Neuroscience and Biobehavioral Reviews 87 (2018) 218–232

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Reproducibility and replicability of rodent phenotyping in preclinical studies

Neri Kafkafi[a,*], Joseph Agassi[a], Elissa J. Chesler[b], John C. Crabbe[c], Wim E. Crusio[d], David Eilam[a], Robert Gerlai[e], Ilan Golani[a], Alex Gomez-Marin[f], Ruth Heller[a], Fuad Iraqi[a], Iman Jaljuli[a], Natasha A. Karp[g], Hugh Morgan[h], George Nicholson[i], Donald W. Pfaff[j], S. Helene Richter[k], Philip B. Stark[l], Oliver Stiedl[m], Victoria Stodden[n], Lisa M. Tarantino[o], Valter Tucci[p], William Valdar[o], Robert W. Williams[q], Hanno Würbel[r], Yoav Benjamini[a]

[a] Tel Aviv University, Israel
[b] The Jackson Laboratory, United States
[c] Oregon Health & Science University, and VA Portland Health Care System, United States
[d] INCIA, Université de Bordeaux and CNRS, France
[e] University of Toronto, Canada
[f] Instituto de Neurociencias CSIC-UMH, Alicante, Spain
[g] Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, UK
[h] Harwell Research Center, UK
[i] University of Oxford, UK
[j] Rockefeller University, United States
[k] University of Müenster, Germany
[l] University of California, Berkeley, United States
[m] VU University Amsterdam, Netherlands
[n] University of Illinois at Urbana-Champaign, United States
[o] University of North Carolina at Chapel Hill, United States
[p] Istituto Italiano di Tecnologia, Italy
[q] University of Tennessee Health Science Center, United States
[r] University of Bern, Switzerland

ARTICLE INFO

Keywords:
Reproducibility
Replicability
GxE interaction
Validity
Data sharing
False discoveries
Heterogenization

ABSTRACT

The scientific community is increasingly concerned with the proportion of published "discoveries" that are not replicated in subsequent studies. The field of rodent behavioral phenotyping was one of the first to raise this concern, and to relate it to other methodological issues: the complex interaction between genotype and environment; the definitions of behavioral constructs; and the use of laboratory mice and rats as model species for investigating human health and disease mechanisms. In January 2015, researchers from various disciplines gathered at Tel Aviv University to discuss these issues. The general consensus was that the issue is prevalent and of concern, and should be addressed at the statistical, methodological and policy levels, but is not so severe as to call into question the validity and the usefulness of model organisms as a whole. Well-organized community efforts, coupled with improved data and metadata sharing, have a key role in identifying specific problems and promoting effective solutions. Replicability is closely related to validity, may affect generalizability and translation of findings, and has important ethical implications.

1. Introduction

In recent years the scientific community, pharmaceutical companies, and research funders have become increasingly concerned with the proportion of published "discoveries" that could not be replicated in subsequent studies, and sometimes could not even be reproduced in reanalysis of the original data. Such evidence is increasingly seen as a problem with the scientific method, impugning the credibility of science as a whole. Prominent institutions and journals, including the National Institutes of Health (NIH), the National Academy of Science

* Corresponding author at: Tel Aviv University, Department of Statistics and Operations Research, Ramat Aviv, Tel Aviv 66978, Israel.
E-mail address: nkafkafi@post.tau.ac.il (N. Kafkafi).

https://doi.org/10.1016/j.neubiorev.2018.01.003
Received 25 October 2016; Received in revised form 13 December 2017; Accepted 11 January 2018
Available online 31 January 2018

(NAS), *Science*, and *Nature*, have recently reconsidered of their policies due to this issue. However, there is still confusion and controversy regarding the severity of the problem, its causes, and what should be done about it, how, and by whom.

In the field of rodent phenotyping, failure of replicability and reproducibility had been noted even before such concerns were widespread, and currently the NIH considers the problem to be especially prevalent in preclinical research. The issue seems further tied to several other complicated methodological challenges, such as handling the potentially complex interaction between genotype and environment, defining and measuring proper behavioral constructs, and using rodents as models for investigating human diseases and disorders. Reproducibility and replicability are crucial in all fields of experimental research, but even more so in animal research, where the lives and welfare of the animals are valuable for ethical reasons, and should not be wasted for inconclusive research. In January 2015, researchers involved in the study of reproducibility and replicability gathered at Tel Aviv University to discuss these issues. These researchers came from various disciplines including genetics, behavior genetics, behavioral neuroscience, ethology, statistics, bioinformatics and data science.

The present paper consists of eight sections, each dedicated to a central theme. In each section we attempt to summarize the consensus opinion or most widely held views on the topic, while also representing more controversial positions. While offering examples, recommendations and insights in multiple contexts, we avoid making a list of guidelines that would be too definitive, given the current state of knowledge and consensus. Full conference proceedings are available as a set of video clips (see links in the acknowledgements section). All authors agree that this paper reflects the complexity of replicability and reproducibility issues, even when restricted to a single area of research, yet it also points at practical ways to address some of these issues.

2. Reproducibility and replicability in general science: a crisis?

The ability to verify empirical findings wherever and whenever needed is commonly regarded as a required standard of modern experimental science. This standard was originally established in the 17th century, by Robert Boyle and other scientists of the Royal Society according to their motto *nullius in verba* (“take nobody’s word”). These pioneers of experimental science regarded the ability to replicate results as an acid test differentiating science from one-time “miracles”. Their criterion for a scientific fact was (following a then common judicial dogma of two witnesses required for a valid testimony) something measured or observed in at least two independent studies (Agassi, 2013). In a case that may have been the first debate over the replicability of a scientific discovery, the Dutch scientist Christiaan Huygens noted a phenomenon related to vacuum in Amsterdam, and was invited to Boyle’s laboratory in London in order to replicate the experiment and show that the phenomenon was not idiosyncratic to his specific laboratory and equipment (Shapin and Schaffer, 1985). Ronald Fisher generalized the Royal Society criterion to more than two replications in his 1935 classic “The Design of Experiments”, writing: “we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results” (Fisher, 1935, p.14). This quote illustrates how the common method of statistical significance, already when it was first conceived, was closely tied with the concept of replicating experimental results. This concept served science well throughout its history, but non-replicable results have surfaced more often in recent years, attracting much attention.

In the field of rodent phenotyping, the problem has in fact always been present, and was recognized in the influential study by Crabbe et al. (1999) before it was noticed in many other fields. However, the issue is by no means unique to rodent phenotyping. For instance, difficulties in replicating discoveries when dissecting the genetics of complex traits in humans motivated the move to far more stringent statistical threshold guidelines proposed by Lander and Kruglyak (1995).

Some notorious recent examples of poor credibility in general science include non-replicable methods of cancer prognosis (Potti et al., 2006, refuted by Baggerly and Coombes, 2009, and retracted), “voodoo correlations” in brain imaging (Vul et al., 2009), “p-value hacking” (Simmons et al., 2011) and Excel coding errors that affected global economic policies (Pollin, 2014). A large community effort (Open Science Collaboration, 2015) recently attempted to replicate the findings of 100 papers in several leading psychology journals, and reported that 64% of the replications did not achieve statistical significance (but see Gilbert et al., 2016). A similar replication project in the field of cancer research (Errington et al., 2014) has just reported preliminary results: of 5 attempted replications, two were replicated, one clearly failed to replicate, and two were unclear due to technical considerations (Nosek and Errington, 2017). The current situation is sometimes referred to as the “credibility crisis”, “replicability crisis” (e.g., Savalei and Dunn, 2015), or “reproducibility crisis” (e.g., Peng, 2015) of recent science, and led prominent scientific journals and institutes to reconsider their policies (Landis et al., 2012; Nature Editorial, 2013; Collins and Tabak, 2014; McNutt, 2014; Alberts et al., 2015). Collins and Tabak specifically mentioned preclinical studies as prone to reproducibility and replicability problems, and Howells et al. (2014) blame the recurrent failure of drug candidates in clinical trials on lack of rigor in preclinical trials. Yet aside of general useful recommendations such as increasing sample sizes, including both sexes when possible, and improving statistical education, it is not clear what the new policies should be.

Ironically, there is currently no scientific consensus even over the name of the problem and the meaning of basic terms, confusing the discussion even further (Goodman et al., 2016). The terms replicable, reproducible, repeatable, confirmable, stable, generalizable, reviewable, auditable, verifiable and validatable have all been used; even worse, in different disciplines and fields of science, these terms might have orthogonal or even contradictory meanings (Kenett and Shmueli, 2015; Goodman et al., 2016). Following the now common term “Reproducible Research” in computer science (Diggle and Zeger, 2010; Stodden, 2010, 2013), a useful distinction was offered by Peng (2011), Peng (2015) and (Leek and Peng, 2015): “reproducibility” is concerned with reproducing, from the same original data, through reanalysis, the same results, figures and conclusions reported in the publication. “Replicability”, in comparison, is concerned with replicating outcomes of another study, in a similar but not necessarily identical way, for example at a different time and/or in a different laboratory, to arrive at similar conclusions in the same research question. We will use the above distinction in the remaining sections. However, note that other researchers recently suggested a similar distinction with the opposite terminology (Kenett and Shmueli, 2015). The NIH now uses the catch-all term “rigor” to denote adequacy or even goodness of experimental design, metadata, and analytic methods that should hopefully lead to higher rates of replicability and reproducibility (Lapchak et al., 2013).

Another categorization (Stodden, 2013) distinguishes between empirical reproducibility, computational reproducibility and statistical reproducibility. (Stodden, 2010, 2013) suggested that computational reproducibility is currently the most problematic. When viewing the objective of the scientific method as “rooting out error”, the deductive branch of mathematics (statistics included) has already developed its standards for mathematical proof, and the empirical branch (life sciences and animal phenotyping included) has already developed its standards for hypothesis testing and method reporting. It is computation-based research that has yet to develop its own standards for reproducibility, including data and code sharing (Stodden et al., 2013).

Ostensibly, science should not require trust in authority – it should be “show me”, not “trust me” (Stark, 2015). Yet in reality, most scientific publications today amount to saying “trust me”. The typical scientific paper does not give access to the raw data, the code, and other