Received: 26 February 2022 | Revised: 25 July 2022 | Accepted: 14 August 2022

DOI: 10.1002/bimj.202200061

Biometrical Journal

SHORT COMMUNICATION

Statistical review of animal trials—A guideline

Sophie K. Piper[1,2,3] | **Dario Zocholl**[1,3] | **Ulf Toelch**[4] | **Robert Roehle**[1,3,5] | **Andrea Stroux**[1,3] | **Johanna Hoessler**[6] | **Anne Zinke**[6] | **Frank Konietschke**[1,3]

[1]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

[2]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Berlin, Germany

[3]Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany

[4]Berlin Institute of Health at Charité - Universitätsmedizin Berlin, QUEST Center for Responsible Research, Berlin, Germany

[5]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Clinical Trial Office, Berlin, Germany

[6]Landesamt für Gesundheit und Soziales, Referat für gesundheitlichen Verbraucherschutz, Berlin, Germany

Correspondence
Sophie K. Piper, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany.
Email: Sophie.Piper@charite.de

Frank Konietschke, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany.
Email: Frank.Konietschke@charite.de

Funding information
BMBF, Grant/Award Number: 01KC1901A; Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG KO 4680/4-1

Abstract

Any experiment involving living organisms requires justification of the need and moral defensibleness of the study. Statistical planning, design, and sample size calculation of the experiment are no less important review criteria than general medical and ethical points to consider. Errors made in the statistical planning and data evaluation phase can have severe consequences on both results and conclusions. They might proliferate and thus impact future trials—an unintended outcome of fundamental research with profound ethical consequences. Unified statistical standards are currently missing for animal review boards in Germany. In order to accompany, we developed a biometric form to be filled and handed in with the proposal at the concerned local authority on animal welfare. It addresses relevant points to consider for biostatistical planning of animal experiments and can help both the applicants and the reviewers in overseeing the entire experiment(s) planned. Furthermore, the form might also aid in meeting the current standards set by the 3+3R's principle of animal experimentation: Replacement, Reduction, Refinement as well as Robustness, Registration, and Reporting. The form has already been in use by the concerned local authority of animal welfare in Berlin, Germany. In addition, we provide reference to our user guide giving more detailed explanation and examples for each section of the biometric form. Unifying the set of biostatistical aspects will help both the applicants and the reviewers to equal standards and increase quality of preclinical research projects, also for translational, multicenter, or international studies.

KEYWORDS
animal experiment, Animal Welfare Act committee, ethics approval, preclinical research, statistical planning

1 | INTRODUCTION

Ethical approval of any experiment involving living organisms is a necessity. It requires justification of need and moral defensibleness through harm benefit assessment of the study. Beyond medical and ethical considerations, design, sample size calculation, and the statistical analysis plan of the experiment constitute important review criteria (Kilkenny et al., 2009; Vollert et al., 2020). According to the German Animal Welfare Act as a federal law, each federal state needs a local animal welfare authority to approve all §8 experiments involving live vertebrates or cephalopods. This law also requires that these animal experiment applications be reviewed by an independent §15 Animal Welfare Act Commission (AWAC) as an advisory body. AWACs have the option of submitting an opinion to the regulatory authorities with their main recommendations for approval or rejection of the application, but the final review and decision rests with the authority concerned. This differs from the tasks of the ethics committees established for clinical trials, which have approval powers. Up to now trained biostatisticians are still underrepresented members of AWACs. An unpublished national survey revealed that only 8 out of 34 (∼24%) active committees for animal research in Germany have appointed a statistician in 2020 (information for one additional committee was not available). Moreover, the general application forms for animal trials vary nationwide and differ in the nature of statistical questions and the depths of statistical requirements.

In comparison with clinical trials, preclinical experiments often split up into several subexperiments across several model systems to support knowledge claims. Due to the complex study structures statistical evaluation demands expertise. Robust results depend on meaningful and efficient statistical planning of all subexperiments. Here, *efficiency* relates to the trade-off between ethical concerns (reduction of animals) and reliability of the statistical methods (in the sense of uncertainty reduction) (Festing et al., 1998; Festing, 2018). Often, animal testing is conducted with very small sample sizes, for example, 10 or even less animals per group (Bonapersona et al., 2021). On the contrary, many statistical methods require medium or large sample sizes for an accurate type-1 error rate control. In case of (very) small sample sizes and when test assumptions are not met, the test procedure might be liberal and thus might tend to overreject the null hypothesis. Errors made in the statistical planning and data evaluation phase can have severe consequences on both results and conclusions (Strech & Dirnagl, 2019). They might proliferate and thus impact future trials—an unintended outcome of fundamental research with profound ethical consequences. Therefore, animal experiments must be efficient in both medical and statistical ways (Ludolph et al., 2010). However, the complexity and the exploratory nature of the studies typically makes the statistical planning a rather challenging but nevertheless nonnegligible task (Kimmelman et al., 2014; du Sert et al., 2020b). Existing and well-established clinical trial criteria like randomization, blinding, preregistration, interim analysis are also applicable in preclinical phases (Festing & Altman, 2002; du Sert et al., 2020a; Vollert et al., 2020). All in all, statistical aspects are key quality criteria and every AWAC should be aware of their importance.

It was our initiative in 2018 to collaborate with the concerned local authority of animal welfare in Berlin, Germany, since we identified an urgent need to improve application quality during our work for the local AWAC as well as our consultancy service at the Charité university hospital (we consulted 116 applications for animal experiments in 2020). We have frequently observed application forms for animal trials that were unclear in design, imprecise in stating the research question and hypotheses, deficient in reporting of statistical planning and unfortunately also sometimes simply inappropriate by pasting text blocks from former applications.

With this paper, we aim to raise awareness for the need of statistical expertise in AWACs and introduce a form sheet addressing relevant points to consider for biostatistical planning of animal experiments. The form has already been in use by the concerned local authority of animal welfare in Berlin, Germany, and is available online (in German) https://www.berlin.de/lageso/_assets/gesundheit/publikationen/biometrie-formblatt.docx. Unifying the set of biostatistical aspects will help both the applicants and the reviewers to equal standards and increase quality of preclinical research projects, also for translational, multicenter, or international studies (Festing & Altman, 2002). Furthermore, the form might also aid in meeting the current standards set by the **3+3R**'s principle of animal experimentation **R**eplacement, **R**eduction, **R**efinement (Guhad, 2005) as well as **R**obustness, **R**egistration, and **R**eporting (Strech & Dirnagl, 2019).

In the following, all criteria will be explicitly presented and explained in Section 2 followed by an specific, fictitious exploratory example in Section 3. Furthermore, we discuss first experiences from both applicants and reviewers in

Section 4. In addition, we provide a reference to our user guide giving more detailed explanation and examples for each section of the biometric form (see doi https://doi.org/10.5281/zenodo.7038608 - to be announced).

2 | BIOMETRIC CRITERIA

As mentioned before, we recall important statistical review criteria in this section. All of them are listed in the current biometric planning form used by the concerned local authority of animal welfare in Berlin, Germany. The layout of the form helps both the applicants and reviewers to summarize the primary goal of the study, its (statistical) planning, sample size calculation, and verification in a precise, specific, and condense way. In the following, we explain the individual items briefly:

1. **Goal of the (sub)trial**: A precise description of the (primary) study goal is key for successful implementation and review. This is reflected in item 13 of the ARRIVE guidelines 2.0: "Objectives: Clearly describe the research question, research objectives and, where appropriate, specific hypotheses being tested" du Sert et al. (2020b).
   Typical goals in the field of preclinical studies are investigation of biological or pathophysiological processes, drug testing, dose-finding, toxicological screenings, or generating novel animal models. Less often the goal is confirmatory testing of a treatment effect. At this stage, indicating whether the trial is "exploratory" or "confirmatory" is mandatory. Whereas exploratory research aims at generating new hypothesis, confirmatory research will test these novel hypotheses in a robust manner. Even though there is a gradient between the two, researchers should be aware that confirmatory studies are characterized by increased reliability and validity of experimental evidence. This is particularly important with regard to the translation of results from animal models into clinical contexts, a step which we requires greatly increased validity of preclinical studies, as one of the authors has discussed more elaborate in another publication (Drude et al., 2021). The distinction between research conducted in exploratory or confirmatory mode has an influence on all biostatistical criteria. One important aspect is the balance between Type I errors (false positives) and Type II errors (false negatives). In many biomedical fields, prior probabilities of a hypothesis being true are low. With low numbers of experimental units (i.e., in most cases animals) in exploratory studies the chances of a Type II error are high. This carries the risk of discarding viable hypotheses too early in a research trajectory. That means stringent null hypothesis testing with a focus on standard p-value thresholds in exploratory research will eventually result in a low positive predictive value. Research conducted in exploratory mode should thus focus on high internal validity by, for example, reducing risk of bias. Whether an exploratory result is based on a true (i.e., biologically meaningful) effect or is false positive is almost impossible to justify. Evidence in such experiments should be judged on effect sizes and uncertainty rather than stringent p-value thresholds. Confirmatory experiments are then needed to collect robust evidence with regard to a specific hypothesis.
   In our statistical consultations with applied researchers, we often recommend studying Ioannidis (2005) for why exploratory research suffers from low positive predictive value and Kimmelman et al. (2014) and Mogil and Macleod (2017) to grasp the distinction between exploratory and confirmatory research.
2. **Primary endpoint** of each subexperiment:
   The primary endpoint is the outcome measure that is used to answer the primary research question and it is also the endpoint the study is powered for. It should be specified precisely 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 observations are made at several time points.
3. **Description of the study design**: A detailed description of the study design is necessary to characterize whether it is suited to reach the study goals. Design aspects as well as methods against bias (blinding and randomization) should be taken into account.
   (a) **Design**: A detailed description of the study design is fundamental. A flowchart is a useful and helpful tool to illustrate design and workflow of all experimental groups involved in the trial (e.g., https://www.nc3rs.org.uk/experimental-design-assistant-eda). Moreover, the role of any control group should be explained in detail (positive, negative, vehicle, etc.).
   (b) **Blinding**: Blinding means that information about treatment allocation is withheld from certain or all investigators involved with the aim of minimizing information bias and enhancing the study quality. Applicants should distinguish between blinding for conduct of the experiment, assessment of outcome, and analysis. Procedures should be described in detail as in most cases blinding involves several parties that need to be coordinated. In case

blinding is not applicable, the reasons must be stated and further methods against bias should be suggested and justified.

(c) **Randomization**: Randomization describes the process of randomly allocating the treatments to the study subjects with the aim of minimizing selection bias and distributing potentially influential parameters evenly across groups. Systematic differences between experimental divisions may induce bias and thus might impact the results and aid false conclusions. Furthermore, since animals are typically kept in shared cages, possible cluster effects should be taken into account. This applies also to physiological factors like, for example, weight that may have an influence on the primary outcome. Here, block randomization is necessary. As for blinding, details of the randomization procedure have to be provided. If randomization is not applicable, the reasons must be stated and further methods against bias and confounding should be provided and justified.

In preclinical research, randomization seems not to be as fully integrated as in clinical phases and its impact is often underestimated. We recommend Hirst et al. (2014) as illustration for the importance of randomization in animal stroke studies. Some very important concepts of randomization, which are often unclear to the researchers in their practical consequences and implications, are discussed in Festing (2020).

4. **Sample size calculation**: The exact number of experimental units allocated to each group as well as the total number of animals in each experiment should be specified (corresponding to ARRIVE guideline item 2a) (du Sert et al., 2020b). Here, the experimental unit is defined as the unit that is randomly and independently allocated to intervention groups. This is essential to prevent pseudo-replication and assures proper definition of degrees of freedom in the analyzes. For an in-depth discussion of the differences between biological, experimental, and observational units, see Lazic et al. (2018). Moreover, details of the sample size calculation should include the following information:

(a) **Confirmatory trial**: A summarizing statement giving the name of the statistical test used for sample size calculation, the chosen significance level (α) with justification if and how multiple testing was adjusted for, the desired power ($1-\beta$) with β being the type II error rate, whether a one- or a two-sided will be used, and, the physiologically relevant (minimal) effect size with regard to the primary endpoint that is planned to be confirmed. For the latter, it is important to state how this effect size was derived, that is, which expected mean and standard deviation (SD) per group are assumed. If possible, this should be backed up by at least one published reference. Since effect size estimates from exploratory studies are often inflated (Colquhoun, 2014) these should be conservatively evaluated if used for subsequent confirmatory trials. In addition, the software used to calculate the sample size (including version number) should be given. Here, a screenshot of the calculation can be helpful for the reviewer. As confirmatory trials are often very similar to previous exploratory studies, care should be taken to outline the exact goal of such a trial. A simple repetition of an experiment with the same number of experimental units is from an animal welfare and evidence generation point of view not desirable (Piper et al., 2019). It needs to be outlined how the confirmatory evidence is complementing exploratory evidence by stating if and how evidence of the two studies will be jointly (e.g., meta-analytically, sceptical p-value (Held et al., 2022)) analyzed. Here, as in all experiments involving animals, ethical considerations such as reducing the number of tested animals need to be carefully balanced (Strech & Dirnagl, 2019).

(b) **Exploratory study**/ pilot study or orientation trial/ preliminary technical test:

If possible, a priori sample size calculation should be done giving all required details stated above for confirmatory trials. Justification of the sample size can also be based on feasibility within a given time frame and laboratory or on the smallest effect size of interest (Danzinger et al., 2022). For pilot experiments like surgical training, technical feasibility, or purely qualitative endpoints, we recommend at least three animals. Adjustment for multiple comparisons is not necessary in our view, but a cautious interpretation of the exploratory p-values should be warranted. If a formal calculation of the sample size is not possible, it should at least be explained how the sample size was derived. Moreover, it should be stated what confidence interval width for the effect estimate (e.g., group means, survival rates, or the proportion of a binary endpoint) can be achieved with the chosen sample size. This is also to get a notion for what size of potential group differences the study is sensitive for or with what precision estimates can be obtained.

In clinical pilot trials that wish to estimate mean and variance of a metric endpoint, 12 samples per group have been recommended (Julious, 2005). As variance is potentially smaller in genetically almost identical rodents compared to humans, we refrain from recommending a fixed minimum sample size. If the trial aims at estimating an effect size for future confirmatory trials, the width of the two-sided $(1-\alpha)$ Wald-type confidence interval $CI = [Mean \mp \frac{1.96}{\sqrt{n}} \cdot SD]$, which is $2 \cdot \frac{1.96}{\sqrt{n}} \cdot SD$, is a useful criterion for sample size justification. For example, its width is 1 SD if

$n = 16$ for any arbitrary metric endpoint, as pointed out by Festing (2018). Note, any value smaller than $n = 16$ will result in an interval, which is wider than 1 SD.

Whereas some exploratory studies potentially inform confirmatory trials, many exploratory studies aim to gain a mechanistic understanding often requiring a considerable number of, for example, knockout strains. In such cases, animal numbers are rapidly increasing potentially exacerbated by complicated breeding schemes. Here, a balance has to be struck whether less animals per group will also yield meaningful insights.

Caveat: p-Values derived from exploratory experiments are not suited to enable decisions toward further experiments like confirmatory studies. Ideally, go/no-go criteria for engaging in confirmatory studies should be defined a priori (Albers & Lakens, 2018; Drude et al., 2021).

In addition, irrespective of "the exploratory or confirmatory" nature of the study the number of required reserve animals or dropouts due to premature death, incorrect interventions, and so forth, should be given. Very high dropout rates may be a reason to reject the proposal and could be controlled for by stopping rules. Reporting a dropout rate only is not sufficient but the absolute number of additional animals should be stated explicitly. Of note, the absolute number of dropouts is often falsely calculated. If the estimated dropout rate r, for example, 0.2 corresponding to 20% and the required sample size n resulting from the sample size calculation are given, the correct number of additional animals needed would be $n_{dropout} = roundup(\frac{n}{(1-r)} - n)$ or for the total number needed: $n_{total} = n + n_{dropout} = roundup(\frac{n}{(1-r)})$.

5. **Statistical Analysis**:

A brief summary of the planned methods of analysis in this (sub)trial should be given including descriptive statistics. Most importantly, this section should contain the analysis of the primary endpoint, which should be in line with the statistical test used to calculate the sample size. Moreover, variables with a relevant effect on the endpoint, baseline measurements, confounder, repeated measures (several measures often over time on the same experimental unit), or site and cluster effects must be considered. The point of repeated measures often requires specific attention. In some experiments, it is not clear whether measures are actually repeated in the same experimental unit, or whether the design is in fact factorial, for example, when animals have to be killed in order to collect the data at each time point. It is important that the experimental design makes this very clear and that the statistical analysis is in line with the experimental design.

Particularly when small sample sizes preclude inclusion of aforementioned parameters (overparameterization), the specific model choice should be discussed. It must also be reported how missing data will be dealt with or why no missing data are expected. If several subgroups are being compared, it should be declared if and how the analyzes will be adjusted for multiple testing. Furthermore, all secondary analyzes should at least be briefly reported. This could be, for instance, any exploratory subgroup analyzes, sensitivity analyzes, further statistical modeling, and graphical illustrations planned.

6. **Is there a logical or sequential order of the experiments planned?**

In contrast to clinical trials, preclinical studies (animal testing) are an ensemble of multiple (sub)experiments, which may or may not condition each other. Results of one experiment can have an impact on subsequent experiments, for example, on the dosage applied, the operational setting used, the number of subgroups investigated, or the time point of investigation, but it may also be the case that a specific (sub)experiment is only conducted if a preceding experiment has shown a specific signal. If experiments follow a logical order and/or a sequential or adaptive planning, any conditions that lead to a go/no-go decision as well as their impact on further experiments must be stated clearly. Such conditions can but do not have to be of statistical nature. In some cases, a strict biological/medical justification might be sufficient, for example, to stop based on exceeding prespecified thresholds on established scores, or, for instance, if in other ways a treatment turns out to be not tolerable. Furthermore, 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. Stating such conditions beforehand is highly preferable from an ethical point of view, since it makes the total number of animals used in all experiments more transparent and may save a significant number of animals.

Although this is rarely the case in preclinical research, group sequential testing in the strict statistical sense might be planned on (Neumann et al., 2017). In contrast to the aforementioned kind of sequential planning, group sequential tests allow to terminate data collection early for reasons of efficacy or futility based on statistical criteria, which are calculated from data on the very same experiment. This distinction may seem trivial to the experienced statistician but in our experience it is not necessarily obvious for the applied researcher, and within the review process of applications

TABLE 1 Summary table

Group	Primary endpoint	Expected effect	Reference for expected effect	Effect size	Drop-out rate	Sample size (including Dropouts)
Treatment (GX)	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						
Day 14						
Day 21						
Control		$mean_{21d} = 30\%$ (SD: 33%)			10%	$8/0.9 \approx 9$ per subgroup
Day 3						
Day 14						
Day 21						
					Total:	$9 \times 6 = 54$

or statistical consultations it must be made clear that early stopping based on p-values without rigorous statistical planning is not admissible. Experiments with group-sequential planning require clear statements about decision rules, boundaries, as well as minimum, maximum, and expected sample sizes under the null and alternative hypothesis. Elaborate adaptive designs are becoming increasingly prominent in clinical trials and may have multiple design features, such as response-adaptive randomization, dynamic borrowing, adaptive enrichment, sample size recalculation, or shared controls. In our perception, the scientific community has not yet adopted these developments for animal experiments, and the current concepts of sequential experiments are statistically much more basic. More statistical research in this direction is required before it can be expected that animal experiments may be conducted according to a complex adaptive design.

7. **Summary of the sample size calculation for each (sub)trial**:
   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 are summarized in a table. An example is provided in Table 1.
8. **Signatures**: Statistical expertise during the planning phase of the experiment will enhance the quality of the trial. Having the form signed by a biostatistician involved in the planning likely reduces the number of revisions and inquiries by the concerned local authority of animal welfare. In our form, the signature is facultative. Furthermore, indicating the day the biometrician has seen the (last) version of the proposal might be helpful.
9. **Preregistration**: In addition, though this was not included in the biometric from sheet, we highly recommend to preregister the planned trial. Preregistration is an important step towards reproducible and efficient research. So far it cannot be legally required by the animal welfare authorities at the time of application. In preregistrations, important information about the experimental design and analysis is prespecified and digitally saved in an appropriate portal. One such portal is operated by the Bundesinstitut für Risikobewertung (Federal Insitute for Risk Assessment) https://www.animalstudyregistry.org/. Currently, information within a preregistration is under embargo for up to 5 years and can only be accessed by the investigators and later editors and reviewers. Only after exemplary publication the preregistration is accessible for third parties making scooping near impossible. Preregistrations also do not preclude exploratory research but help distinguish confirmatory from exploratory aspects of a study.

3 | SPECIFIC EXAMPLE FOR AN EXPLORATORY TRIAL

The following is a fictitious example of how to fill in our biometric planning form. If several experiments are planned within one experiment, the form should be filled in for each experiment in consecutive numerical order separately.

We chose an exploratory experiment in basic research which shall investigate the influence of a growth factor GX on fracture healing. No specific preliminary data are available in mice, but there is evidence that increased GX levels are associated with reduced bone formation and that regulation of GX could be used to influence fracture healing. To assess the dynamics of the healing process, the bone fraction of the callus volume will be measured by μCT at 3, 14, and 21 days and compared with GX-deficient control animals. See Figure 1 for an illustration of the design. In addition, histological and immunohistological analyzes, flow cytometry (FACS analyzes) and assessments of immunologically relevant organs such as spleen, lymph nodes, and bone marrow are planned so that the animals have to be killed at each time point.

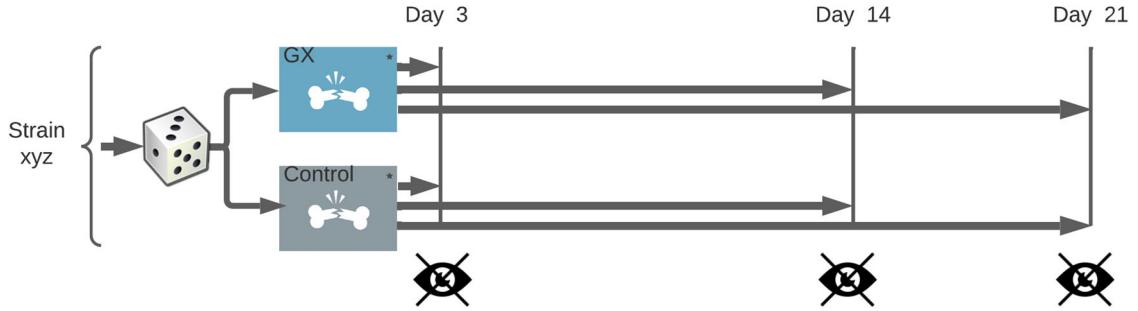

FIGURE 1 Design flowchart for our exploratory example: A 2 × 3 design. Two treatment groups of mice (GX treatment/ GX depleted controls) are measured in three subgroups on day 3, 14, and 21 after osteotomy, respectively. Crossed eyes symbolize the investigator is blinded to treatment allocation. Build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

The header of the biometric form summarizes general information:
Biometric Form
Title of the research project: Influence of growth factor GX on bone healing
Experiment number: 1
Animal species: *Mus muculus* (Strain:xyz)
Number of animals per investigated group in this experiment: 8
Total number of animals (including dropouts) in this experiment: 48

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):
   The aim of this animal trial is to analyze 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. **Primary endpoint of the experiment** (with unit of measurement, measurement method, and time of measurement):
   Primary endpoint is 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. **Description of the study design** (e.g., which groups are compared, which interventions are performed, when is the outcome measured?)
   (a) **Design**: (if possible with flowchart)
   This is a 2 × 3 Design. Mice are randomized into two treatment groups (Group A: GX treatment in the initial healing phase; Group B: GX depleted controls) and 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, while bone healing in the osteotomy gap is analyzed on day 21 by µCT. See Figure 1 for an example design flowchart for this setting built with www.lucidchart.com.
   (b) **Blinding**: (e.g., double-blind or evaluation blinded; if no blinding is done, please explain why)
   For FACS analyzes and evaluation of µCT images, the observer will be blinded, not knowing the corresponding treatment groups.
   (c) **Randomization**: (if no randomization is used, please explain why)
   Within each strain, mice are randomized into treatment and control as well as "day 3," "day 14," and "day 21" subgroups using randomization lists stratified for male and female mice using the R-package `randomizeR`.
4. **Sample Size Calculation**:
   (a) **For a confirmatory trial**: not applicable here.
   (b) **Exploratory trial, pilot study, or orientation test/technical preliminary test**:
      (i) **Sample size calculation or explanation** (e.g., feasibility, precision of estimation that can be achieved with given sample size):

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. 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 *SD*s of about 0.2–0.3 in this outcome measure. Assuming a common *SD* of 0.33, a two-sided 95% confidence interval for the difference in means will extend by 0.323 from the observed difference in means, which would be sufficient to describe this effect for the first time. Multiple testing is not adjusted for here and our focus lies on estimating effect sizes with 95% confidence intervals.

(ii) **The required number of reserved animals/dropouts** (due to premature death, faulty interventions, etc. Specify a dropout rate and reason as well as the required absolute number of animals): 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 six additional animals for this trial.

(iii) **Software used for sample size calculation** (including version number): nQuery version 9.1.1.0.

5. **Statistical analysis**: (e.g., type of statistical modeling, adjustment for potential bias, adjustment for multiple testing, secondary analysis, handling of missing values?): Data analysis will be exploratory and mainly descriptive with the report of mean and *SD* 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. Here, *p*-values are exploratory only and do not allow for confirmatory generalization. Adjustment for multiple testing is deliberately omitted. The focus of our analyzes lies on estimating effect sizes with 95% confidence intervals.
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): Subgroups of animals will be subsequently examined on day 3, 14, and 21. No specific prerequisites have to be fulfilled for these and no go/no-go scenarios are intended as long as the stress on each animal is acceptable (according to scoring sheet XYZ).
7. **Summary table for the sample size planning**:

4 | DISCUSSION

For clinical trials (including early and later phase), ethical review boards are well established (Buchner et al., 2019; Doppelfeld & Hasford, 2019). Thorough reviews include verification of the need of the study, its correct implementation along with proper risk/benefit assessments. Especially, the latter involves attesting correct choice and usage of the statistical methods applied for planning and analysis of the data. Whereas ethics commissions for clinical trials often appointed a trained statistician (Rauch et al., 2020), this is usually not the case for ethics commissions with emphasis on preclinical and animal trials. About 80% of the current German animal ethic commissions do not involve a biostatistician (unpublished survey in 2020 with missing information of one federal state). In principle, according to §42 of the regulations on the welfare of animals used for experiments or for other scientific purposes (Tierschutz-Versuchstierverordnung - TierSchVersV) they consist of scientists and members appointed by animal welfare organizations. All members of the AWAC must be able to understand and judge animal experiment applications based on their experience. Further requirements for participation in AWACs are legally not regulated. Moreover, unified standards for statistical planning are currently missing, but there is high demand for statistical skills (including scientific planning and analysis and the revision by the authorities). This impression was confirmed by many statisticians on several national conferences, where we presented our biometrical form. In Germany, the general application forms for animal trials as well as the required reporting of biometrical planning in particular vary nationwide, depending on the federal state, where the animal trial is conducted. Ultimately, the authorities decide individually which forms are used. For example, the biometric form which is currently used in Tuebingen (Baden-Wuerttemberg) differs from the one used in Berlin, in terms of design aspects like randomization, blinding, and the impact of sequential order of experiments. Moreover, we put emphasis on explicitly distinguishing between exploratory and confirmatory studies. This is important, because biometrical planning of confirmatory studies should be as thoroughly and precise as for clinical trials (Festing & Altman, 2002). Especially, type I error inflation due to multiple hypothesis testing needs to be addressed in confirmatory animal trials. In exploratory (hypothesis generating)

settings, on the other hand, adjustment for multiple testing is not necessary in our opinion as long as *p*-value interpretation is done with caution and marked as "exploratory" in any publication as required by the ARRIVE guidelines (Kimmelman et al., 2014; du Sert et al., 2020b). We further note that other guidelines for animal welfare exist, see, for example, the OECD (Organization for Economic Cooperation and Development) guideline for acute oral toxicity (OECD, 2001). The biometrical planning form used in Berlin until 2020 distinguished between "orientation study" and "comparison study" only, which was not intuitive to the applicant and did not touch the aspect of confirmatory versus exploratory statistical comparisons. Other biometrical planning forms currently in use in Germany distinguish three or four of the following study types: (a) Technically required preliminary experiment in which the animals serve to obtain the material and are not themselves used in the experiment; (b) Hypothesis-generating trial in which no further specified hypotheses are to be tested (pilot experiment, basic research); (c) Hypothesis-testing experiment, and as in Tuebingen (Baden Wuerttemberg), for example, also (d) educational studies.

Overall, review boards work completely independent due to federal regulations without having unified criteria (Jörgensen et al., 2021). In general, missing of the latter complicates the work of reviewers and does not guarantee an unbiased review process upon known criteria for the applicants. The situation particularly impedes planning and conducting of multicenter preclinical trials where every involved research institution must apply for approval to their concerned local authority.

We here introduced a guideline indexing biostatistical criteria that should be reported by applicants to receive ethics approval. Our biometric form sheet has been implemented by the concerned local authority of animal welfare in March 2019 on a voluntary basis and its use is obligatory since January 2021. After 2 years, we can already summarize that our form sheet helps reviewers in assessing research goals and trial design also in rather complex preclinical trials. Applicants, on the other side, have not always received the new biometric form sheet with enthusiasm. For those who do not have sound training in statistics, it is sometimes overwhelming how detailed they are now expected to explain their primary research objective, sample size planning and statistical analysis—which used to be handled much less strictly. Usually, several hours of consultation work are needed before the form is ready and can be signed by the consulting statistician. In 2020, we consulted 116 application forms. Average overall time documented in our consultancy data base was 219 min per application with an average time of 83 min in meetings with the applicant and another 136 min of work for the iterative process of preparation and revision. Most often the distinction between exploratory and confirmatory research is not clear to applicants.

We have had a lot of discussions about a suitable sample size for exploratory settings in the past. In our opinion, three to five animals are sufficient to show, for instance, technical or surgical feasibility in a pilot trial. While some of the authors of this paper have used the suggestions of Julious (2005) to justify 12 animals per group to estimate mean and *SD* for a pilot study with metric endpoint, others oppose that this suggestion is for clinical trials in humans—and variation among humans is potentially larger compared to animals. We do not feel comfortable in recommending one minimum effect size. Instead, the applicant should have a notion for what size of potential group differences the study is sensitive for or with what precision estimates can be obtained.

From February to December 2020, roughly 150 application forms for animal experiments have been reviewed by the animal welfare committee of the concerned local authority of animal welfare in Berlin. Of these applications, about 35% used the new biometric form, which was recommended but not required at that time. If the form was not used, about 90% of the applications had statistical deficiencies that demanded major revision. With the new form, deficiencies were less frequent (about 70%), though still far from satisfactory. We believe this is based on the fact that most applications were completed without the help of a biostatistician. We are optimistic that deficiencies will decrease further in 2021, as the form is now mandatory and early assistance from a biostatistician is strongly recommended. In our consultancy and AWAC work, we also found that insisting on completing a biometric check list alone may not have the desired effect on planning quality (Kilkenny et al., 2009; Vollert et al., 2020). Guidelines generally require more than journal and agency endorsement (Hair et al., 2019; Leung et al., 2018; Vollert et al., 2020; Zhao et al., 2020). We therefore advocate early involvement of a biostatistician and sufficient training and guidance for applicants, reviewers, and agencies, and recommend guidelines such as PREPARE (Smith et al., 2018) already for the planning phase of experiments (Festing & Altman, 2002; Ludolph et al., 2010). Analogous to du Sert for the consensus-adopted ARRIVE 2.0 reporting guideline (du Sert et al., 2020a, 2020b), we have written a user guide to educate and support applicants. We believe this will improve the quality of applications and facilitate the implementation of our biometric form over time. Herein, we extend the motivation and explanation for each point in the form and added a glossary for nonstatisticians. Moreover, we provide explicit examples for an exploratory and a confirmatory study (see doi https://doi.org/10.5281/zenodo.7038608 - to be announced).

4.1 | Open issues/Further work

A major aim of the presented work is harmonizing the review process in preclinical research. The present review criteria and the form sheet are an important first step. In Germany, the working group Non-Clinical Statistics of the IBS (German Region), for example, has been initiating workshops on statistical planning and ethical review of animal testing and has disseminated the form. We plan further activities and workshops with the help of others in achieving this goal. Our user guide is a living document that will be constantly updated and we encourage readers (applicants as well as biometricians and AWAC members) to add their experiences. Moreover, we advertise its usage among animal welfare agencies and officers.

4.2 | Recommendation

Based on our experience with our biometric form sheet to date, we would like to express some general recommendations for statisticians working in the field of animal experiments, both in terms of statistical consultancy work and the review process in AWACs. There is a lack of statistical support for preclinical researchers, and preclinical researchers often lack the prerequisites to conduct rigorous statistical planning and analysis of their experiments. Therefore, there is a great need for more involvement of statisticians, and more guidance. We have observed that using our biometric form sheet in many cases provides much of the required guidance and recommend its implementation at other AWACs. In addition, preclinical researchers should take advantage of biostatistical support as soon as they start planning their experiments. We also recommend our user guide to support the consultation process. Moreover, we would welcome further initiatives to increase the number of statisticians in AWACs. We would also like to stress the importance of biometrical institutes providing statistical consultancy services to applicants and agencies. Ideally, this will initiate a harmonization process involving all relevant stakeholders such as researchers, animal welfare officers, industry, and biostatisticians.

5 | CONCLUSION

In conclusion, preclinical researcher need biostatistical support when planning their experiments. From our consultancy experiences, we developed and implemented a biometric form sheet to guide applicants. This is a first step to standardize applications and streamline ethical reviews. Finally, we plan for broader dissemination in collaboration with the other animal welfare authorities in Germany.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft grant number DFG KO 4680/4-1. Ulf Toelch was supported by BMBF funding (01KC1901A).

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

AUTHOR CONTRIBUTIONS

Sophie K. Piper, Dario Zocholl, Ulf Toelch, Andrea Stroux, Robert Roehle, and Frank Konietschke designed the biometric form sheet. Sophie K. Piper, Dario Zocholl, Ulf Toelch, Robert Roehle, and Frank Konietschke prepared parts and critically revised the manuscript as well as the user guide. Johanna Hoessler and Anne Zinke acquired background information about the concerned local animal welfare authorities in Germany and carefully revised the manuscript.

DATA AVAILABILITY STATEMENT

Anonymized data about the usage of the biometric form sheet are available at https://doi.org/10.5281/zenodo.5615540.

ORCID

Sophie K. Piper https://orcid.org/0000-0002-0147-8992

Ulf Toelch https://orcid.org/0000-0002-8731-3530
Frank Konietschke https://orcid.org/0000-0002-5674-2076

REFERENCES

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195.

Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R., & Joëls, M. (2021). Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience*, *24*(4), 470–477.

Buchner, B., Hase, F., Borchers, D., & Pigeot, I. (2019). Tasks, regulations, and functioning of ethics committees. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *62*(6), 690–696.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216.

Danzinger, M., Dirnagl, U., & Toelch, U. (2022). Increasing discovery rates in preclinical research through optimised statistical decision criteria. https://www.biorxiv.org/content/10.1101/2022.01.17.476585v1

Doppelfeld, E., & Hasford, J. (2019). Medical research ethics committees in the Federal Republic of Germany: Establishment and integration into medical research. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *62*(6), 682–689.

Drude, N. I., Martinez Gamboa, L., Danziger, M., Dirnagl, U., & Toelch, U. (2021). Science Forum: Improving preclinical studies through replications. *eLife*, *10*, e62101.

du Sert, N. P., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Hurst, V., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … , Würbel, H. (2020a). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLoS Biology*, *18*(7), e3000411.

du Sert, N. P., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., Maccallum, C. J., Macleod, M., … Würbel, H. (2020b). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, *40*(9), 1769–1777.

Festing, M. F. (2018). On determining sample size in experiments involving laboratory animals. *Laboratory Animals*, *52*(4), 341–350.

Festing, M. F. (2020). The "completely randomised" and the "randomised block" are the only experimental designs suitable for widespread use in pre-clinical research. *Scientific Reports*, *10*(1), 1–5.

Festing, M. F., & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal*, *43*(4), 244–258.

Festing, M. F., Baumans, V., Combes, R. D., Haider, M., Hendriksen, C. F., Howard, B. R., Lovell, D. P., Moore, G. J., Overend, P., & Wilson, M. S. (1998). Reducing the use of laboratory animals in biomedical research: problems and possible solutions: The report and recommendations of ECVAM Workshop 29. *Alternatives to Laboratory Animals*, *26*(3), 283–301.

Guhad, F. (2005). Introduction to the 3Rs (refinement, reduction and replacement). *Journal of the American Association for Laboratory Animal Science*, *44*(2), 58–59.

Hair, K., Macleod, M. R., & Sena, E. S. (2019). A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Research Integrity and Peer Review*, *4*(1), 1–17.

Held, L., Micheloud, C., & Pawel, S. (2022). The assessment of replication success based on relative effect size. *Annals of Applied Statistics*, *16*(2), 706–720.

Hirst, J. A., Howick, J., Aronson, J. K., Roberts, N., Perera, R., Koshiaris, C., & Heneghan, C. (2014). The need for randomization in animal trials: An overview of systematic reviews. *PLoS One*, *9*(6), e98856.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jörgensen, S., Lindsjö, J., Weber, E. M., & Röcklinsberg, H. (2021). Reviewing the review: A pilot study of the ethical review process of animal research in sweden. *Animals*, *11*(3), 708.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*, *4*(4), 287–291.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F., Cuthill, I. C., Fry, D., Hutton, J., & Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*, *4*(11), e7824.

Kimmelman, J., Mogil, J. S., & Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, *12*(5), e1001863.

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is "n" in cell culture and animal experiments? *PLoS Biology*, *16*(4), e2005282.

Leung, V., Rousseau-Blass, F., Beauchamp, G., & Pang, D. S. (2018). ARRIVE has not ARRIVed: Support for the ARRIVE (Animal research: Reporting of in vivo experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One*, *13*(5), e0197882.

Ludolph, A. C., Bendotti, C., Blaugrund, E., Chio, A., Greensmith, L., Loeffler, J.-P., Mead, R., Niessen, H. G., Petri, S., Pradat, P.-F., Robberecht, W., Ruegg, M., Schwalenstöcker, B., Stiller, D., van den Berg, L., Vieira, F., & von Horsten, S. (2010). Guidelines for preclinical animal research in ALS/MND: A consensus meeting. *Amyotrophic Lateral Sclerosis*, *11*(1-2), 38–45. PMID: 20184514.

Mogil, J. S., & Macleod, M. R. (2017). No publication without confirmation. *Nature News*, *542*(7642), 409.

Neumann, K., Grittner, U., Piper, S. K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J. P., Kimmelman, J., & Dirnagl, U. (2017). Increasing efficiency of preclinical research by group sequential designs. *PLoS Biology*, *15*(3), e2001307.

OECD. (2001). *Guideline for testing of chemicals: Acute oral toxicity-acute toxic class method*. Guideline 423.

Piper, S. K., Grittner, U., Rex, A., Riedel, N., Fischer, F., Nadon, R., Siegerink, B., & Dirnagl, U. (2019). Exact replication: Foundation of science or game of chance? *PLoS Biology*, *17*(4), e3000188.

Rauch, G., Hafermann, L., Mansmann, U., & Pigeot, I. (2020). Comprehensive survey among statistical members of medical ethics committees in Germany on their personal impression of completeness and correctness of biostatistical aspects of submitted study protocols. *BMJ Open*, *10*(2), e032864.

Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A., & Brattelid, T. (2018). Prepare: Guidelines for planning animal research and testing. *Laboratory Animals*, *52*(2), 135–141.

Strech, D., & Dirnagl, U. (2019). 3Rs missing: Animal research without scientific value is unethical. *BMJ Open Science*, *3*(1), e000048.

Vollert, J., Schenker, E., Macleod, M., Bespalov, A., Wuerbel, H., Michel, M., Dirnagl, U., Potschka, H., Waldron, A.-M., Wever, K., Steckler, T., van de Casteele, T., Altevogt, B., Sil, A., Rice, A. S. C., & The EQIPD WP3 study group members. (2020). Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Science*, *4*(1), e100046.

Zhao, B., Jiang, Y., Zhang, T., Shang, Z., Zhang, W., Hu, K., Chen, F., Mei, F., Gao, Q., Zhao, L., Kwong, J. S. W., & Ma, B. (2020). Quality of interventional animal experiments in Chinese journals: Compliance with ARRIVE guidelines. *BMC Veterinary Research*, *16*(1), 460.

How to cite this article: Piper, S. K., Zocholl, D., Toelch, U., Roehle, R., Stroux, A., Hoessler, J., Zinke, A., & Konietschke, F. (2022). Statistical review of animal trials—A guideline. *Biometrical Journal*, 1–12. https://doi.org/10.1002/bimj.202200061