Journal of Integrative Neuroscience

J. Integr. Neurosci. **2024**; 23(2): 30
https://doi.org/10.31083/j.jin2302030

Editorial

Reproducibility in Preclinical *in Vivo* Research: Statistical Inferences

Charmaine J.M. Lim[1], Sanna K. Janhunen[2], Gernot Riedel[1,*]

[1]Institute of Medical Sciences, University of Aberdeen, AB25 2ZD Aberdeen, UK
[2]Organon R&D Finland, 20520 Turku, Finland
*Correspondence: g.riedel@abdn.ac.uk (Gernot Riedel)
Academic Editors: Yoshihiro Noda and Rafael Franco
Submitted: 23 November 2023 Accepted: 30 November 2023 Published: 4 February 2024

The crisis of irreproducible preclinical *in vivo* research does not only come about from poor study design and replication failures. Rather, a core problem is statistical inference and the over-simplification of statistical models, flawed interpretation, and the communication of results. The actual statistical method is rarely the issue [1]. Instead, statistical concepts, such as the upmost careful and precise terminologies, in-depth understanding of the assumptions and specifications of distributions and random and fixed effects, and the applications of mixed models [2] that are the most overlooked, misunderstood and misrepresented issues in preclinical *in vivo* research. The current logic of (and first decision in) interpreting outcomes in translational neuroscience is regrettably still derived from the use of dichotomous null-hypothesis significance testing (NHST). NHST and the use of the term 'statistical significance' has been deeply ingrained into scientific practice [3], with unfounded resistance against other alternative statistical approaches.

Nowadays, NHST automatically dictates that statistically significant findings must hold true, and any insignificant results must therefore be disregarded. The original intention of NHST and *p*-values to invite further scrutiny of results [4,5] seems irreversibly lost. The scientific community inexplicably values them highly, despite being conceptual themselves and carrying theoretical baggage that has, over the years, been misappropriated and misunderstood [6]. Perhaps getting rid of *p*-values could rid this baggage but this has been the argument for a century with little success because everyone knows that everyone uses dichotomous statistics. If anything, our reliance on *p*-values has increased over time [7].

Indeed, the use of the term "statistically significant", or any variation thereof (such as "statistically different", "$p < 0.05$" or the use of asterisks in figures and tables), should be made redundant. Studies focusing solely on *p*-values and statistical significances as determinants of the importance of a finding should certainly be sent back. Statistical significance generally does not equate to scientific or clinical significance and 'non-significant' findings do not equate to evidence of absence [8]. Having two labels to an outcome can also coerce *p*-hacking, selective reporting, and publication bias [9]. All results must, in fact, be reported and discussed regardless of their statistical significance/*p*-values. Stakeholders can exacerbate the publication bias since they tend to lean towards statistical evidence that aids in making clear 'yes/no' decisions.

Biological/neuroscientific data is generally comprised of noisy signals and uncertainties. Considering the estimates (e.g., hazard ratios, interval ratios, mean differences), confidence intervals, observed effects and limits, can help to interpret the compatibility of their values with the data and potentially the corresponding clinical relevance of the findings. For example, the interval estimate can be constructed to express uncertainty by several approaches: in a frequentist approach, *p*-values are complemented by a confidence interval in every null-hypothesis test; in Bayesian paradigms, credible intervals or support intervals can express uncertainty [10]; and in randomisation-based approaches, uncertainty can be quantified by bootstrapped intervals [11]. Values which are qualitatively very different (based on the width of the interval estimate) can suggest that the estimate is very noisy and that firm conclusions should be avoided.

Behaviour in animal models most definitely leads to mixed and noisy results that may not be translatable to humans. Behavioural proxies are complex and a culmination of multifaceted, often subtle, systems that may be present in different species in specific ways [12]. It is not unlikely that subtle behavioural changes are often missed, or data anomalies are given undue consideration. A Bayesian approach can express probabilistic statements such as there is a probability of 0.75 that the risk ratio is <0.9. Estimation statistics can also divert the attention from qualitative "yes/no" answers to the quantitative question of "how different?" by conveying magnitudes and uncertainties [11,13] in simple data sets. Estimation statistics, however, requires caution and is simply not enough when interpreting complex behaviour. Estimation statistics can be evaluated with other statistical models such as Principal Component Analysis or clustering methods that provide a simple output for multiple complex systems by means of scoring each test to represent each behavioural trait of interest [14]. Bootstrapping is another particularly reliable way in small and highly skewed datasets to estimate standard errors and confidence intervals without relying on assumed probability distributions [15].

Of course, Bayesian or estimation statistics seem like new and exciting alternatives, but they are the cryptocurrency equivalent in statistics: everyone knows of them but

only has a vague notion and cannot be sure of exactly what they mean and where or how they can be implemented. And just like *p*-values and NHST, Bayesian and estimation statistics can be used inappropriately if their conceptual foundations are not understood without understanding their uncertainty. The problem, ultimately, does not lie in producing labels or the use of any other statistical measures such as intervals or Bayes factors as a means to dichotomise study outcomes. But for a start, we could remove significance thresholds, which are arbitrary anyway, and the use of dichotomous statistical measures to address issues of replicability, reduce significance chasing, *p*-hacking or data dredging, to bring about less publication bias, inflated effect sizes, and thereby produce more reliable research.

So what now? Unfortunately, this editorial cannot provide an ultimate solution to replacing the overused phrase of 'statistical significance' nor recommend a one-size-fits-all approach to any statistical inference. The principles for the use of statistics are solid and readily accessible; yet translational behavioural neuroscience remains stubbornly dominated by sub-standard or out-dated strategies in statistical analysis with the corollary that reproducibility and replicability are low. Greater pressure and education must be placed on investigators, journal editors, reviewers and funding bodies to rigorously demand and enforce the reproducibility of the research. Preclinical *in vivo* research is the foundation for the development of high-quality clinical therapies and diagnostics. Investigators need to abandon the long-existing tradition of underestimating the complexity of animal behaviour and overestimating their intuition that feeds their reluctance to accept potential flaws in their statistical methodologies or data. As researchers, we want to publish gold, and all that is required are coherent research questions and plans, solid statistical methods, and truthful conclusions.

Author Contributions

CL, GR and SJ wrote the article. CL conducted a literature review. All authors reviewed and edited the manuscript and have read and approved its final version. All authors have participated equally and sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest. Gernot Riedel is serving as one of the Editorial Board members of this journal. We declare that Gernot Riedel had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to Yoshihiro Noda and Rafael Franco.

References

[1] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019; 567: 305–307.

[2] Bello NM, Renter DG. Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. Journal of Dairy Science. 2018; 101: 5679–5701.

[3] Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in Human Neuroscience. 2017; 11: 390.

[4] Fisher R. Statistical methods and scientific induction. Journal of the Royal Statistical Society: Series B (Methodological). 1955; 17: 69–78.

[5] Edgeworth FY. Methods of Statistics. Journal of the Statistical Society of London. 1885; 181–217.

[6] Goodman SN. Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. American Statistical Association. 2019; 73: 26–30.

[7] Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016; 315: 1141–1148.

[8] Altman DG, Bland JM. Absence of evidence is not evidence of absence. British Medical Journal. 1995; 311: 485.

[9] Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114: 3714–3719.

[10] Kruschke JK, Liddell TM. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review. 2018; 25: 178–206.

[11] Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. Nature Methods. 2019; 16: 565–566.

[12] Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349: aac4716.

[13] Calin-Jageman RJ, Cumming G. Estimation for Better Inference in Neuroscience. eNeuro. 2019; 6: ENEURO.0205-19.2019.

[14] Harrison DJ, Creeth HDJ, Tyson HR, Boque-Sastre R, Isles AR, Palme R, *et al*. Unified Behavioral Scoring for Preclinical Models. Frontiers in Neuroscience. 2020; 14: 313.

[15] Bland JM, Altman DG. Statistics Notes: Bootstrap resampling methods. British Medical Journal. 2015; 350: h2622.