 21 days and compared to GX depleted control animals. This will be explored in two different strains of mice that model fracture healing in an immunologically young (strain A) or an experienced adaptive immune system (strain B), respectively. See figure 1 for an illustration of the design. In addition histological and immunohistological analyses, FACS analysis and evaluations of immunologically relevant organs such as spleen, lymph nodes and bone marrow are planned so that animals have to be killed at each time point.

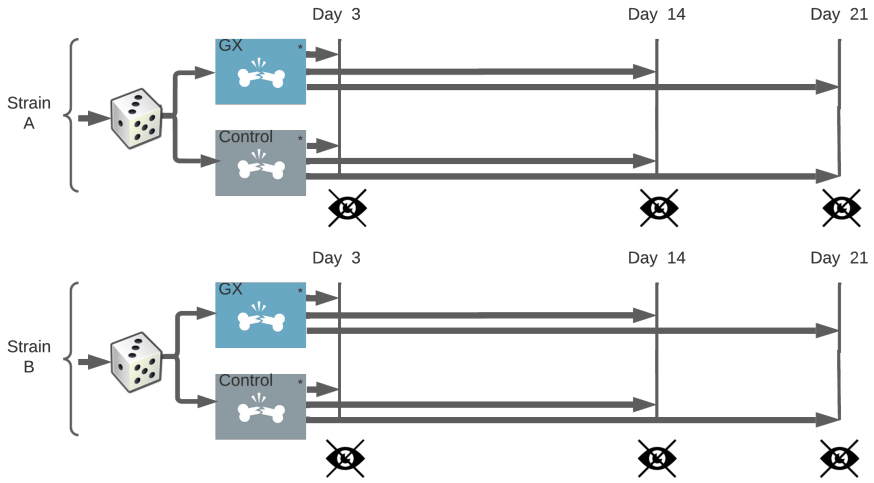

Figure 1: Design flowchart for our exploratory example: A 2x2x3 Design. Two strains of mice in two treatment groups (GX treatment/ GX depleted controls) are measured in three subgroups on day 3, 14 and 21 after osteotomy, respectively. Crossed eyes symbolize the investigator is blinded to treatment allocation. Build with www.lucidchart.com.

1.4.2 Specific confirmatory example

A (fictive) example for a confirmatory study in preclinical research has the goal of confirming a previously found effect with higher reliability.

To study hypertrophy and subsequent heart failure (HF) a transverse aortic constriction (TAC) surgery is conducted in mice. In this HF model, an increase of a certain metabolite has been observed post TAC. In order to decrease the metabolite a catalyst is over-expressed via adeno-associated virus (AAV) mediated gene transfer. The extent of HF is measured via left ventricular ejection fraction (LVEF), which is the primary endpoint. The 2x3 design consists of the operation treatment (TAC/sham) and the vector treatment (AAV+gene, AAV_wo_gene, saline) resulting in 6 groups (5 control groups). The hypothesis is that the specific catalyst will reduce the metabolite and thus increase LVEF. There is already data from an initial study consisting of 8 mice in the TAC_AAV+gene group and 8 mice in the AAV_wo_gene group. In this study only male mice were tested. Goal of the current experiment is to confirm previous findings, add also female mice, add complete control groups (for the full 2x3x2 design).

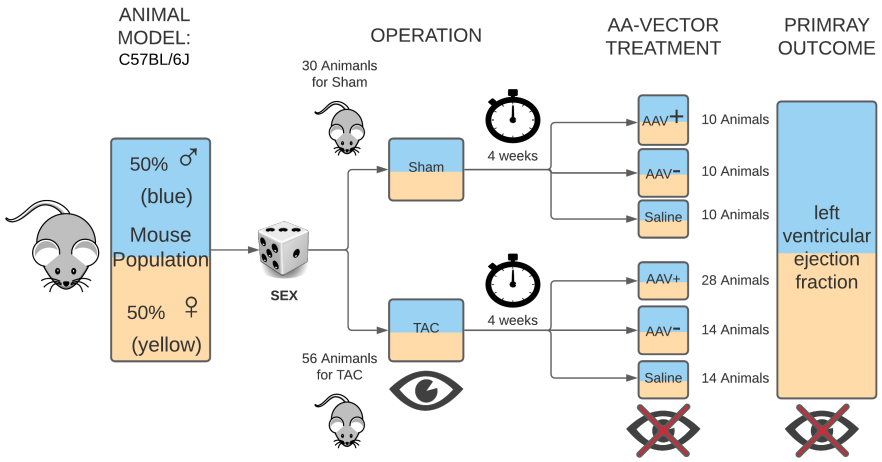

Figure 2: Design flowchart for our confirmatory example: A 2x3x2 Design: Two operation modes (sham/TAC) are measured in three treatments (AAV+gene, AAV_wo_gene, saline) in both sexes (male/female). Crossed eyes symbolize the investigator is blinded to treatment allocation. Build with www.lucidchart.com.

2 Research question

As described previously in chapter 1.4, methodology substantially differs between exploratory and confirmatory experiments. The specific research question or aim of an experiment determines whether it is exploratory or confirmatory in nature. It is therefore essential to define the research question as explicit and precise as possible before any further elaborated planning of the experiment. A common problem is that research questions are stated in a too general way. In order to perform any meaningful statistical sample size planning, it must be possible to translate the research question into a specific statistical hypothesis. Though one can have the same hypothesis for an exploratory as for a confirmatory experiment, only the letter allows confirmatory generalization of results of a statistical test (statistical significance) if adequate sample size planning has been done before. The p-value of exploratory experiments does not allow confirmatory generalization of results.

2.1 Examples

2.1.1 Specific exploratory example

The aim of this animal trial is to analyse the influence of the growth factor GX on fracture healing with special consideration of the immunological experience of the adaptive immune system.
This is an exploratory study aiming at a first quantification of effect sizes. The specific exploratory hypothesis of this experiment is: GX administration in the initial healing phase increases bone fraction of the callus volume at day 21 compared to control mice depleting GX using antibodies.

2.1.2 Specific confirmatory example

The aim of this confirmatory animal trial is to validate that the specific catalyst will reduce the metabolite and thus increase the left ventricular ejection fraction (LVEF) measured via echocardiography. The specific hypothesis to be tested is: LVEF measured 20 days after virus injection is higher in TAC AAV+ mice compared to TAC AAV- mice. Moreover, we would like to explore that the findings are also valid in female mice, and that the findings persist in complete controls.

3 Primary endpoint(s)

All outcome measurements (= endpoints) should be clearly defined a priori (see ARRIVE guideline: item 6 Percie du Sert et al. (2020)). The primary endpoint is the outcome measure that is used to quantify the effect of main interest and answer the primary research question. For confirmatory studies, the primary endpoint is also the outcome measure that is used to determine the sample size. The experiment will be sufficiently powered for the primary endpoint, only, in confirmatory trials. All secondary endpoints can be analysed in an exploratory (mainly descriptive) fashion only. Ideally, there is only one primary outcome measure. This could also be a combination of different e.g. behavioral tests. In this case, the primary outcome could be a specific proportion of behavioral tests failed or a certain constellation of test parameters that are considered to build the primary outcome of interest as a binary variable (constellation reached yes/no). Occasionally, there are two or three primary research questions and thus primary endpoints. Then multiple testing needs to be accounted for by adjusting the significance level in the sample size calculation.
The primary endpoint - as all other endpoints - should be specified precisely and objectively with the exact measurement method, the unit of measurement and the specific point in time when the measurement for the primary research question is taken. The latter is especially important if the study has a longitudinal design and measurements/observations are made at several points in time (i.e. measurements are not independent).

3.1 Examples

Examples for a primary endpoint could be the signal to noise ratio of a fluorescent marker in a specific region of interest normalized to a background region (without unit), the sensitivity of a certain test, the tumor size in mm^3 after 12 weeks of treatment, the occurrence of heart failure within 3 months, progression free survival in days, or changes in a performance score 8 weeks after after treatment.

3.1.1 Specific exploratory example

In our exploratory example the primary endpoint is the bone fraction of the callus volume measured by µCT at day 21. It is calculated as the relative proportion of bone volume (BV) to total callus volume (TV) in percent over a volume of interest in the callus (BV/TV in %).

3.1.2 Specific confirmatory example

In our specific confirmatory example the primary endpoint is the left ventricular ejection fraction (LVEF) in percent measured 20 days after virus injection via echocardiography.

4 Study design

The study design comprises an overview of the study concept (including objectives) as well as the general workflow. In general, it contains information about the type of animal and strain, the number of groups and their constellation (e.g. are male and female animals assigned to different treatment groups, are animals/study units observed over time, etc.). Further, each measurement and its timing should be included and explained, e.g. in a table with explanatory text or a figure. A schematic figure of the workflow is very advisable (see, e.g., www.lucidchart.com, or https://eda.nc3rs.org.uk/). Thus, the study design should be a clear and easily understandable description of the study and is fundamental for both a well devised proposal and an equitable review. In animal testing, most study designs are rather complex and designed with sequential manners in the way that the involvement of further animals/units or even further experiments might depend on the outcome of the respective trial. A complete description of the study design should involve these information as well as a detailed list of stopping or continuing criteria. Two very important design aspect to yield unbiased results (i.e to exclude systematic errors, see also https://catalogofbias.org/) and to ensure equality of handling and observation are randomisation and blinding. Randomisation is the random allocation of animals to their treatment groups. It is advisable to use a computer generated randomisation list and to keep the list disclosed until randomisation for a specific animal occurs. A dedicated person for handling the randomisation list migth be necessary. randomisation in (flexible) blocks is also advisable to reduce foreseeability of allocation. Further, the randomisation should be stratified by important prognostic factors in order to generate comparable groups (with the only difference being the treatment) if possible. Many different software packages and online tools for the generation of randomization lists exists (e.g., R-package *randomizeR* or https://www.randomizer.at/ or the experimental design assistant https://www.nc3rs.org.uk/experimental-design-assistant-eda). The treatment group of the animals should be blinded to the investigator, if possible, to ensure equal observation of all animals.

4.1 Examples

4.1.1 Specific exploratory example

This is a 2x2x3 Design. Two strains of mice in two treatment groups (GX treatment in the initial healing phase/ Gx depleted controls) are measured in three subgroups on day 3, 14 and 21 after osteotomy, respectively. The effect on the inflammatory response is determined by FACS analysis on day 3 and 14, and bone healing in the osteotomy gap is analysed on day 21 by µCT. See figure 1 for the design flowchart of this setting.

Blinding: For FACS analyses and evaluation of µCT images the observer will be blinded, not knowing the corresponding treatment groups.

Randomisation: Within each strain, mice are randomized into treatment and control as well as "day 3", "day 14" and "day 21" subgroups using randomisation lists stratified for male and female mice prepared with the R-package *randomizeR*.

4.1.2 Specific confirmatory example

The study flow diagram for this study with a 2x3x2 group design is shown in Figure 2. Two operation modes (sham/TAC) are measured in three treatments (AAV+gene, AAV_wo_gene, saline) in both sexes (male/female).

Blinding: The conduct of operation procedures will not be blinded. Nevertheless, the assessment of the outcome (echocardiography) will be done blinded. Analysis will be conducted according to the analysis plan outlined here.

Randomisation: We use a stratified randomisation scheme with animal sex and weight in 3 classes with an allocation ratio of 2:1 for TAC groups and controls prepared with the R-package *randomizeR*.

5 Sample size planning

5.1 Sample size planning for exploratory experiments

5.1.1 Goal and general principle

In exploratory research, previous evidence might not exist or if it does it is potentially associated with high uncertainty due to low number of experimental

units. Sample size planning can thus be based on estimating effect sizes with a certain precision (i.e. 95% confidence interval). Similar experiments or literature can serve as a proxy for the range of values to expect. Effect sizes from previous studies (also own), however, often carry the risk of effect size inflation (Colquhoun (2014)), meaning those effects published are likely overestimated. In exploratory analyses it is advisable to define a minimum effect size of interest that the experiment can detect with a certain power. The motivation for this effect size can be based on previously found effect sizes, biological relevance, but also on feasibility by stating what effect size can be sufficiently detected when only a certain amount of animals is available. Importantly, calculations should be done how power changes if the true (not the measured) effect size is actually lower than anticipated. Our general recommendation is that a successful exploratory experiment should be followed up by a confirmatory study that is based on the findings of the initial study considering effect shrinkage (Drude et al. (2022)).

5.1.2 Specific exploratory example

Since this is the first experiment ever to investigate the role of the growth factor GX for fracture healing, no specific assumptions about the effect size can be made, and the sample size is justified via the precision of the estimation. We plan to use eight animals per subgroup and time point resulting in 8x12=96 animals. This is a pragmatic choice based on our experience with similar exploratory experiments. This sample size allows for a sufficient precision of the estimation if the variance is not unexpectedly high: previously, we have observed standard deviations of about 0.2 to 0.3 in this outcome measure. Assuming a common standard deviation of 0.3, a two-sided 95% confidence interval for the difference in means will extend by 0.32 from the observed difference in means, which would be sufficient in our opinion to describe fracture healing effects of GX for the first time.
With 8 animals per group, a two-tailed t-test with significance level 0.05 has a power of 80% to detect a standardised effect size (Cohen's d) of 1.56 using the planning software G*Power. This corresponds, for example, to an expected difference in means of at least 0.5 with an estimated pooled standard deviation (SD_{pooled}) of 0.32 given the formula Cohen's d$= (\mu_1 - \mu_2)/(SD_{pooled})$. For any smaller effect sizes we have less than 80% power with 8 animals per group. Multiple testing is not adjusted for here and our focus lies on estimating effect sizes with 95% confidence intervals (CI).

Dropouts: In our experience with this mouse and intervention, there is a dropout

due to inflammatory reaction in about 10% of animals. We thus need $n_{dropout} = roundup(\frac{8}{(1-0.1)} - 8) = 1$ reserve animal per group resulting in a total of 12 additional animals for the entire trial and an overall number of 108 animals including reserve animals.

5.2 Sample size planning for confirmatory experiments

5.2.1 Goal and general principle

For confirmatory experiments, the goal of the sample size calculation is to ensure that the trial has sufficient power to detect a potentially meaningful effect under a given type 1 error rate. The statistical power represents the chances of a "true positive" detection given that an effect actually exists. Statistical power ranges from 0 to 1 (or 100%) and is typically desired to be at least 0.8 (or 80%). A statistically significant finding is given if the p-value calculated from the data is below the predefined significance level α. In that case the null hypothesis ("there is no effect") is rejected in favour of the research or alternative hypothesis ("there is an effect"). The higher the statistical power, the lower the probability of making a type 2 error β by wrongly failing to reject the null hypothesis, or in other words the lower the chances of a "false negative" finding.
Importantly, only the effect on the primary endpoint is considered for the sample size calculation. Secondary endpoints do not play a role in the sample size calculation for a confirmatory experiment. Therefore the trial will not necessarily be powered for their detection, although methods exist to make additional claims on secondary endpoints. Nonetheless, these have to be prespecified.

The specific calculations that are performed for the sample size calculation depend on the corresponding appropriate statistical test, but the general procedure usually remains the same. The following items are involved and have to be stated explicitly *a priori* (before data are acquired) after having specified the null and the research (alternative) hypothesis:

1. Type I error rate, or the 'significance level': the probability of rejecting the null hypothesis although it is actually true.

2. Power: the probability of rejecting the null hypothesis given the alternative hypothesis is true. This is equal to (1- β) with β being the type II error rate.

3. Effect size: a quantification of the strength of the effect of interest in the primary endpoint. A popular effect size to quantify the difference between two means is Cohen's d but many other effect size measures exist for different situations, e.g. for binary outcomes, survival and comparisons of more than two groups.

4. Sample size: required number of animals per group needed to detect the specified effect size with the desired power and at the significance level chosen.

From these four items, the researcher must specify **three** to be able to calculate the fourth. Typically, type I error rate, power and effect size are specified to calculate the sample size. In principle, any of these items can be calculated given the other three are provided. Thus, with a specific sample size in mind, a prespecified type I error rate and the desired power given, one can also calculate for which minimal effect size the trial is powered for.

In practice, type I error rate and power are often set to 5% and 80%, respectively, and the only factor in the calculation that requires considerable justification is the effect size. How the effect size is specified depends on the statistical test and the software that is used. In many cases it is possible to derive the effect size from the assumed data for the primary endpoint under the alternative hypothesis, e.g. the expected probability of the event to occur, or the expected group means and standard deviations for the control and experimental group. It can sometimes be difficult to define a realistic assumption about the effect size, particularly if limited knowledge about the subject is available. Instead, it may be appropriate to define the effect size as the minimal clinically meaningful or clinically relevant effect that one would like to be able to detect with the specified statistical power. The basis and calculation of the effect size should be given in detail containing relevant references.
If it is not even possible to specify the effect size based on prior knowledge and publications, it is likely that the trial is actually an exploratory study for which the described confirmatory approach is not appropriate.

5.2.2 Specific confirmatory example

In our initial study, we detected an effect size of $d = 1.5$. This measure is derived from *Applicant et al. 2008* Figure 3A: We detected a mean of $\mu = 5$ and a standard deviation of $sd = 2$ in the AAV+gene group and a mean of $\mu = 3$ with an $sd = 2$ in the AAV_wo_gene group. This yields an effect size of 1.5 ($d = \frac{\mu_{AAV+} - \mu_{AAV-}}{\sqrt{\frac{sd^2_{AAV+} + sd^2_{AAV-}}{2}}}$). Due to the pilot character of the exploratory study, we

assume an effect shrinkage (Drude et al. (2022)) and thus assume an effect size $d = 1.0$ for our confirmatory study. We base our power analysis on our main comparison from the exploratory study and conduct a t-test for a mean difference between μ_{AAV+} and μ_{AAV-}. From G*Power(3.1) we calculated that we need 46 animals (23 per group) for the simple comparison (p-value threshold=.05, elevated power for a confirmatory study $\beta = .9$). We aim to have equal number of male and females per group and a dropout rate of 10% (see below), that is we will need 26 animals per group. We have additional controls that are conceptionally very similar. So we will reduce the number of animals in these groups as follows: 28 animals will be distributed to the two TAC control groups (AAV_wo_gene group/saline) (14 each to preserve sex balance). 30 animals for the three sham operation groups with 10 animals in each group. This will lead to an unbalanced design, but will be outweighed by the reduction of number of animals in the control groups. We expect no differences between sham control groups (mean difference $< 10\%$).
We anticipate a droput rate of 10%. Examples for such a drop out rate can be found in the following publication: *Applicant et al. 2012*

6 Statistical analysis

The section "Statistical analysis" of the biometric sheet should briefly the summarise the statistical analysis plan.
In a confirmatory study, the focus is on the analysis of the primary endpoint in terms of sample size calculation, statistical modeling, and inferential methods used for analysis. In an exploratory study, a more comprehensive description of the planned strategies may be provided, especially if several equally important research questions are addressed. The more precisely the analysis plan is outlined, the more reliable and credible the results of the experiment will be. In this sense, a statistical analysis plan is one of the most effective tools to avoid "fishing" for satisfactory results.
Besides choosing the appropriate statistical methods, another aspect of the statistical analysis is the interpretation of potential statistical quantities. When analysing treatment effects, the "significance" of test results (i.e. the p-value) should not be the sole base for decisions and discussions (Colquhoun (2014)). Clinical relevance and precision of estimates (i.e. confidence intervals) also play an important role. A non-significant treatment effect may still be worth further investigation if it is clinically relevant. On the contrary, if an effect is highly significant but not clinically relevant, a further investigation is at least questionable.

In the following, some common aspects of the statistical analysis are outlined. It depends on the specific experiment, which of these are relevant and whether the list is exhaustive.

- **Descriptive statistics.** Usually an important part of the statistical analysis are the descriptive, summary measures. Here, the specification of absolute and relative frequencies for categorical data, mean with standard deviation for sufficiently normally distributed metric data or median with limits of the interquartile range [25th and 75th percentile] for quantitatively skewed data are common standards. In addition, the number of missing values and, if applicable, number of and reasons for drop-outs should be explicitly stated.

- **Primary analysis method.** Deviations from the statistical test used in the sample size calculation should be avoided or must be explained. Possible situations may be a lack of information: For example, a sample size may be calculated using an unpaired t-test and the analysis is planned using an ANCOVA model including treatment group (which would be the primary treatment effect) and baseline measurements (of the continuous outcome variable). The baseline adjusted effect of the treatment might be unknown in the planning phase, so using the t-test is a conservative approach for sample size calculation in such a situation. Further, including the baseline measurement in the actual analysis later might increase the power due to less residual variation.

- **Multiple testing.** If more than one primary endpoint is tested or if claims for secondary endpoints are also important, the type I error (i.e. the probability of finding a false positive result) increases if no proper adjustments are undertaken. Therefore, the planned analysis should account for the type-1 error rate inflation. For this multiple testing problem many solutions exist (e.g. control of family wise error rate, or false discovery rate, hierarchical testing) and the choice of an appropriate approach might dependent on the research question and aim of the trial. The problem of multiple testing is less pronounced in exploratory trials than in confirmatory trials, since in exploratory trials usually the aim is to identify all possibly relevant signals and false positive findings are less severe. Nevertheless, excessive testing in exploratory trials should be avoided and the focus should also be on clinical relevance and precision of estimates For situations with high dimensional data, where there are much more dependent variables than independent observations to be analyzed, multiple testing needs to be accounted for also in exploratory trials.

- **Alternative testing.** It is also possible (and advisable) to include alternative tests if certain prerequisites for tests are not fulfilled (e.g. if the data do not follow a normal distribution). This will increase the transparency and reliability of the analysis further.

- **Repeated measures.** It is often the case that measurements are not independent from each other, e.g. because the same individuals are measured multiple times or because animals that were kept in the same cage are more similar to each other than to animals from another cage. If possible, these cluster effects (observation clustered in a single animal and animals clustered in a cage) should be accounted for in the analysis. If not accounted for, this might lead to severe bias (Aarts et al. (2014)).

- **Handling missing values.** If information on important measurements is missing, this can negatively affect the statistical power and introduce bias. Depending on the mechanism and amount of missing values, a simple complete case analysis might be sufficient, e.g. if information is missing completely at random in only few occasions. In more complex cases, more sophisticated statistical approaches might be needed (e.g. multiple imputation).

- **Adjustment for confounders.** Relevant confounders should be accounted for in the analysis. Since this additionally accounts for variance in the outcome, it usually increases power of the statistical analysis. Such confounders can be any parameters that differ between the groups being compared, such as age and sex in some experiments.

- **Secondary analyses.** While the sample size planning is usually conducted for a single primary analysis, there may be multiple secondary goals of the experiments. Per definition, the sample size is not sufficiently powered for the secondary analyses, and thus these are exploratory with a hypothesis generating character or purely supportive for the findings of the primary analysis. The description of the secondary analyses should at least state the endpoints and the planned statistical methods.

6.1 Examples

6.1.1 Specific exploratory example

Data analysis will be exploratory and mainly descriptive with the report of mean and standard deviation or median and limits of the interquartile range

[25th and 75th percentile] for continuous data and absolute and relative frequencies for ordinal and nominal variables. In addition, the number of missing values and deceased animals is explicitly stated. Differences between experimental groups are reported with 95% confidence intervals. All p-values are exploratory only and do not allow for confirmatory generalisation. Adjustment for multiple testing is deliberately omitted. The focus of the evaluation is on estimating effect sizes with 95% confidence intervals. Exploratory analyses will be marked as such in the publication.

6.1.2 Specific confirmatory example

Descriptive statistics are given as mean and standard deviation or median and limits of the interquartile range [25th and 75th percentile] for continuous data and absolute and relative frequencies for nominal variables. In addition, the number of missing values and deceased animals is explicitly stated. Differences between experimental groups are reported with 95% confidence intervals.
We first calculate a linear model with LVEF as dependent variable and our treatment and sex as independent variables. Although sample size was calculated using the simple t-test, a linear regression will be used for analysis to account for sex as additional co-variate. In this linear model, our main contrast of interest is the post hoc comparison of TAC AAV+ mice vs. TAC AAV- mice. In a follow-up analysis, we will increase the number of factors in the treatment variable to account for the different controls (i.e. instead of treatment and two controls, we will have six factor levels). The experiment is not powered for interactions between independent variables, but we will conduct an exploratory analysis into the interaction between treatment and sex. This will be used for deciding for future experiments and whether this needs to be explored further. All exploratory analyses will be marked as such in the publication.

7 Prerequisites and consequences for subsequent (sub-)experiment(s)

Often animal studies are planned with a distinct order of single experiments. Results of one experiment can have an impact on subsequent experiments, e.g. on the dosage applied, the operational setting used, the number of subgroups investigated, or the time point of investigation. If such a sequential order of the experiments is planned, it should be stated explicitly which conditions from previous (sub-)experiments have to be fulfilled in order to start with the

present (sub-)experiment. Moreover, any conditions that lead to a go/no-go decision must be stated clearly, as well as their impact on further experiments and analyses. Such conditions can, but do not have to be of statistical nature. In some cases, a strict biological/medical justification may be sufficient. For example, stopping an experiment can be based on exceeding pre-specified thresholds on established scores, or if in other ways a treatment turns out to be not tolerable or not effective. Further, there might be arguments from a design perspective to not perform an experiment: if the experiment is clearly a follow-up on a subsequent experiment, it might make no sense to perform the second experiment if the result of the first was negative.
It must be noted that experiments which are stopped before the planned sample size was reached yield biased data. Therefore it is bad research practice to terminate an experiment early once a significant p-value was found (or the data look bad so that a significant p-value appears unlikely), even when the motivation is to save animals. Unplanned changes in the experimental plan must always be described in publications, so that the magnitude of bias can be judged. Although not commonly performed in preclinical research, it is possible to plan an experiment with group sequential testing in the strict statistical sense (Neumann et al. (2017)). Such a planning allows for early stopping based on statistical parameters, e.g. the p-value, under control of type I error and power. In this case, a clear reporting of time points and nature of the interim analyses is required, i.e. under which conditions the experiment will be terminated and how results of this experiment affect the rest of the trial.

7.1 Examples

7.1.1 Specific exploratory example

Subgroups of animals in strain A and B will be subsequently examined on day 3 and 21 irrespective of results in the other strain. No specific prerequisites have to be fulfilled and no go/no-go scenarios are intended as long as the stress on the animals is acceptable (stress score below xxx). The experimental groups with the endpoint FACS analyses on day 14 are only performed if either on day 3 or on day 21 a signal was observed. Otherwise, a relevant signal seems unlikely and these experimental groups are discarded. There is no formal statistical definition of ”a relevant signal” but a brief interim report will be provided to the regulatory authority to justify continuation/stopping of the experiment.

7.1.2 Specific confirmatory example

No sequential approach is planned. Based on the preceding studies, no severe safety issues are expected. In case of unforeseen problems that make it necessary to terminate some parts or the whole experiment early, the regulatory authorities will be notified and the circumstances will be described in the corresponding publication(s).

8 Summary table for the 2 examples

For a detailed overview of the final design, the calculated sample size (including possible dropouts) for each group, assumed effect sizes and/or effects to be estimated as well as the overall (total) number of animals planned are summarized in a table. Examples are provided in Table 2 and 3.

Table 2: Summary table for the exploratory example.

Group	Primary Endpoint	Expected Effect	Reference for expected effect	Effect Size	Drop-Out Rate	Sample Size (including Dropouts)
Strain A: Treatment (GX) day 3 day 14 day 21	BV/TV in %	$mean_{21d}$= 80% (SD: 33%)	Name et al. (2020)	Cohen's d=1.506	10%	$8/0.9 \approx 9$ per subgroup day 3,14 and 21
Strain A: Control day 3 day 14 day 21	BV/TV in %	$mean_{21d}$= 30% (SD: 33%)			10%	$8/0.9 \approx 9$ per subgroup day 3,14 and 21
Strain B: Treatment (GX) day 3 day 14 day 21	BV/TV in %	$mean_{21d}$= 80% (SD: 33%)	Name et al. (2020)	Cohen's d=1.506	10%	$8/0.9 \approx 9$ per subgroup day 3,14 and 21
Strain B: Control day 3 day 14 day 21	BV/TV in %	$mean_{21d}$= 30% (SD: 33%)			10%	$8/0.9 \approx 9$ per subgroup day 3,14 and 21
					Total:	9 x 12 = 108

Table 3: Summary table for the confirmatory example
Numbers have been adjusted for group sizes to be approximately equal between treatment (AAV+/TAC)and the two control conditions (TAC control has 2 subgroups; the sham operated group has three sub control groups).

Group	Primary Endpoint	Expected Effect	Reference for expected effect	Effect Size	Drop-Out Rate	Sample Size (including Dropouts)
AAV+ TAC Operation	LVEF in %	$mean_{LVEF}$= 50% (SD: 22%)	Name et al. (2020)	Cohen's d ≈ 0.9	$< 10\%$	$28/0.9 \approx 32$
AAV-/Vehicle TAC Operation	LVEF in %	$mean_{LVEF}$= 30% (SD: 22%)			$< 10\%$	$28/0.9 \approx 32$ 14 per subgroup + drop out
AAV+/AAV-/Vehicle Sham Operation	LVEF in %	$mean_{LVEF}$= 70% (SD: 22%)			$< 10\%$	$30/0.9 \approx 33$ 10 per subgroup + drop out
					Total:	32+32+33=97

9 Appendix: Glossary

9.1 Glossary

Bias

The term bias describes *systematic* deviations or differences between results of the study and the true population value. An example could be that estimations for a population will potentially be biased if only male mice are analyzed instead of male and female mice. There are many sources of bias, like allocation bias, observer bias, attrition bias or confounding (https://catalogofbias.org/biases).

Biological replicate

A biological replicate is a measurement on the level of the biological unit. The biological unit (BU) is the entity about which inferences are made. If you test only one cell line in an experiment, inferences can be made only about this cell line as biological unit. It is still unclear in how far this extrapolates to other cell lines or humans in general.

Blinding

Blinding describes whether certain people involved in the study have knowledge about allocation of groups. Single-blind refers to the study subjects not knowing whether they are in the treatment or control group, double-blind applies when neither the study subjects nor the treating and outcome assessing researcher do have this information. Triple blinding implies that also the analysing researchers do not have any knowledge about group allocation. Blinding is meant to reduce bias, specifically observer bias. Sometimes, blinding may not be possible by design.

Clinical relevance

An effect (e.g. a difference) is clinically relevant if it is significant from a medical point of view. The assessment is therefore not at the discretion of statistics, but must be evaluated or assessed or determined by medical professionals.

Cohen's d

Cohen's d is a common effect size measure when the difference of two group means is of interest. It is defined as the difference of means divided by the pooled standard deviation of the two groups, i.e. $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2}}$, where $s^2 = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2}$. There, the groupwise variance is defined as $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X}_i)^2$. For one-sample problems this effect size reduces to $d = \frac{\overline{X}-\mu}{\sqrt{s^2}}$, where μ is the assumed true population mean and s^2 the empirical variance of the data.

Confidence interval

The confidence interval is an interval of (un-)certainty. It covers the true population value with probability of $1-\alpha$ (see Significance Level). Contrary to the p-value, the confidence interval reflects the variation of the data and is directly interpretable with respect to the effect. Furthermore, increasing sample size leads to narrower confidence intervals and therefore increasing the precision of the estimate.

Confirmatory study

A study aimed at corroborating empirically specific relationships between defined factors, based on previous (exploratory) observations. Results from con-

firmatory studies may verify a hypothesis and require adequate sample size planning a priori.

Confounder

A distortion that modifies an association between an exposure and an outcome because a factor is independently associated with the exposure and the outcome (https://catalogofbias.org/biases/confounding/). Importantly, a confounder variable is not a consequence of the experimental intervention. In experiments with small sample sizes confounding can be substantial despite randomization and constitutes a source of random error and bias.

Effect

A (statistical) effect is a statistical parameter that is used to quantify and summarize the endpoint of a trial. For example, an effect could be the group mean at a specific point in time, a difference of means, a correlation, an odds ratio or a risk ratio.

Effect shrinkage

Estimated effects from exploratory trials are often overestimated due to small number of subjects, missing prior information about potential effect sizes and publication bias (Colquhoun (2014)). Thus, for planning a confirmatory trial one would assume a smaller ("shrunk") effect than was observed in the prior exploratory trials (Drude et al. (2022)). A rule of thumb for shrinkage is to use approximately $2/3$ of an exploratorly oberved effect size.

Effect size

The terms effect, strength of the effect, effect size are used synonymously. It should be used to illustrate and describe the practical relevance of statistically significant results instead of the p-value. Different effect measures exist to calculate the effect size depending on the variable's scale. An effect size should be derived considering not only a relevant size of an effect but also a variation when observing this effect i.e., the relevant effect in units of the standard deviation. This effect size refers to a statistic which estimates the magnitude of an effect.

Endpoint

Endpoint is synonymous to outcome or measurement of interest (e.g. systolic blood pressure). It describes a variable of interest that the trial wants to investigate. There should be only one primary endpoint and possibly one or more secondary endpoints.

Estimate

An estimate is a value that summarizes an observed sample and is used to approximate the true population value. For example, if a sample mean is equal to 3.14, this specific value is one estimate for the population mean, based on the respective sample used.

Experimental unit

The experimental unit (EU) is the entity that is randomly and independently assigned to experimental conditions, e.g. the cage or each single animal. The EU makes up the sample size (N). The observational unit (OU), on the contrary, is the entity on which observations (or measurements) are made, e.g. several mice in one cage or several organs in one mouse. If observational units are not the experimental unit and thus mistaken as sample size (N), pseudoreplication is introduced and studies are likely underpowered.

Exploratory study

A study aimed at investigating possible relationships between different factors without having strong previous assumptions or statistical hypotheses. An exploratory study may be used to identify confounders (e.g., physiological parameters relevant to the research question). The result of an exploratory study may allow for generation of a statistical hypothesis that can later be tested in a confirmatory study.

Familywise error rate (FWER)

The FWER is the probability of having *at least one* false significant result from multiple hypotheses. This is relevant in multiple testing. Intuitively, this can be thought of as the type 1 error of the collection of hypothesis, or stated colloquially „probability of having some false significant result“.

Fishing for significance

Also known as p-hacking, fishing for significance describes that *only* significant results are reported after multiple testing and analyses have been performed without respective adjustments.

G*Power

G*Power is free software that can be used to perform sample size calculations on your computer (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower).

Hypothesis test

The hypothesis test provides a decision on the basis of a predetermined criterion (e.g. the significance level) whether the null hypothesis can be rejected on the basis of the data and thus the research (or alternative) hypothesis can be accepted or whether the null hypothesis cannot be rejected on the basis of the data.

Independence

(Stochastical) Independence is given if two or more events/outcomes/random variables do not affect or imply each other. For example, two different people flipping a coin each are independent because the toss of one does not affect the toss of the other one. On the other hand, measuring a clinical parameter in one subject at two different points in time is most likely dependent (e.g. high baseline measurements could lead to high follow up measurements).

Internal validity

Internal validity refers to how far measurements in an experiment reflect causal conclusions or mechanisms. That is, do we really measure what we want to measure? A proper experimental design will aim for high internal validity reducing potential risks of bias by introducing for example randomisation and blinding.

Multiple testing

Multiple testing describes the situation where several hypotheses of interest are tested. Without proper adjustment it leads to type 1 error inflation. For

example, this occurs when multiple groups are included in the trial and several pairwise comparisons are conducted. Or when several primary endpoints are compared between groups.

Null hypothesis

In statistics, the null hypothesis (also H_0) is an assumption that may be rejected using an hypothesis test. For superiority studies, the null hypothesis usually states that on average there is no effect, no difference or connection between the groups to be compared. Its counterpart is the alternative or research hypothesis. See also statistical hypothesis.

Objectives

The objectives are related to the hypotheses and therefore to the primary and secondary endpoints. The primary objective is the same as the primary research question and describes what the study is planning to answer.

Power

The probability of rejecting the null hypothesis given the alternative hypothesis is true. This is equal to (1- β) with β being the type 2 error rate.

Precision

Precision refers to the precision of an estimator or an estimate. Thus, it describes how confident we can be about a certain finding, typically given by a confidence interval (where the precision can be measured by the width of the respective interval), the standard deviation or standard error of the estimate.

Preclinical research trajectory

A preclinical research trajectory comprises a cumulative series of experiments (including exploratory and confirmatory) that generate evidence to enable a decision to carry a newly developed intervention forward to clinical testing.

P-value

The p-value is the conditional probability to observe the observed or more extreme (i.e. for example even larger) differences if the null hypothesis (there

are no differences between the groups) would apply. The condition that the null hypothesis is valid must be given in order to interpret the p-value as the named probability (therefore ”conditional” probability). It can be interpreted as the probability of obtaining the observed data if the null hypothesis holds, which means that small p-values are indicative of the alternative hypothesis.

Randomization

A procedure in which subjects (for example, participating patients) are randomly assigned to different treatment groups using a randomization mechanism. This is intended to distribute known and unknown person-related confounders equally between e.g. therapy and control groups.

Reliability

Reliability refers to the consistency in a measurement. In a broader scientific context, this means that a result is reliable if it is consistently replicated. A reliable measurement is not necessarily valid; for example, if results are reproducible, but do not reflect the studied disease pathology. The sample size directly influences the reliability of a result as uncertainty about the measured effect is (in most cases) decreased with increased sample size.

Replicability

Replicability is the ability to obtain similar results throughout a scientific experiment by maintaining the exact same conditions, experimental design, etc.

Replication

The process of confirming previous empirical evidence by means of e.g., using either the same or closely resembling methods, including additional controls and/or conditions, performing previous experiments in different labs or analysing a larger number of samples or animals than in the original study, with the aim of increasing the reliability and reaffirming the reliability of previously observed results. In the framework of preclinical research trajectories, replications are considered part of a confirmation process. Given that a series of experiments is needed to confirm a hypothesis about a directional relationship, replications incrementally solidify evidence supporting (or refuting) an initial claim.

Reproducibility

Reproducibility refers to the understanding of experimental procedures and analyses in such detail that researchers can engage in a replication. Through this researchers can distinguish between variability in results that arise from methodological differences and variability due to sampling variability.

Research hypothesis

The research hypothesis (also called alternative hypothesis, H1) is the counterpart of the null hypothesis and claims in superiority studies that there is an effect, a difference or a connection between the groups being compared.

Sample size or number of cases

The number of cases (n) or the sample size is the number of independent experimental units from a population in a study. In order to estimate statistical parameters with a given accuracy from a sample, sample size calculation/planning is required.

Sensitivity

This term is typically related to diagnostic and prognostic models and describes the proportion of true positives, i.e. the rate of positive diagnoses/predictions, given that the true value is positive as well.

Significance level

The decision limit (α) when a statistical test result is considered significant (when p-value $\leq \alpha$). The significance level is also called the probability of error (α error or type 1 error) of a statistical hypothesis test. It is the maximum tolerated error probability with which the test may erroneously decide in favour of the alternative hypothesis if the null hypothesis was true in the population.
Usually a significance level $\alpha = 0.05$ is defined, i.e. a maximum of 5 false positive statistically significant test results out of 100 tests on different samples from the population is tolerated.
If the null hypothesis is rejected with a significance level of 5%, there is a 5% probability of error, i.e. 5% of type 1 errors are made.

Smallest effect size of interest (SESOI)

The SESOI is the smallest effect size that is considered theoretically and/or practically interesting and can be taken to justify the sample size for a given experiment. To determine the SESOI, previous evidence as well as practical aspects (e.g., feasibility) may be considered.

Specificity

This term is typically related to diagnostic and prognostic models and describes the proportion of true negatives, i.e. the rate of negative diagnoses/predictions, given that the true value is negative as well.

Statistical hypothesis

The statistical hypothesis is a reformulation of the research question to connect it to a statistical hypothesis test. A statistical hypothesis must satisfy at least two conditions: (i) There has to be a statistical parameter contained (e.g. the mean, proportion, odds ratio, regression coefficient, event rate...) and (ii) together with its counterpart hypothesis it must exhaust the whole possible parameter space. The latter can be confirmed if either the null or the alternative hypothesis must be true with no third option available (e.g. $\mu \leq 3$ vs. $\mu > 3$).

Statistical inference

Statistical inference describes the process of inferring information from a sample to a population via statistical tools. A non-significant treatment effect may still be worth further investigation if it is clinically relevant. On the contrary, if an effect is statistically significant but not clinically relevant, a further investigation is at least questionable.

Statistical significance

The result of a statistical test is called statistically significant if the calculated p-value is less than the pre-defined significance level α. In this case, the sample data are less likely given the null hypothesis was true than the previously defined probability of error for a false positive result (usually 5%), such that the null hypothesis is rejected and the alternative hypothesis is accepted to be true. A confirmatory generalization of a statistically significant test result (meaning that the alternative hypothesis is valid in the entire population) is

only meaningful in confirmatory experiments. Exploratory experiments do not allow confirmatory generalization of results. A non-significant treatment effect may still be worth further investigation if it is clinically relevant. On the contrary, if an effect is statistically significant but not clinically relevant, a further investigation is at least questionable.

Technical replicate

Technical replicates are replicates regarding one biological unit and can form several levels across a hierarchy (dependent experiments on the same day, or independent experiments on different days, but with the same cell line, for example). Important hierarchy levels are the experimental unit and the observational unit. As this term can be ambiguous as to the level addressed the exact level should be stated when introducing technical replicates.

Type 1 error

In hypothesis testing, a type 1 error occurs when the null hypothesis (drugs act in the same way) is wrongly rejected and the alternative hypothesis (drugs act differently) is accepted, although in reality the null hypothesis is true (corresponds to a false positive result). The maximum error probability for type 1 error is determined with the significance level α (usually 0.05 or synonym 5%) under the assumption that the null hypothesis applies.

Type 1 error inflation

This error inflation occurs in multiple testing situations. If no adjustments are made the familywise error rate increases with an increasing number of hypothesis tests performed. As a result, the probability of obtaining false-positive results increases.

Type 2 error and power

Whereas the type 1 error rate describes the rate of falsely rejecting the null hypothesis, the type 2 error describes the rate of not rejecting the null hypothesis which is actually false. This might be, for example, if a statistical test does not detect an actual effect. Complementary, the power of a statistical test describes its ability to detect effects, i.e. proportion of correct rejections of the null hypothesis. Thus, if the type 2 error is β, the power is $1-\beta$, typically required to be 80%.

References

Emmeke Aarts, Matthijs Verhage, Jesse V Veenvliet, Conor V Dolan, and Sophie Van Der Sluis. A solution to dependency: using multilevel analysis to accommodate nested data. *Nature neuroscience*, 17(4):491–496, 2014.

David Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*, 1(3):140216, 2014.

Ulrich Dirnagl. Resolving the tension between exploration and confirmation in preclinical biomedical research. *Good Research Practice in Non-Colinical Pharmacology and Biomedicine. Springer*, pages 71–79, 2019.

Natascha Ingrid Drude, Lorena Martinez-Gamboa, Meggie Danziger, Anja Collazo, Silke Kniffert, Janine Wiebach, Gustav Nilsonne, Frank Konietschke, Sophie K Piper, Samuel Pawel, et al. Planning preclinical confirmatory multicenter trials to strengthen translation from basic to clinical research–a multi-stakeholder workshop report. *Translational Medicine Communications*, 7(1):1–13, 2022.

Jonathan Kimmelman, Jeffrey S. Mogil, and Ulrich Dirnagl. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. *PLOS Biology*, 12(5):e1001863, May 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001863. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001863.

Konrad Neumann, Ulrike Grittner, Sophie K Piper, Andre Rex, Oscar Florez-Vargas, George Karystianis, Alice Schneider, Ian Wellwood, Bob Siegerink, John PA Ioannidis, et al. Increasing efficiency of preclinical research by group sequential designs. *PLoS biology*, 15(3):e2001307, 2017.

Nathalie Percie du Sert, Viki Hurst, Amrita Ahluwalia, Sabina Alam, Marc T Avey, Monya Baker, William J Browne, Alejandra Clark, Innes C Cuthill, Ulrich Dirnagl, et al. The arrive guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, 40(9):1769–1777, 2020.

Sophie K. Piper, Dario Zocholl, Ulf Toelch, Robert Roehle, Andrea Stroux, Johanna Hoessler, Anne Zinke, and Frank Konietschke. Statistical review of animal trials—a guideline. *Biometrical Journal*, n/a(n/a), 2022. doi: https://doi.org/10.1002/bimj.202200061. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202200061.