SPONSORED BY THE

Federal Ministry of Education and Research

BIH QUEST Center for Responsible Research

Guidance for planning and conducting confirmatory preclinical studies and systematic reviews

This document provides guidance on multicenter confirmatory preclinical studies and systematic reviews. It is based on a review of the literature and interaction with the projects currently funded under the call "Confirmatory preclinical studies – quality in health research" by the German Federal Ministry of Education and Research (BMBF).

The information herein was compiled by the DECIDE project conducted at the BIH QUEST Center for Responsible Research. DECIDE investigates confirmatory studies as well as systematic reviews and develops best practices together with the above mentioned BMBF-funded groups and other stakeholders and experts.

This document represents the current research status and should be seen as a guide that covers main issues but may require adoption to specific cases.

Minimum requirements to start a confirmatory study

Ideally, preliminary evidence should fulfill minimum criteria that warrant the step to a multicenter confirmatory study. Unless field-specific restrictions apply, those minimum criteria should (at least in part) be fulfilled before engaging in a confirmatory study.

- Blinding of previous experiments was at least performed for assessment of outcome, particularly if this outcome is subjective (not assessed by an algorithm or machine)
- Experimental units were randomized (independently and randomly allocated to experimental conditions)
- (Expected) Animal attrition and outlier removal of previous experiments should be clearly reported to inform inclusion/exclusion criteria for the confirmatory study
- The primary outcome of previous experiments needs to be clearly specified
- Basic quality management procedures in place like SOPs and published, complete protocols for complicated, non-standard experimental procedures; quality check for important medicinal products like antibodies, cell lines, etc. were conducted
- In-house replication of main finding to substantiate the highly uncertain initial findings
- Model is disease relevant and limitations are acknowledged
- Initial experiments/in-house replication were conducted on a sufficient number of experimental units to estimate effect size and uncertainty of primary outcome sufficiently to conduct a sample size calculation for the confirmatory study

Experimental Design of confirmatory studies

The experimental design of the confirmatory study should reflect the initial design of the exploratory study. Crucial factors like primary outcome, essential methods, and model should stay constant

between exploratory and confirmatory phase. Deviations need to be spelled out, motivated, and strengthen validity. To strengthen validity, limited extensions are possible. That is, a confirmation is not necessarily a direct replication of the initial experiment. It is rather a test of the underlying knowledge claim and should enable decisions for future steps and translation into clinical contexts. For deviations regarding the number of experimental units see statistical analysis and planning below.

General

- A flow chart[1] including randomization and blinding scheme is helpful for reviewers.

Blinding

- Blinding procedures should be stated for all stages: (i) allocation of animals (to minimize selection bias), (ii) the experiment (experimenters should be blinded to the intervention), (iii) the outcome assessment (e.g., Imaging or behavioral testing), and (iv) the analyses
- If blinding is not possible on all levels a rational should be given (e.g., experimenter cannot be blinded due to obvious differences in phenotypes of the different groups)

Randomization[6]

- Method of randomization needs to be stated including method/tool used for randomization sequence generation
- If randomization options are limited for the experimental group a rationale needs to be provided e.g., due to social transfer of pain in rodents
- Outcome relevant physiological parameters (like bodyweight or sex) should be included in the randomization scheme (blocking)
- If a subsample of animals is taken out of the experiment for specific analysis earlier, the selection of these animals should also follow a randomization scheme. (Consider immortality bias, i.e., if a substantial proportion of animals has reached a humane endpoint before a subsample analysis is performed)

Outcome measures

- All outcome measures need to be clearly defined a priori (e.g., tumor growth or behavioral change) and preregistered
- The primary outcome measure (on which sample size calculation was performed) needs to be stated and preregistered
- Secondary outcome measures can serve as converging/supporting evidence

Control groups

- Depending on the experimental objective control groups should include positive and negative controls where possible.
- If available a comparator from standard clinical care should be included and a new treatment/intervention tested for non-inferiority

Inclusion/ exclusion criteria

- Need to be defined a priori (= pre-registered)
- Should refer to animal welfare considerations (humane endpoints for the individual animal), animal model (e.g., phenotype or threshold for disease characteristic) **and** data (e.g., a priori defined threshold when data will be excluded or due to technical issues)
- Outliers should be defined across all groups and not only for a specific treatment group[13]

(Animal) Model

- Limitations of the model should be stated
    - What parts of e.g., a disease does the model reflect?
    - Are (clinical) biomarkers and/or diagnostics like imaging considered?

Quality control

- Measures to assure the quality of medicinal product(s), cells, animal model, phenotypes should be in place
- Strategy for the harmonization of protocols as well as trainings between different labs needs to be described

Systematic Heterogenisation

External validity is strengthened by introducing systematic heterogenisation. Even though experiments need to be standardized between centers, systematic heterogenisation potentially reveals important interactions of the genotype and the environment that also affect treatment outcomes[14] It is thus important that a) there is awareness of environmental differences between centers (e.g. standard food and cage sizes) b) introduce a *limited* number of systematic heterogenisation factors between centers. This can be done on the environmental level (varying age, light regime, timing of experiments) or on the genetic level (different strains, homo vs heterozygous animals). Here a balance needs to be found between strengthening external validity through systematic heterogenisation and comparability of centers. If a factor has been previously identified to influence the primary outcome, it should be included in the heterogenisation scheme. Alternatively, complementing experiments within one center probe external validity by varying crucial factors mentioned earlier. Again, a balance between feasibility and knowledge gain needs to be struck here. Including both sexes in each center should be the norm (except for sex specific diseases).

Statistical analysis and Study Planning

Every project should consult a (bio-)statistician in the planning phase. Ideally, the statistician is available for potential consultations throughout the research phase. Analyses should adhere to the pre-specified plan and particularly demanding analyses should be closely monitored or conducted by the consulting statistician.

- Primary outcome should be defined a-priori before starting the confirmatory trial and should be the same (or similar) as in the exploratory study.
    - If devices are used from which several outcome measures can be extracted (e.g., gait analysis), pre-registering the specific outcome measure as a defined primary outcome is crucial.
- Experimental unit must clearly be stated (the unit that is randomly and independently allocated to experimental conditions)[2];
- sample size calculation (see below) must be based on experimental unit
- To be able to evaluate outcome of exploratory study, give mean values and corresponding standard deviations ($M \pm SD$) of all groups (intervention and control(s)) as well as experimental units per group (N) that these values are based on. Researchers should be careful to not mistake standard error of the mean (SEM) with standard deviations especially when communicating results from exploratory studies to study statisticians for the confirmatory trial.
- For deriving an effect size such as Cohen's d, a *pooled* standard deviation must be calculated across experimental groups.
- For outcomes which are presented as percentages to facilitate cross-study comparison and standardization consider providing raw values as well. Percentage value show floor and ceiling effects with regards to their variances when values approach 0% or 100%.

Sample size estimation for confirmation

We recommend not to use exploratory effect size (ES) estimate for sample size estimation for subsequent experiments (within-lab replication or multicenter study). The exploratory ES is often based on a small sample and likely overestimates the true underlying effect.[7,8]

- Instead, we recommend defining a smallest effect size of interest (SESOI) or using a shrinkage factor by which the exploratory ES is reduced. This results in larger sample sizes and a higher power to detect effects that are smaller than the exploratory ES
- SESOI and shrinkage factor need to be substantiated.
    - When employing a SESOI, it should be stated whether the SESOI reflects a biologically relevant effect (to argue for a specific mechanism of action) or is of clinical relevance to e.g. predict efficacy of an intervention. This might vary depending on the research question. In the latter case, the choice of a SESOI should be informed by a clinician.
    - The use of a particular SESOI or shrinkage factor may be based on different considerations (e.g., previous evidence, practical constraints/feasibility). Those considerations should be transparently reported
- Account for systematic heterogenization (see above). Sample size estimation needs to account for larger biological variability in the results in the confirmatory compared to the exploratory experiments. This can also be reflected in a shrinkage factor (see above)
- In more sophisticated study design such as using blocked or stratified randomization, for example, the blocking or stratification factor needs to be incorporated as a covariate in the statistical model.

- For the reasons stated above, using smaller or similar sample size in the confirmatory trial as compared to the exploratory trial is not recommended and should be justified explicitly[9].

Control groups

- If two or more control groups are pooled for comparison against treatment group:
  - Establish that control groups are similar before pooling.
  - Pre-specify equivalence margin.
  - If similarity cannot be established: Which comparison (control – intervention) is most relevant to answer the research question?

Patient derived in-vitro models

- Obtain and store meta-data of donors. Such meta-data should contain basic patient data like age and gender and additionally disease relevant information. This can be incorporated into extended statistical model and analysis (may be a source of variance).
- Make sure structure of statistical models reflect repeated measures and biological units/observational units properly (e.g. through random effects structure in mixed models)
- Outlier criteria should be pre-defined.
- What are inclusion criteria for donors – how does this sample of donors relate to the superpopulation (e.g. all patients) of donors?

Subgroup analysis, sensitivity analysis

- Consider clustering of data (e.g. mice of same litter) in subgroup analysis to detect deviations (e.g. prenatal disturbances).
- Sensitivity analysis could include different imputation methods e.g. multiple imputations, which can then be compared to the standard complete case analysis that might suffer from selection bias.
- Sensitivity analysis can also include incorporating baseline-values for outcome measures to assess robustness.
- Presenting sex-disaggregated data is recommended for meta-analytic purposes[10].
  - However, simple comparisons between male and female animals are potentially misleading if no formal test of interaction was performed and particular care must be taken to describe results according to the hypotheses that were addressed in a statistical test[11].
  - If an interaction analysis is performed, results should be described as exploratory if the sample size calculations are based on main effects only. Be aware that power is usually extremely low to detect significant interaction effects[12].

Multicenter Studies

Specific Aspects for multicenter studies

- Based on the primary outcome of your study determine core results that need to be replicated across centers
    - Core results are outcomes that reflect your main effect and ideally are gold standard measures for your paradigm (some examples: Stroke: lesion volume; Cancer: cancer volume across time and survival; Cardio: left ventricular ejection fraction measured via gold standard method)
- Balance cost and effort against the evidence generated when deciding which outcomes to replicate across several labs
- The number of external labs should be limited (ideally not more than 2)[3]
- External labs should be experienced in the models/methods you use
- (External) labs can perform complementary analyses that need not be replicated across all labs but provide insights into patho-/disease mechanisms
- Plan time and resources for protocols, work instructions and standard operating procedures generation/standardization between labs.
    - Protocols need to be closely aligned and differences between labs documented. Protocols need high level of detail and reviewers in all centers
- Compare controls/ baseline of your primary outcomes to ensure similarity between centers
- If animal experiments are planned make sure animal permit applications are aligned across centers or plan enough time for the application of all centers
- Protocols and data storage and analysis should be administered centrally
- Data formats need to be standardized across centers
- Pre-define center-allocation scheme in randomization process. Prioritize covariates (e.g. bodyweight is more important than sex), which must be balanced across centers in case any imbalances arise.

Systematic Reviews

Valuable hints:

- It is recommended that researchers will attend / have attended training in (preclinical) systematic review and / or have a systematic review methodologist listed on the review team
- Search for already existing systematic reviews in the field of interest
- Demonstrate need for the review
- Complete SYRCLE protocol template
- Describe strategies for data management and sharing and dissemination of results
- Specify whether an information specialist (e.g., librarian) will be / has been consulted in the development of the search strategy
- Specify at least 2 databases that will be searched, unless expressly justified
- Specify that screening, data extraction and risk of bias for each study will be carried out by two reviewers, independently (or equivalent measure to ensure quality)

- Specify how the findings from the assessment of internal validity will be used to inform the conclusions of the review
- Discuss how external /translational validity will be assessed, i.e., the features of the included studies that are specifically relevant to the assessment of the external / translational validity of the animal models in the disease of interest and how these relate to the translational potential of the findings (to humans)
- Specify how the findings from the assessment of external / translational validity will be used to inform the conclusions of the review
- Describe how the outcomes selected for assessment are relevant to the outcomes of the human disease being modelled
- If meta-analysis is planned, specify whether a statistician / methodologist will be / has been consulted
- Specify what reporting guidelines will be used for any publications (PRISMA preclinical is under development, but the main PRISMA guidelines are relevant)
- Specify which systematic review data management system will be used, i.e. not Excel
- Specify how, where and when data and code will be made freely available (expectation that these will be shared without restriction unless expressly justified). Should adhere to FAIR data principles.
- Financial plan should include project management costs
- Work and financial plans should also include adequate time for training all staff who will be working on the review and piloting searches, screening, and data extraction forms to ensure quality standards
- The protocol should be registered (PROSPERO for reviews related to human health) and / or published (e. g. on OSF or equivalent with DOI and timestamp) prior to the start of data extraction.
    - This can be a work package and is not mandatory at the stage of proposal writing

Two example protocols for guidance:

https://openscience.bmj.com/content/5/1/e100135

https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-3-48

References

1. du Sert, N. P. *et al.* The Experimental Design Assistant. *Nature Methods* **14**, 1024–1025 (2017).

2. Lazic, S. E., Clarke-Williams, C. J. & Munafò, M. R. What exactly is 'N' in cell culture and animal experiments? *PLOS Biology* **16**, e2005282 (2018).

3. Voelkl, B., Vogt, L., Sena, E. S. & Würbel, H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLOS Biology* **16**, e2003693 (2018).

4. Voelkl, B. *et al.* Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience* **21**, 384–393 (2020).

5. Clayton, J. A. Studying both sexes: a guiding principle for biomedicine. *FASEB J* **30**, 519–524 (2016).

6. Festing, M. F. W. The “completely randomised” and the “randomised block” are the only experimental designs suitable for widespread use in pre-clinical research. *Scientific Reports* **10**, 17577 (2020).

7. Colquhoun, D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* **1**, 140216–140216 (2014).

8. Colquhoun, D. The False Positive Risk: A Proposal Concerning What to Do About p-Values. *null* **73**, 192–201 (2019).

9. Piper, S. K. *et al.* Exact replication: Foundation of science or game of chance? *PLoS Biol* **17**, e3000188 (2019).

10. Heidari, S., Babor, T. F., De Castro, P., Tort, S. & Curno, M. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev* **1**, 2 (2016).

11. Garcia-Sifuentes, Y. & Maney, D. L. Reporting and misreporting of sex differences in the biological sciences. *eLife* **10**, e70817 (2021).

12. VanderWeele, T. J. & Knol, M. J. A tutorial on interaction. *Epidemiologic Methods* **3**, 33–72 (2014).

13. André, Q. (2021). Outlier exclusion procedures must be blind to the researcher’s hypothesis. *Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001069

14. Voelkl, B., Altman, N.S., Forsman, A. *et al.* Reproducibility of animal research in light of biological variation. *Nat Rev Neurosci* **21,** 384–393 (2020). https://doi.org/10.1038/s41583-020-0313-3