Conceptual approaches to sample size calculation and their effect on assessing replication success.

Relevance

"No publication without confirmation" is a demand gaining support by the preclinical research community to foster reproducibility of scientific findings and increase the certainty of selecting the most promising hypotheses for further (clinical) testing (Mogil and Macleod 2017; Kimmelman, Mogil, and Dirnagl 2014). A crucial factor in designing confirmatory studies is the choice of sample sizes, which especially in research involving animals, has to follow a throughout harm-benefit and feasibility assessment. Reported practices from large-scale preclinical replication projects reflect the lack of a clear consensus as to whether, and how to incorporate the often highly uncertain estimates from exploratory research into the sample size calculation for confirmatory studies (Errington et al. 2021; Amaral et al. 2019; Drude 2022). Adding to the complexity of conducting and assessing replication studies, many definitions of replication success are dependent on the precision of effect estimates, and in turn highly dependent on sample size choices (Errington et al. 2021). We assume different conceptual starting points for sample size calculation influence the probability to declare replication success. With a simulation study, we aim to assess which approach best balances correct decision rates to declare replication success or failure and additional animals used.

Background

Animal experiments represent an essential intermediate step between basic molecular biomedicine and clinical research. The critical evaluation of the results from animal experiments aims to minimize harm and maximize benefit for patients. Failures in clinical translation however have raised doubts about the reliability of animal research (Ioannidis et al. 2014; Begley and Ellis 2012). Current animal experiments with a low number of animals in mainly single-center exploratory studies are characterized by low statistical power and low positive predictive values (Bonapersona et al. 2021; Colquhoun 2014). In consequence, evidence generated in preclinical animal studies is not decision-enabling for the start of clinical trials. As an improvement strategy, scientists propose to establish a preclinical trial structure consisting of separate phases of hypothesis-generating exploratory research, within-lab and, if warranted by the evidence, multi-center replication with similar requirements on preregistration, robustness of design, methods and statistical analysis as clinical trials. (Drude

et al. 2021; Mogil and Macleod 2017; Strech and Dirnagl 2019). In confirmatory research, an emphasis lies on the reliability and conclusiveness of the findings, thus animal numbers need to be increased to reduce the chance of spurious findings. So far, it is unclear how animal numbers in confirmatory studies should be estimated as reported effect sizes from exploratory studies will be highly uncertain and potentially inflated due to low power, selection and publication bias (Colquhoun 2014; Ioannidis 2008; Button et al. 2013; Holman et al. 2016). Alternative approaches have been developed within the field of psychology (Perugini, Gallucci, and Costantini 2014; Simonsohn 2015; Lakens 2021) but their suitability for preclinical sciences has yet not been assessed. The central goal of a confirmatory preclinical trial is to judge on the validity and relevance of a research hypothesis. With respect to interventional trials aiming at clinical translation, if the judgement is correct, patients either are protected from unnecessary harm or can be more certain about a clinical benefit of a new intervention. Therefore, both the correct declaration of replication failure and replication success should be seen as beneficial to patients and in accordance with the 3R principles. In consequence, the decision criteria should function as an accurate diagnostic tool. The most commonly employed criteria is (re-)achieving statistical significance and equal directionality of the effect estimate in the replication. Alternatively, researchers could assess whether an effect size larger than the smallest effect size of interest was observed in addition to statistical significance (Danziger, Collazo, Dirnagl, and Toelch 2022). Criteria driven by cumulative assessment of both the original and replication study include fixed-factor meta-analysis or a reverse Bayesian approach with the sceptical p-value as a measure of replication success (Held, Micheloud, and Pawel 2021; Errington et al. 2021).

Research Questions

The first part of my project consists of a descriptive summary of approaches to sample size calculation and replication success measures in three large-scale preclinical replication efforts, namely, the Reproducibility Project: Cancer Biology, the Brazilian Reproducibility Initiative and the Confirmatory Preclinical Studies project. We are interested in assessing:

- How are sample size calculations for preclinical replication studies performed?
- (How) is information from the exploratory effect estimate and its uncertainty incorporated?
- What assumptions do researchers make regarding the stability of the effect size?
- How do researchers justify these choices?

Fig.1 provides some preliminary visualization of relative sample sizes for the three projects.

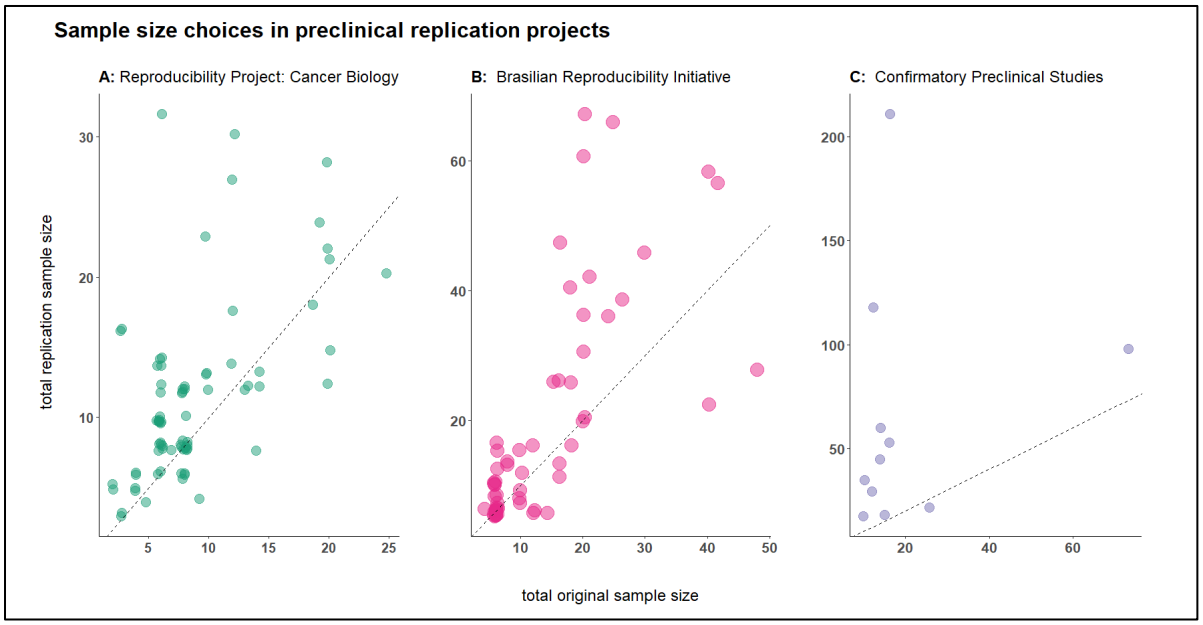

Fig 1: A: Reproducibility Project: Cancer Biology, data from 77 replicated *in vivo*/*in vitro* experiments, 80 % power; **B:** Brazilian Reproducibility Initiative, data from 58 replicated *in vitro*/*in vivo* experiments, 95 % power; **C:** Confirmatory Preclinical Studies.

The second part of the project consists of an *in-silico* assessment of the relationship between sample size calculation approach and correct decision rates under different replication success criteria. We explore all possible combinations outlined in **Table 1**. We aim to explore the following questions:

- For each sample size calculation approach: How many times does the result of the sample size calculation fall below two animals per group, reflecting a missing need to start a confirmatory study. How many times is this decision correct (effect size is relevant)?
- For each sample size calculation approach: How many times does the result of the sample size calculation indicate more than 100 animals per group, pointing towards feasibility problems?
- How often is a correct decision regarding follow-up on the hypothesis of interest taken? This is the case if:
    - Replication success is declared and true underlying effect size is relevant. (Continued line of research for a valid claim).
    - Replication failure is declared and true underlying effect is irrelevant. (Decision switch towards abandoning the research hypothesis).
- How often is an incorrect decision regarding follow-up on the hypothesis of interest taken? This is the case if:
    - Replication failure is declared and true underlying effect size is relevant.
    - Replication success is declared and true underlying effect is irrelevant.

- How does the rate of correct decisions relate to the average absolute number of additional animals used in the replication study compared to initial study?

Methods

Dataset: Our data set contains data of original experiments from three different preclinical research projects: Reproducibility Project: Cancer Biology (RPCB), Brazilian Reproducibility Initiative (BRI) and Confirmatory Preclinical Studies (CPS). Experiment-level data is freely available on OSF for RPCB, was shared through a co-author for the BRI and was obtained from project proposals for CPS.

Exclusion criteria for original experiments:

- no Cohen's d or standardized mean difference presented as effect size estimate
- original sample size not given
- original sample size per group > 100
- no statistical significance achieved in the original study

Extracted data:

- original total sample size
- original 95 % confidence interval

Sample size calculation & replication success criteria: We will than re-calculate sample sizes for a two-sample design with one-sided testing using the following approaches:

a. Sample size calculation based on the exploratory effect size, power = 80 %, significance level = 0.05. The underlying assumption is that an effect of the same size can be shown in the confirmatory study. However, given the highly inflated effect sizes in preclinical research, this approach is likely to lead to underpowered replication studies.
b. Sample size calculation based on i) the lower 80 % confidence bound and power = 80 %, significance level = 0.05. This method was introduced by Perugini et al. 2014 within the field of psychology and coined "safeguard power analysis"(Perugini, Gallucci, and Costantini 2014).
c. Sample size calculation based on a reverse Bayesian approach proposed by Held et al. incorporating a skeptical prior in addition to the exploratory effect estimate and its uncertainty (Held 2020). We will incorporate an anticipated effect size shrinkage of 25%, power = 80 %, significance level = 0.05.
d. Sample size calculation based on the smallest effect size of interest (SESOI) set at Cohen's $d_1 = 0.5$ (for $d_{orig} = 0.5 - 1.5$); Cohen's $d_2 = 1.5$ (for $d_{orig} = 1.5 – 2.5$); $d_3 = 2.5$ (for $d_{orig} > 2.5$), power = 50 %, significance level = 0.05. In this approach, researchers have to

define a lower threshold for a relevant and realistic effect size *a priori* based on their field-specific knowledge.

The *in-silico* approach enables us to set an underlying true effect size, when simulating outcome values for the replication experiments. We will investigate three scenarios:

Scenario 1: The ratio between replication and original effect size (or relative effect size) equals 0.5, which represents a 50 % shrinkage of the effect estimate towards the null. This scenario reflects a magnitude error in the original study. It is empirically plausible but rather optimistic (Errington et al. 2021). Here, we assume that the shrunken effect estimate would still be of scientific relevance and therefore worthwhile detecting.

Scenario 2: We simulate a null-effect of treatment.

Scenario 3: We simulate a 125 % shrinkage of the effect estimate, which reflects a sign error of the original study and a change in directionality of the treatment effect. This was empirically observed in some completed replication studies of the RPCB and is especially important from a patient-harm reduction perspective.

The replication success criteria are listed in Table 1 and follow common data-driven approaches described (A, C) or alternative methods (B, D) incorporating external information such as the smallest effect size of interest or a skeptical prior (Errington et al. 2021; Danziger, Dirnagl, and Toelch 2022; Held, Micheloud, and Pawel 2021). For the skeptical p-value approach, we will use the R-package *ReplicationSuccess* based on Held et al. (Held 2020). The simulation and analysis will be programmed in R.

Sample size calculation approach	**Replication success criterion**
A: Replication study powered for the effect size obtained in the original study at 80% and 95% respectively.	**A:** statistical significance (alpha = 5%) and same direction of effect
B: Replication study powered at 50% for the smallest effect size of interest (SESOI). $SESOI_1$ = 0.5 (for orig. effect size: 0.5 -1.5); $SESOI_2$ = 1.5 (for orig. effect size: 1.5 – 2.5); $SESOI_3$ = 2.5 (for orig. effect size: > 2.5)	**B:** statistical significance (alpha = 5%) and ≥ SESOI
C: Replication study powered at 80% for the lower 80% confidence bound obtained from the original study.	**C:** meta-analysis of original and replication study (fixed-effect), significant p-value (alpha = 5%)
D: Replication study powered for reverse Bayesian approach (skeptical p-value), with an effect size shrinkage estimate of 25%	**D:** "significant" skeptical *p*-value (golden level)

Table 1: Sample size calculation approaches and replication success criteria.

Project team, project plan, dissemination strategy

I am working on this project together with Meggie Danziger and Ulf Tölch. We are currently preregistering the project on OSF as an open-ended preregistration. We have so far applied the exclusion criteria to the three datasets and outlined the simulation code, which will be written in R. We aim to share data and code on OSF and Github to allow for the reproducibility of study results. However, for BRI and CPS we have to discuss with the respective researcher teams first, on how data sharing can be accomplished. We aim to publish our manuscript with LabAnimal or Royal Society for Open Science.

Time table

	10-2022	11-2022	12-2022	01-2023	02-2023	03-2023	04-2023
Preregistration on OSF							
Descriptive summary of RPCB, BRI & CPS							
Programming simulation							
Analysis and visualization of results							
Sensitivity analysis							
Manuscript writing							

References

Amaral, Olavo B, Kleber Neves, Ana P Wasilewska-Sampaio, and Clarissa FD Carneiro. 2019. “The Brazilian Reproducibility Initiative.” Edited by Peter Rodgers, Timothy M Errington, and Richard Klein. *ELife* 8 (February): e41602. https://doi.org/10.7554/eLife.41602.

Begley, C Glenn, and Lee M Ellis. 2012. “Raise Standards for Preclinical Cancer Research.” *Nature* 483 (7391): 531–33.

Bonapersona, V., H. Hoijtink, R. A. Sarabdjitsingh, and M. Joëls. 2021. “Increasing the Statistical Power of Animal Experiments with Historical Control Data.” *Nature Neuroscience* 24 (4): 470–77. https://doi.org/10.1038/s41593-020-00792-3.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Colquhoun, David. 2014. “An Investigation of the False Discovery Rate and the Misinterpretation of *p* -Values.” *Royal Society Open Science* 1 (3): 140216. https://doi.org/10.1098/rsos.140216.

Danziger, Meggie, Anja Collazo, Ulrich Dirnagl, and Ulf Toelch. 2022. “Balancing sensitivity and specificity in preclinical research” bioRxiv 2022.01.17.476585; doi: https://doi.org/10.1101/2022.01.17.476585. https://doi.org/10.1101/2022.01.17.476585.

Drude, Natascha Ingrid. 2022. “Planning Preclinical Confirmatory Multicenter Trials to Strengthen Translation from Basic to Clinical Research – a Multi-Stakeholder Workshop Report.” July 29, 2022. https://doi.org/10.21203/rs.3.rs-1855244/v1.

Drude, Natascha Ingrid, Lorena Martinez Gamboa, Meggie Danziger, Ulrich Dirnagl, and Ulf Toelch. 2021. “Improving Preclinical Studies through Replications.” Edited by Peter Rodgers, Catherine Winchester, and Hanno Wuerbel. *ELife* 10 (January): e62101. https://doi.org/10.7554/eLife.62101.

Errington, Timothy M, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021. “Investigating the Replicability of Preclinical Cancer Biology.” Edited by Renata Pasqualini and Eduardo Franco. *ELife* 10 (December): e71601. https://doi.org/10.7554/eLife.71601.

Held, Leonhard. 2020. “A New Standard for the Analysis and Design of Replication Studies.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183 (2): 431–48. https://doi.org/10.1111/rssa.12493.

Held, Leonhard, Charlotte Micheloud, and Samuel Pawel. 2021. “The Assessment of Replication Success Based on Relative Effect Size.” *ArXiv:2009.07782 [Stat]*, April. http://arxiv.org/abs/2009.07782.

Holman, Constance, Sophie K. Piper, Ulrike Grittner, Andreas Antonios Diamantaras, Jonathan Kimmelman, Bob Siegerink, and Ulrich Dirnagl. 2016. “Where Have All the Rodents Gone? The Effects of Attrition in Experimental Research on Cancer and Stroke.” *PLOS Biology* 14 (1): e1002331. https://doi.org/10.1371/journal.pbio.1002331.

Ioannidis, John P. A. 2008. “Why Most Discovered True Associations Are Inflated.” *Epidemiology* 19 (5): 640–48. https://doi.org/10.1097/EDE.0b013e31818131e7.

Ioannidis, John P A, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, and Robert Tibshirani. 2014. “Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis.” *The Lancet* 383 (9912): 166–75. https://doi.org/10.1016/S0140-6736(13)62227-8.

Kimmelman, Jonathan, Jeffrey S. Mogil, and Ulrich Dirnagl. 2014. “Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation.” Edited

by David R. Jones. *PLoS Biology* 12 (5): e1001863. https://doi.org/10.1371/journal.pbio.1001863.

Lakens, Daniel. 2021. “Sample Size Justification.” PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf.

Mogil, Jeffrey S., and Malcolm R. Macleod. 2017. “No Publication without Confirmation.” *Nature* 542 (7642): 409–11. https://doi.org/10.1038/542409a.

Perugini, Marco, Marcello Gallucci, and Giulio Costantini. 2014. “Safeguard Power as a Protection Against Imprecise Power Estimates.” *Perspectives on Psychological Science* 9 (3): 319–32. https://doi.org/10.1177/1745691614528519.

Strech, Daniel, and Ulrich Dirnagl. 2019. “3Rs Missing: Animal Research without Scientific Value Is Unethical.” *BMJ Open Science* 3 (1): bmjos-2018-000048. https://doi.org/10.1136/bmjos-2018-000048.