From replicability to generalizability

Arroyo Araujo, Maria

DOI:
10.33612/diss.325014460

General introduction

María Arroyo-Araujo

Groningen Institute of Evolutionary Life Sciences (GELIFES), University of Groningen (Groningen, the Netherlands)

REPRODUCIBILITY AND REPLICABILITY OF RESULTS

The development of new therapeutic targets relies heavily on the results of preclinical research. For this reason, it is of the utmost importance that preclinical findings are reproducible and replicable. In brief, reproducible results say about the feasibility to obtain consistent results using the same input data, methods, code, analysis (*i.e.,* computational reproducibility). Thus, reproducibility is closely linked to transparency and does not have a say about the correctness of the computation (*e.g.,* if there is an error in the experimental design and the study is replicated, the same erroneous result will be replicated). On the other hand, replicability means obtaining consistent results after collecting new data by using comparable methodologies (1,2). The relevance of replicable findings in preclinical studies lays on maximizing the potential of these findings towards the development of therapeutic strategies for the target (clinical) population. Specifically, replicability of results from scientific research has served as a way to operationalize truth as it suggests that the phenomenon under examination can be detached from the specific circumstances at which it was originally assessed (2,3). Unfortunately, over the past decades there have been numerous accounts of poor scientific replicability both across and within labs, certainly preclinical research and rodent phenotyping studies are not the exception (4,5). For instance, the landmark multi-center study by Crabbe et al., (1999) (6) showed that the variability across laboratories was larger than within laboratories following protocol standardization and harmonization across sites. The consequences of this 'replicability crisis' impact both the scientific community as well as the general public. For example, researchers may be unintentionally misled by inconclusive and/or inaccurate findings steering research towards slow, non-efficient, treatment development for clinical trials. Other costs as a result of poor replicability of results include the waste of financial and other resources, ethical concerns that come with the use of animals for inconclusive/uninformative research, as well as the delay in development of new therapeutic treatments. Thus, there is an urgent need to identify the underlying causes of irreplicable scientific findings, thereby reduce variability across and within laboratories, and ultimately improve the scientific value of preclinical studies for the development of novel therapeutic strategies that could benefit patients and their families.

Sources and countermeasures

The possible sources for irreplicable results are numerous and differ on their potential to help gaining knowledge. According to this classification, irreplicable results that are helpful to gain knowledge are consequence of studying complex systems with imperfect knowledge and tools and they represent a normal part of the scientific process rather than mistakes. In contrast, irreplicable results that are not helpful for gaining knowledge

come from shortcomings in the design, performance and reporting of studies; these can be honest mistakes or deliberate misconduct (1).

When it comes to replicability of results, it is necessary to consider the study's internal validity: a study is said to have internal/causal validity when one can assure that the outcome obtained was caused by the experimental manipulation and not by any other source of variability (7). To ensure this causal relationship, experimental designs should account for unknown sources of variability (*i.e.,* noise) that could influence the effect of the experimental manipulation. This can be achieved by research practices that minimize the risk of bias (*e.g.,* blinding of groups/treatments, randomization of subjects, etc.) and thus, prevent possible confounding. Hence, results from studies with internal validity are more likely to be accurate and replicable.

Other sources of irreplicable results and/or low interval validity that are classified as unhelpful to gain knowledge are: publication bias, underpowered studies, p-hacking, and p-HARking (Hypothesis After Results). Such suboptimal research practices are indeed some of the best known sources of irreplicable results (4) and have a tight link to research integrity. According to the national survey on research integrity (NSRI) performed to Dutch researchers across fields and academic ranks, 50% of responders have engaged in at least 1 of 11 questionable research practices (QRPs) surveyed (8). In addition to QRPs, falsification, fabrication and plagiarism (FFP) also affect replicability; these sources are associated to researcher integrity. Such researcher misconduct is more rare in frequency than QRPs and so its impact on the literature is smaller (9); nevertheless, 4% of responders engaged in data fabrications at least one time over the previous three years. The results of the NSRI indicate higher prevalence of QRPs than previously reported elsewhere (9,10), indicating that there is a need for a change in the current research culture. As a matter of fact, the authors of the NSRI also explored possible explanatory factors to incur in QRPs and misconduct. Publication pressure was identified as the main driver to engage in QRPs; this pressure is likely to drive researcher towards 'cherry picking' their results towards positive findings which are more easily published than negative ones. Selective reporting biases the body of knowledge contributing heavily to the replicability crisis.

In terms of the academic rank, being a PhD student or junior researcher increased the likelihood of engaging in any QRPs, while scientific norm subscription decreased the odds of QRPs and FF. Together, these findings suggest that better training and mentoring from the supervisors' side is needed to improve the scientific performance of young researchers.

Certainly, there is an urgent need to foster responsible research practices among researchers; besides the integrity of the research and researcher, it is also crucial to improve how research in performed, thus, the work in this thesis will focus on the impact that experimental designs and reporting practices have on the variability of results between studies/laboratories.

It is important to keep in mind that although results reproducibility and replicability are a way to confirm scientific findings, this does not necessarily mean that results can be extrapolated to different contexts (*i.e.,* results generalizability) (11). For example, if a study is successfully replicated across independent samples of male mice, these results may not be informative for female mice. Therefore, in order for results to be generalizable and, thus, likely replicable, they must be sufficiently robust, as will be further discussed in the next section.

Robustness of data

In order to replicate results, the outcomes should be consistent albeit the changes implied when re-running the same experiment in somewhat different times/conditions/populations. Data that is resilient to experimental variation (e.g., environmental and genetic variability) will be more likely to generalize to other contexts (i.e., results will have external validity); thus, they will more likely be reproduced (12).

Experimental design

One way in which the robustness of data can be established is through the experimental design as this sets the boundaries for the contexts to which the results may be able to generalize. Currently, best scientific practices advocate for standardizing the animal subjects and their environment by keeping their properties constant overall (12). While standardizing environmental factors is believed to reduce background noise, when taken to an extreme it will provide results that are only informative for, and replicable under the same circumstances in which they were obtained (*i.e.,* idiosyncratic results) (13–16). This is known as the 'standardization fallacy' (11). One of the main setbacks of rigorous standardization of experimental subjects and their environments is that it fails to incorporate the changes in the expression of a phenotype in response to the environmental influences (12). This makes results more likely to be replicated in the same/similar contexts/settings/times than in novel ones (*i.e.,* less robust). Therefore, experimental designs that incorporate diversity of experimental conditions will result in data that is more likely resilient to diverse contexts. However empirical evidence is needed to support this claim.

Representativeness of study samples

Another way to address robustness of data is to increase the representativeness of the study sample. Representativeness of a study population in the context of preclinical research implies that the study population incorporates the biological variability of the population of interest (3). However, rigorous standardization practices aim to reduce the variability of subjects within a study, which would potentially draw the experimental sample further from the target population. In other words, this approach may decrease the representativeness of the study sample and the likelihood of replicating the results under slightly different conditions due to compromised generalizability (2,11).

A way to improve the representativeness of study populations is by diversifying the rearing/husbandry conditions and/or population characteristics (*e.g.,* age, sex, genetic background) within a study; in other words, creating a more heterogeneous population. In this way the between study variability would decrease as each individual study accounts for the unavoidable differences of phenotypic expression between studies/labs (15,17–20).

Altogether, when aiming to produce replicable results that are informative to a target population, one must minimize the risk of bias and ensure the robustness of data. This can be achieved by practices such as blinding and randomization, and likely by diversifying the experimental conditions and population.

TOWARDS RESEARCH (REPLICABILITY AND REPRODUCIBILITY) IMPROVEMENT

As stated in the section above, it has become clear that the way preclinical experiments are usually planned and conducted should be modified if we intend to improve data quality and data interpretation. Specifically, there is room for improvement in the transparency of reporting of studies, particularly, at the level of the experimental design and the representativeness of the study samples. Towards this end, there have been numerous efforts contributing to improve the replicability and reproducibility of preclinical results by promoting rigorous research practices and informative statistical methods such as the ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines to improve the quality of reporting in animal research (21), and the development of initiatives like SYRCLE (Systematic Review Center for Laboratory animal Experimentation) to guide and provide tools to improve the appraisal of systematic reviews and meta-analyses of preclinical data (22). Another initiative is the most recent creation of the open-access Platform for the Exchange of Experimental Research Standards (PEERS) (23) that aims

to provide researchers with a guide on the factors that are most relevant for their experimental design. Overall, the improvement of preclinical studies aims to produce preclinical results that are more accurate, meaningful and informative for translational researchers. For this reason, it is sensible to explore the possible experimental factors affecting data variability in preclinical studies. Certainly, in this way, research practices and study protocols may be adapted accordingly to enhance accuracy, reproducibility and replicability of results across sites.

AIMS AND THESIS OUTLINE

This thesis discusses and tests different perspectives from which data quality and replicability of preclinical results are affected by research practices, and possible ways how to counteract this.

In **Chapter 2,** three different labs part of the EU-AIMS (European Autism Interventions) consortium conducted the same pharmacological experiment in the Shank2-Knockout (KO) rat autism model. The approach taken was to align the apparatus, protocol and data analysis across sites to evaluate their effect on the variability of outcomes between laboratories. Towards this end, the different behavioral outcomes as well as the impact of a pharmacological manipulation were compared across sites. The work described in **Chapter 3** is part of the EQIPD (European Quality in Preclinical Data) consortium aimed to promote research practices that enhance data quality in preclinical studies. This chapter summarizes a three-stage study that evaluated the effects of experimental protocols that differed in the degree of standardization within-laboratory and harmonization across seven labs from academia and industry. The aim of this study was to evaluate impact of protocol harmonization in the variability of results between laboratories.

In **Chapter 4** we investigate the replicability of the behavioral phenotype of the Fragile-X mental retardation mouse model (Fmr1-KO) through a systematic review and meta-analysis. This analysis includes a report on transparency of reporting in light of data quality and replicability of results. Based on the results of this meta-analysis, a study was carried out to assess the behavioral phenotype of the Fmr1-KO. In addition, we explored the possible influence of specific environmental factors that are in fact often not reported in scientific literature. We assessed whether these factors can act as a source of variability thereby potentially contributing to poor replicability. The results from this behavioral study is presented in **Chapter 5**.

As mentioned above, part of the work described in this thesis was carried out as part of the EQIPD consortium. Related to this, in **Chapter 6** we describe the implementation of the quality system (QS), developed by EQIPD members, in an academic lab setting. This chapter exemplifies how the use of this tool can promote rigorous research practices to boost preclinical data quality in academia and industry. More detailed information regarding the EQIPD-QS can be found in **Appendix 1** of this thesis. Finally, **Chapter 7** discusses the overall findings of this thesis, provides conclusions and future perspectives on data quality and replicability of preclinical data.

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. [Internet]. Washington, DC: The National Academies Press.; 2019. Available from: https://doi.org/10.17226/25303.
2. Kafkafi N, Agassi J, Chesler EJ, Crabbe JC, Crusio WE, Eilam D, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci Biobehav Rev. 2018 Apr;87:218–32.
3. Kukull WA, Ganguli M. Generalizability: the trees, the forest, and the low-hanging fruit. Neurology. 2012 Jun 5;78(23):1886–91.
4. Bishop D. Rein in the four horsemen of irreproducibility. Nature. 2019 Apr;568(7753):435–435.
5. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLOS ONE. 2009 Nov 30;4(11):e7824.
6. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science. 1999 Jun 4;284(5420):1670–2.
7. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959 Mar;56(2):81–105.
8. Gopalakrishna G, Riet G ter, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PLOS ONE. 2022 Feb 16;17(2):e0263023.
9. Fanelli D. Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci. 2018 Mar 13;115(11):2628–31.
10. Xie Y, Wang K, Kong Y. Prevalence of Research Misconduct and Questionable Research Practices: A Systematic Review and Meta-Analysis. Sci Eng Ethics. 2021 Jun 29;27(4):41.
11. Würbel H. Behaviour and the standardization fallacy. Nat Genet. 2000 Nov;26(3):263.
12. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020 Jul;21(7):384–93.
13. Paylor R. Questioning standardization in science. Nat Methods. 2009 Apr;6(4):253–4.
14. Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009 Apr;6(4):257–61.
15. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nat Methods. 2010 Mar;7(3):167–8.
16. Wahlsten D, Metten P, Phillips TJ, Boehm SL, Burkhart-Kasch S, Dorow J, et al. Different data from different labs: lessons from studies of gene-environment interaction. J Neurobiol. 2003 Jan;54(1):283–311.
17. Bodden C, Kortzfleisch VT von, Karwinkel F, Kaiser S, Sachser N, Richter SH. Heterogenising study samples across testing time improves reproducibility of behavioural data. Sci Rep. 2019 Jun 3;9(1):8247.
18. Farrar BG, Voudouris K, Clayton N. Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research [Internet]. PsyArXiv; 2020 Aug [cited 2022 Jan 3]. Available from: https://osf.io/2vt4k
19. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biol. 2018 Feb 22;16(2):e2003693.
20. von Kortzfleisch VT, Karp NA, Palme R, Kaiser S, Sachser N, Richter SH. Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' Sci Rep. 2020 Oct 6;10(1):16579.

21. Sert NP du, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biol. 2020 Jul 14;18(7):e3000411.
22. Hooijmans CR, Draper D, Ergün M, Scheffer GJ. The effect of analgesics on stimulus evoked pain-like behaviour in animal models for chemotherapy induced peripheral neuropathy- a meta-analysis. Sci Rep. 2019 Nov 26;9(1):17549.
23. Sil A, Bespalov A, Dalla C, Ferland-Beckham C, Herremans A, Karantzalos K, et al. PEERS — An Open Science "Platform for the Exchange of Experimental Research Standards" in Biomedicine. Front Behav Neurosci [Internet]. 2021 [cited 2022 Feb 12];15. Available from: https://www.frontiersin.org/article/10.3389/fnbeh.2021.755812