Observations from Statistical Review Editors: A Commentary

Matt Spick[1], Jan Higgins[2], Cynthia L. Green[3], Roland Matsouaka[3], Daniel B. Shin[4], Russell P. Hall III[5] and Nophar Geifman[1]

JID Innovations (2024) **4,** 100302; doi:10.1016/j.xjidi.2024.100302

INTRODUCTION

Reproducibility and replicability are crucial components of the scientific method, but they may be compromised when there are inherent issues related to a study and analytic choices such as statistical errors or misalignments between the study's objectives and implementation. Indeed, statistical errors and misunderstandings contribute to low reproducibility and replicability, hindering independent verification or changes in the direction of research (McNutt, 2014). Such problems can easily occur in health science, where there are many confounding factors and low prior odds of genuine findings (Ioannidis, 2005). Guidelines for statistical reporting that can minimize these issues are well-established but are not always followed. In January 2023, to help address these challenges in a more targeted way, *JID Innovations* established a statistical review board as part of its overall editorial process, nominating editors with expertise in statistical analysis and data science (Hall, 2023). All submissions to the journal are reviewed by 1 of the statistical review editors to provide specialist evaluation and feedback on study design, statistical tests, and analyses, as well as bioinformatic aspects of the manuscript. In this commentary, common themes identified by statistical review editors in their peer reviews are brought forth along with comments that are made during the routine peer review process, to highlight prevalent issues in statistical methodologies and reporting seen in submissions to *JID Innovations*. The goal of this commentary is to propose easy steps that authors can take to inform study design at the outset of any data-driven project, reduce the number of potential revisions to statistical methodology and presentation in the original submission, and ultimately improve the reproducibility and replicability of the work published in *JID Innovations*, with the added benefit of a more efficient submission process.

METHODS

All peer reviews are logged centrally by *JID Innovations*, including the full text of the reviewer comments and suggestions. Manuscripts that had completed the review process were examined for the purpose of this commentary; the overall list was then filtered to include manuscripts with reviews by a statistical review editor. In total, 42 articles were extracted alongside the statistical editor reviews, and the resulting qualitative dataset was then reviewed using a reflexive thematic analysis approach (Saunders et al, 2023). These themes were then discussed as a group to identify common issues among the peer reviewers relating to study design, methodology as well as statistical tests, analyses, and reporting.

RESULTS AND DISCUSSION

Figure 1 provides a visualization of a data analytics process as might be employed for basic, translational, and clinical research, highlighting the intermediate feedback mechanisms and reporting matters that underpin the majority of the issues identified in this work. These are also included, along with less common but still important errors often encountered during the peer review process (Table 1).

Study design and recruitment

Cohort sizes were a frequent source of problems in the reviewed manuscripts. This can be addressed partly by better choices of statistical test or model, but can also be dealt with by power analysis at the design stage to ensure that the final sample size n would be appropriate to answer the research question. There were also bias problems in recruitment, for example, a failure to recruit appropriate controls, a lack of racial or ethnic diversity, and lack of reporting on female participation. Such issues are not limited to small studies: even large-scale biobank recruitment leads to populations that do not reflect the age or ethnic profiles of their wider populations (Beesley et al, 2020; Schoeler et al, 2023). These differences, which might typically include a skew toward older, higher-income cohorts, and lower-deprivation cohorts (Fry et al, 2017), are difficult to overcome, especially when there is a voluntary component to participation. In almost all cases, the issues identified in recruitment would best be dealt with by involving expert statistical advice at the study design stage. In addition, data management and stewardship should also be addressed at the design stage rather than as an afterthought, ideally by following the FAIR (Findable,

[1]*School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom;* [2]*JID Innovations, Cleveland, Ohio, USA;* [3]*Department of Biostatistics and Bioinfroamtics, Duke University School of Medicine, Durham, North Carolina, USA;* [4]*Department of Dermatology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA; and* [5]*Department of Dermatology, Duke University School of Medicine, Durham, North Carolina, USA*

Correspondence: Nophar Geifman, School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford GU2 7XH, United Kingdom. E-mail: n.geifman@surrey.ac.uk

Cite this article as: JID Innovations 2024;4:100302

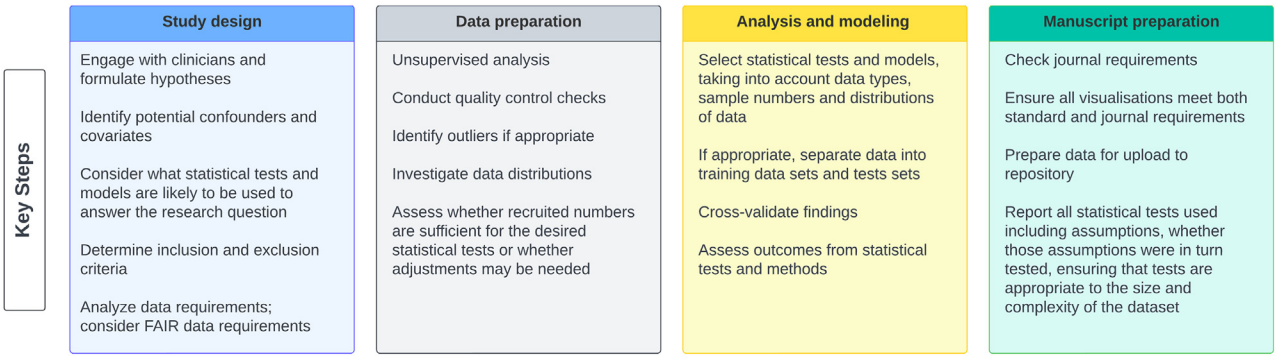

Figure 1. Key steps in the data analytics process as applied to health data science, created in Lucid (lucid.co). FAIR, Findable, Accessible, Interoperable, and Reusable.

Accessible, Interoperable, and Reusable) guiding principles (Wilkinson et al, 2016).

Appropriateness of test statistics, analyses, and models

There are many statistical methodologies that can be employed for data analysis in the context of a health or biomedical study. The appropriateness of each approach is determined by the size of cohorts recruited, assumptions around the data distribution, and the research question to be answered. Nevertheless, often, an unsuitable method was chosen in the reviewed manuscripts. For example, one submitted manuscript employed multivariable analyses for small sample sets featuring high multicollinearity, where simpler models such as a penalized ridge regression model (eg, Least Absolute Shrinkage and Selection Operator regression) or even simple univariable regression may have been better suited for the data. Conversely, reviewers also noted the use of overly simplistic models, such as using *t*-tests for multiple time points where a mixed-effects regression model for longitudinal data would have been more appropriate. Another issue raised for several submitted manuscripts was the application of standard workflows, without an appropriate assessment of whether the requirements of such workflows were met. A common example is the assumption of normal or log-normal distributions, when appropriate formal tests for normality should be employed, and nonparametric methods should be considered as an alternative. A less obvious example is the use of false discovery rate (FDR) correction as a blanket means of producing highly conservative results, often when variables are not fully independent, and common FDR methods such as Benjamini–Hochberg are too conservative. Authors may overlook adjusting for multiple comparisons altogether, possibly as it may not be obvious because it may appear in subsequent comparisons following an ANOVA test. The reasons for multiple corrections can stem for a number of analytic decisions or based on the study design and objectives (Li et al, 2017). In some instances, methods were not reported at all, and test statistics were referred to only as adjusted. Lack of detail regarding statistical tests and models naturally hinders reproducibility and replicability. Most problems grouped under this theme could easily be addressed by avoiding complex models and solutions except where they are clearly required to solve the problem at hand, by always considering whether the assumptions that underly standard workflows are being met in the author's dataset, and by being explicit in Methods sections about the exact tests and assumptions made.

Reporting and presentation

The statistical review editors found that a significant number of submitted manuscripts failed at this fundamental task. The most common error under this theme was a failure to report confidence intervals, which in some manuscripts with small cohorts could be wide. Reporting of *P*-values or other statistics to an inappropriate level of detail was also frequently mentioned by the reviewers, especially in works analyzing small cohorts; reporting excessive numbers of decimal places does not make an author's results more precise! A further fundamental mistake—seen surprisingly often—was a failure to properly present graphs and other figures. Labels for axes were often missing or incomplete, and authors frequently did not provide self-contained figure legends that contained all the information required to interpret a figure, without having to go through the text of a paper for explanations. Other presentational errors requiring correction included imprecise language, for example, the use of the term multivariate analysis when multivariable analysis would be more accurate: the former describes analysis of multiple inputs (independent variables) and their association with multiple outputs (dependent variables), whereas the latter describes multiple inputs being used to describe a single output, for example, a multiple (thus multivariable) linear regression analysis (Hidalgo and Goodman, 2013). More positively, the majority of issues under this final theme could be addressed by simply truncating reported test statistics to avoid spurious accuracy, by always reporting appropriate confidence intervals for inferential statistics, and by following journal guidelines more closely on figure presentation.

CONCLUSIONS

Reproducibility and replicability are core to the philosophy of *JID Innovations*, and this underpinned the journal's decision to introduce a statistical review board. Fifteen months after the introduction of specialist reviewers, a number of common themes have emerged relating to statistical and data issues, many of which could have been addressed during the design of the study or the manuscript drafting stage by following the recommendations given under each thematic heading. Other

Table 1. Simple Errors for Authors to Avoid

Study design
Failure to identify important covariates or confounders
Sample sizes inappropriate to answer the research question robustly
Issues around unrepresentative recruitment
Failure to identify in advance how data will be shared with the wider community, for example, by repository or a general failure to observe FAIR data standards
Selection of an inappropriate control group (eg, sample type, tissue type, disease, assay)
Data preparation
Failure to investigate data distributions, especially testing for normality
Not matching transformations (log transformation, z-scaling, pareto scaling) to the assumptions made
Lack of documentation of approach to missing values or exclusions due to technical issues
Analysis and modeling
Inappropriate use of tests to compare groups at multiple time points (ie, using cross-sectional methods for longitudinal data)
Lack of separation of roles: Ideally, analyses should be conducted by a statistician and not by the investigator—failure to separate roles can reduce reproducibility.
Methods "overkill", especially using multiple methods and then selecting the one with greatest significance
Inappropriate use of machine learning, especially when a dataset is not big enough
Failure to revisit appropriateness of statistical tests given recruitment numbers actually achieved
Manuscript preparation
Reporting values to inappropriate levels of accuracy—no more than 3 decimal places—and use of exponential notation if more accuracy is needed (eg, genome studies with many comparisons)
Ensuring correct use of terminology, for example, distinguishing multivariate from multivariable analysis, failing to specify repeated measures ANOVA versus standard ANOVA, using interquartile range (when reporting the difference between the percentiles) versus 25th−75th percentiles (when reporting the percentile values themselves)
Not including a limitations paragraph
Not including a detailed statistical methods section, as part of methods and materials
Poor plot-type choices, for example, using line plots for categorical data
No labels on plot axes or labels that do not include units
Poor color selection in plots, for example, low contrast or colors that are inappropriate for colourblind readers
If using nonparametric methods, data should be presented using nonparametric methods (eg, median [Q1−Q3] or box-plots or dot plots [actual data points]). Similarly, parametric methods should be used to present and compare normally distributed data.
Legends should be comprehensive and allow for interpretation of figures without reference to the main text; as a minimum, legends should specify sample size, how data are presented (eg mean or median, SEM or SD, what are the components of a box whiskers plot), and how data are compared (eg, *t*-test or ANOVA). Symbols should be used for *P*-values in figures that represent $P < .05$, $P < .01$, and $P < .001$, and legends should also state the specific statistical test utilized

Abbreviations: FAIR, Findable, Accessible, Interoperable, and Reusable; Q1, quartile 1; Q3, quartile 3.

points, both major and minor, to be considered before submission to reduce the number of required revisions are set out for convenience in Table 1. The growth in accessibility of data science tools, whether online such as MetaboAnalyst or Reactome or offline in the form of data science libraries such as scikit-learn or TidyModels, has opened up data analysis to a wider audience, including clinicians, biologists, and other users. Nonetheless, we at *JID Innovations* continue to see the role of dedicated statisticians and data scientists as crucial in study design; analysis; manuscript submission; and, of course, peer review.

DATA AVAILABILITY STATEMENT

The data analyzed for this study comprised peer review comments provided to authors—by the statistical editors—and as such are considered confidential. Deidentified, aggregated data can be made available by contacting JH at jan.higgins@sidnet.org

KEYWORDS

Dermatology, Peer review, Skin science, Statistics

ORCIDs

Matt Spick: http://orcid.org/0000-0002-9417-6511
Jan Higgins: http://orcid.org/0000-0002-5342-382X
Cynthia L. Green: http://orcid.org/0000-0002-0186-5191
Roland Matsouaka: http://orcid.org/0000-0002-0271-5400
Daniel B. Shin: http://orcid.org/0000-0002-4974-2561
Russell P. Hall III: http://orcid.org/0000-0001-7621-4935
Nophar Geifman: http://orcid.org/0000-0003-2956-6676

CONFLICT OF INTEREST

NG, CLG, RM, and DBS are members of the statistical review board for *JID Innovations*. JH is the Managing Editor for *JID Innovations*. RPH is the Editor for *JID Innovations*. The remaining authors state no conflict of interest.

ACKNOWLEDGMENTS

The authors wish to acknowledge all peer reviewers and editorial staff at *JID Innovations*.

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE (AI) OR LARGE LANGUAGE MODELS (LLMS)

The author(s) did not use AI/LLM in any part of the research process and/or manuscript preparation.

REFERENCES

Beesley LJ, Salvatore M, Fritsche LG, Pandit A, Rao A, Brummett C, et al. The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. Stat Med 2020;39:773−800.

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK

Biobank participants with those of the general population. Am J Epidemiol 2017;186:1026–34.

Hall RP 3rd. Replication and reproducibility and the self-correction of science: what can JID innovations do? JID Innov 2023;3:100188.

Hidalgo B, Goodman M. Multivariate or multivariable regression? Am J Public Health 2013;103:39–40.

Ioannidis JP. Why most published research findings are false [published correction appears in PLoS Med 2022;19:e1004085]. PLoS Med 2005;2:e124.

Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol 2017;46:746–55.

McNutt M. Journals unite for reproducibility. Science 2014;346:679.

Saunders CH, Sierpe A, Von Plessen C, Kennedy AM, Leviton LC, Bernstein SL, et al. Practical thematic analysis: a guide for multidisciplinary health services research teams engaging in qualitative analysis. BMJ 2023;381:e074256.

Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. Nat Hum Behav 2023;7:1216–27.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship [published correction appears in Sci Data 2019;6:6]. Sci Data 2016;3:160018.