Noname manuscript No.
(will be inserted by the editor)

Enhancing statistical reproducibility in the biomedical sciences: practical guides for the everyday researcher

Ariel Mundo Ortiz · Bouchra Nasri ·

Received: date / Accepted: date

Abstract Reproducibility continues to be a major goal of biomedical research. However, the field still struggles to use Statistics (a core component of reproducibility) in a reproducible way. In this paper, we present some of the reasons that we believe contribute to this ongoing disconnect between Statistics and reproducible biomedical research, and we also present guidelines aimed to help trainees and researchers to internalize the need of using Statistics in a manner that enhances reproducibility.

Keywords keywordA, keywordB ·

1 Background

The "reproducibility crisis" in science has affected particularly biomedical research , where it is believed that many articles describe results that cannot be replicated[1,2]. Although it is recognized that some indicators of reproducibility have improved in recent years in publications in the field, it is also recognized that there is still a large room for improvement[3].

In recent years, different studies have identified practices that could be implemented to improve various aspects of reproducibility in the field, suggesting (among other recommendations) greater involvement of senior investigators in the data collection/analysis process[4], and the use of reproducible tools and

Ariel Mundo Ortiz
École de santé publique, Université de Montréal,
Centre de recherches mathématiques, Université de Montréal
E-mail: ariel.mundo.ortiz@umontreal.ca

Bouchra Nasri
École de santé publique, Université de Montréal
E-mail: bouchra.nasri@umontreal.ca

workflows[5,6]. However, the incorrect use of statistics is also a key factor that limits reproducibility[7,8].

It is well established that statistics are misused in biomedical research[8], but we believe it is important to also highlight some of the causes that drive this misuse in order if we are to improve and remedy this issue. We believe that an important factor that drives the improper use of statistics is the fact that statistics is a subject that is seen as abstract and obscure by biomedical trainees and researchers. In fact, it is known that statistics is something that worries students[9]. The biomedical field is not exempt from seeing statistics as a "necessary evil" and consequently, continuing to perpetuate systemic issues that affect reproducibility. Overall, this results in a disconnect between what *should* be done to address statistical reproducibility in biomedical research, and the approach that *is* taken by the field.

We believe that presenting guidelines that address some of the issues that the field faces in terms of statistical reproducibility is a necessary step towards improving statistical practices and preparing future generations of scientists to routinely perform research that is open and reproducible.

2 Guidelines

2.1 "Why are we doing this?"

The very abstract and quantitative nature of Statistics often leads students to worry and to adopt a "memorisation" strategy to approach it[10], a phenomenon that trainees in biomedical research are not exempt of. However, memory alone does not suffice to successfully analyze "real" data: it has been shown that researchers often incorrectly analyze experimental designs, and checks that are to be made before experiments are done after experiments have been completed[11].

It is therefore important that trainees focus on understanding the "why" (the rationale) of the statistical tools they intend to use. Indeed, Statistics cannot be appropriately used without understanding the assumptions and the basic theory that underlies them[12]. Researchers tend to think that Statistics are only needed when a statistical test needs to be performed, but the truth of the matter is that Statistics play a crucial role at every stage of research: they are needed to determine the number of observations required to achieve a certain statistical power, as well as to determine the choice of experimental design that needs to be used to answer the question of interest. These are steps that precede a "statistical test", and they need to be taken into careful consideration before any experiments or data are to be collected.

2.2 Learning statistics is necessary

In the previous point we have indicated that researchers need to learn the foundations of Statistics in order to adequately use them. This is a view shared

by many others, that have emphasized the importance of statistical education as a way of addressing the reproducibility crisis[13]. However, the very nature of Statistics makes it distinct from learning physics, or mathematics[10]. It has been suggested that a problem-based learning approach might be best suited to teach Statistics[14], but it is possible that most trainees will not be able to benefit from that approach in the Statistics classes they are required to take (if that is the case). This implies that the most likely option for a trainee is a combination of formal Statistical training (taking classes in a University) and self-guided study. After all, not only there is a wide range of variation among the academic requirements set by each program[15] but it is also possible that trainees are interested in a particular statistical topic without committing the time and resources needed for a full course.

Fortunately, the educational resources of Statistics have increased vastly over the last decade, and nowadays there are multiple materials that cover a wide range of statistical topics without excessive mathematical complexity. For example, those interested in revisiting statistical foundations will find an excellent resource in the work of James et al.[16]. Topics on generalized linear models and generalized linear mixed models (which we believe are extremely important for biomedical researchers to learn, but that might not be covered in the Statistics courses required by their programs) can be found in McCulloch[17], Dobson[18], and Stroup[19].

2.3 You should not aim to do "everybody does"

This point might seem redundant in the light of what we indicated above about the importance of learning Statistics. However, our own experience has shown that a deterrent for trainees to learn Statistics is that it suffices to mimic the analyses they have seen in a paper ("what everybody does"). It is tempting to repeat the analysis presented in a previous study, as it might seem that because the study passed peer review, its methodology (including its statistical analyses) *should* be correct. However, without a clear understanding of the assumptions behind the analysis is correct in the first place. The truth is that that is seldom the case, as Hardwicke et al.[20] showed that most leading biomedical journals do not perform specialized statistical reviews on papers they published (only 23% reported that they did statistical analyses).

Sadly, statistical errors plague biomedical studies[21], and without a solid understanding of the assumptions and limitations of a statistical method, researchers will be unable to critically assess the analyses presented and determine if the methodology presented is applicable in their case. Peer review is not perfect, and researchers need to internalize the fact that repetition is not enough to achieve reproducibility.

2.4 Your statistical analyses need to be reproducible

Biomedical data is complex and nowadays, computational skills are necessary to successfully analyze the datasets that are obtained as a result of experiments[5]. However, there seems to be gap between the computational skills that trainees acquire to collect their data, and the skills they possess to perform statistical analyses of such data. Because of the "memorizing" approach that we have mentioned above, many researchers prefer a "click" approach to perform their statistical analyses (using a program that only requires them to select certain options from a menu). Although a "click" approach is apparently more efficient from a time perspective, the biggest trade-off is that there is no real understanding from the user of what is actually happening behind the scenes[22].

Others have indicated the importance of coding to create a workflow that minimizes the errors associated with manual manipulation[23], an opinion we concur with and believe is equally applicable in the context of statistical analyses. Although statistical analyses can be performed by a myriad of different computational tools (such as R, Python, and Julia), no statistical analysis is complete if it is not reproducible.

There are some tools that are specifically suited to allow reproducible workflows. Following on the recommendations of Brito et al.[5] regarding the use of open source tools to create reproducible workflows, in Table 1 we provide a list of open source tools that are designed to combine text and computations (therefore allowing to create reproducible documents), that are easily accessible to biomedical researchers, that support multiple computational languages, and that have multiple resources (such as examples, books, and guides) that can help researchers familiarize themselves with how they work.

Table 1: Tools that allow for reproducible statistical analyses

Tool	Characteristics	Languages supported	Resources
RMarkdown	Allows to create reproducible documents (notebooks, reports, books, scientific articles) that combine coding and text. Output formats include HTML, PDF, MS Word, Beamer and others	R, Python, SQL, Julia and others	Xie et al.[24], an online version of the book can be found at `https://bookdown.org/yihui/rmarkdown/`), examples can be found at `https://rmarkdown.rstudio.com/gallery.html`

Tool	Characteristics	Languages supported	Resources
Bookdown	Allows to create reproducible documents and follows the same syntax of RMarkdown, but includes added capabilities such as cross-referencing and facilitating the creation of documents (such as books) that are composed of multiple RMarkdown documents.	R, C/C++, Python, Fortran, Julia, SQL, Stan and others	Xie[25], an online version of the book can be found at `https://bookdown.org/yihui/bookdown/` `https://bookdown.org/` contains examples of books created using Bookdown.
Quarto	Publishing system for scientific and technical documents that is compatible with VS Code, RStudio, and Jupyter Notebooks. Documents can be compiled in HTML, PDF, MS Word, Beamer, Shiny, MS PowerPoint, Revealjs presentations, and many others	R, Python, Julia, Observable	`https://quarto.org/` contains multiple examples, tutorials, and use guides

2.5 Models are just that, models

- Biology is complex
- Models are a simplification
- They offer an explanation, but that is not the only explanation

2.6 Significance should not be driven by a p-value

Perhaps the aspect of statistical reproducibility that biomedical researchers struggle the most is the concept of significance and its association with a *p-value* below 0.05. Much has been said about the limitations of statistical tests and p-values, and how it is wrong to dichotomize the "significance" of a result on the basis of a p-value cutoff, and yet, this is still a prevalent practice in the field.

Here, we will not attempt to provide another repetition of the facts that others have so eloquently provided about this topic (we refer the reader to the excellent works of Ziliak and McCloskey[26], Greenland et al.[27], Wasserstein and Lazar[28], and Chia[29] for discussions in detail), but we much rather try to shed some light on why the "p-values <0.05 equals significance" is so prevalent in biomedical research.

We believe that this problem has multiple facets: First, there is the issue of how researchers view statistics, which closely relates to that we described in Point 1. Researchers view Statistics as a black box where multiple obscure terms such as "distribution", "likelihood", "parameters", and many greek letters are mixed up along a language whose technicalities are incomprehensible and confusing. In a sense, it is true that technical statistical language is a completely different beast from the research language that biomedical researchers commonly employ; comprehensibly, trying to learn a new technical language might seem as a daunting task for which researchers, facing already time constraints due to research and academic life, might not feel to have the time or resources to learn. Adding to this problematic, introductory statistical courses and textbooks typically do not discuss the limitations of p-values in a clear way[27].

Second, the dichotomization of significance is the driving force that the field uses to measure research outcomes. In other words, researchers perpetually suffer from "significant-itis"[29] (believing that results are only good if the p-value is <0.05) because such metrics are ubiquitously presented in publications as the correct metric to measure the success or failure of a research outcome.

These two facts then, create a vicious cycle where researchers are perpetually working to find "significant results", and makes trainees believe early on that that is the correct way of validating research outcomes. We believe it is important to remind researchers that statistical significance does not equate to clinical significance[27], that p-values are just a way to determine how the data behaves under the model assumptions, and that they are the result of historical and philosophical choices made by people[30,31].

The other part of the problem is what the view of statistical tests as an obscure and complicat

p-values should not be the final goal of some data, but rather the interpretation by the researcher is what is important.

- p-values and positive results continue to dominate the field
- what a p-value actually tells (maybe example with data)?
- p-value should not be the goal of a study

References

1. Begley CG, Ioannidis JP. Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation research.* 2015;116(1):116-126.
2. Oakden-Rayner L, Beam AL, Palmer LJ. Medical journals should embrace preprints to address the reproducibility crisis. *International Journal of Epidemiology.* 2018;47(5):1363-1365. doi:10.1093/ije/dyy105
3. Wallach JD, Boyack KW, Ioannidis JP. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLoS biology.* 2018;16(11):e2006930.

4. Samsa G, Samsa L. A guide to reproducibility in preclinical research. *Academic Medicine*. 2019;94(1):47-52. doi:10.1097/acm.0000000000002351

5. Brito JJ, Li J, Moore JH, et al. Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience*. 2020;9(6). doi:10.1093/gigascience/giaa056

6. Papin JA, Gabhann FM, Sauro HM, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. *PLOS Computational Biology*. 2020;16(5):e1007881. doi:10.1371/journal.pcbi.1007881

7. Erwin B. Montgomery Jr. *Reproducibility in Biomedical Research*. Elsevier; 2019. doi:10.1016/c2018-0-02296-3

8. Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochemia medica*. 2015;25(1):5-11.

9. Ralston K. "Sociologists shouldn't have to study statistics": Epistemology and anxiety of statistics in sociology students. *Sociological Research Online*. 2019;25(2):219-235. doi:10.1177/1360780419888927

10. Ramsey JB. Why do students find statistics so difficult. *Proccedings of the 52th Session of the ISI Helsinki*. Published online 1999:10-18.

11. Kitchenham B, Madeyski L, Brereton P. Problems with statistical practice in human-centric software engineering experiments. In: *Proceedings of the Evaluation and Assessment on Software Engineering*. ACM; 2019. doi:10.1145/3319008.3319009

12. Marino MJ. Statistical analysis in preclinical biomedical research. In: *Research in the Biomedical Sciences*. Elsevier; 2018:107-144. doi:10.1016/b978-0-12-804725-5.00003-3

13. Patil S, Satagopan J. Building and teaching a statistics curriculum for post-doctoral biomedical scientists at a free-standing cancer center. *CHANCE*. 2022;35(1):56-64. doi:10.1080/09332480.2022.2039036

14. Ekmekci O, Hancock AB, Swayze S. Teaching statistical research methods to graduate students: Lessons learned from three different degree programs. *International Journal of Teaching and Learning in Higher Education*. 2012;24(2):272-279.

15. Gatchell D, Linsenmeier R. Similarities and differences in undergraduate biomedical engineering curricula in the united states. In: *2014 ASEE Annual Conference & Exposition Proceedings*. ASEE Conferences. doi:10.18260/1-2--23015

16. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Springer US; 2021. doi:10.1007/978-1-0716-1418-1

17. McCulloch CE, Searle SR. *Generalized, Linear, and Mixed Models*. John Wiley & Sons; 2004.

18. Dobson AJ, Barnett AG. *An Introduction to Generalized Linear Models, Fourth Edition*. 4th ed. CRC Press; 2018.
19. Stroup WW. *Generalized Linear Mixed Models*. CRC Press; 2012.
20. Hardwicke TE, Goodman SN. How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey. Koletsi D, ed. *PLOS ONE*. 2020;15(10):e0239598. doi:10.1371/journal.pone.0239598
21. Lang T. Twenty statistical errors even you can find in biomedical research articles. *Croatian Medical Journal*. 2004;45:361-370.
22. Deardorff A. Why do biomedical researchers learn to program? An exploratory investigation. *Journal of the Medical Library Association: JMLA*. 2020;108(1):29.
23. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. Bourne PE, ed. *PLoS Computational Biology*. 2013;9(10):e1003285. doi:10.1371/journal.pcbi.1003285
24. Xie Y, Allaire JJ, Grolemund G. *R Markdown*. CRC Press; 2018.
25. Xie Y. Bookdown: Authoring books and technical documents with r markdown. In: *The R Series*. CRC Press; 2016:47-66.
26. Ziliak ST, McCloskey D. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press; 2008.
27. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3
28. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108
29. Chia K-S. "Significant-itis" — an obsession with the p-value. *Scandinavian Journal of Work, Environment & Health*. 1997;23(2):152-154. Accessed December 22, 2022. `http://www.jstor.org/stable/40966624`
30. Huberty CJ. Historical origins of statistical testing practices. *The Journal of Experimental Education*. 1993;61(4):317-333. doi:10.1080/00220973.1993.10806593
31. Freedman D. From association to causation : Some remarks on the history of statistics. *Journal de la Société française de statistique*. 1999;140(3):5-32. `http://www.numdam.org/item/JSFS_1999__140_3_5_0/`