STATISTICS

Sex difference analyses under scrutiny

A survey reveals that many researchers do not use appropriate statistical analyses to evaluate sex differences in biomedical research.

COLBY J VORLAND

Related research article Garcia-Sifuentes Y, Maney DL. 2021. Reporting and misreporting of sex differences in the biological sciences. *eLife* **10**:e70817. doi: 10.7554/eLife.70817

Scientific research requires the use of appropriate methods and statistical analyses, otherwise results and interpretations can be flawed. How research outcomes differ by sex, for example, has historically been understudied, and only recently have policies been implemented to require such consideration in the design of a study (e.g., ***NIH, 2015***).

Over two decades ago, the renowned biomedical statistician Doug Altman labeled methodological weaknesses a "scandal", raising awareness of shortcomings related to the representativeness of research as well as inappropriate research designs and statistical analysis (***Altman, 1994***). These methodological weaknesses extend to research on sex differences: simply adding female cells, animals, or participants to experiments does not guarantee an improved understanding of this field of research. Rather, the experiments must also be correctly designed and analyzed appropriately to examine such differences. While guidance exists for proper analysis of sex differences, the frequency of errors in published research articles related to this topic has not been well understood (e.g., ***Beltz et al., 2019***).

Now, in eLife, Yesenia Garcia-Sifuentes and Donna Maney of Emory University fill this gap by surveying the literature to examine whether the statistical analyses used in different research articles are appropriate to support conclusions of sex differences (***Garcia-Sifuentes and Maney, 2021***). Drawing from a previous study that surveyed articles studying mammals from nine biological disciplines, Garcia-Sifuentes and Maney sampled 147 articles that included both males and females and performed an analysis by sex (***Woitowich et al., 2020***).

Over half of the articles surveyed (83, or 56%) reported a sex difference. Garcia-Sifuentes and Maney examined the statistical methods used to analyze sex differences and found that over a quarter (24 out of 83) of these articles did not perform or report a statistical analysis supporting the claim of a sex difference. A factorial design with sex as a factor is an appropriate way to examine sex differences in response to treatment, by giving each sex each treatment option (such as a treatment or control diet; see ***Figure 1A***). A slight majority of all articles (92, or 63%) used a factorial design. Within the articles using a factorial design, however, less than one third (27) applied and reported a method appropriate to test for sex differences (e.g., testing for an interaction between sex and the exposure, such as different diets; ***Figure 1B***). Similarly, within articles that used a factorial design and concluded a sex-specific effect, less than one third (16 out of 53) used an appropriate analysis.

Notably, nearly half of the articles (24 out of 53) that concluded a sex-specific effect statistically tested the effect of treatment within each sex and compared the resulting statistical significance. In other words, when one sex had a statistically significant change and the other did not,

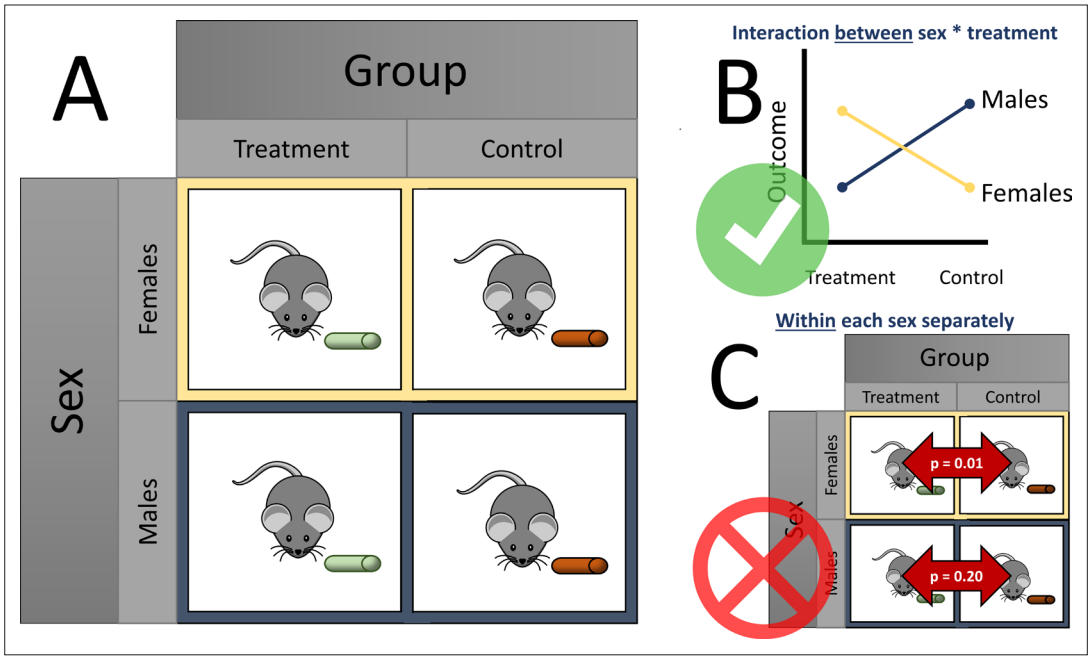

Figure 1. Considering sex differences in experimental design. (**A**) A so-called factorial design permits testing of sex differences. For example, both female (yellow boxes) and male mice (blue boxes) are fed either a treatment diet (green pellets) or control diet (orange pellets). Garcia-Sifuentes and Maney found that 63 % of articles employed a factorial design in at least one experiment with sex as a factor. (**B**) An appropriate way to statistically test for sex differences is with a two-way analysis of variance (ANOVA). If a statistically significant interaction is observed between sex and treatment, as shown in the figure, evidence for a sex difference is supported. Garcia-Sifuentes and Maney found that in studies using a factorial design, less than one third tested for an interaction between sex and treatment. (**C**) Performing a statistical test between the treatment and control groups within each sex, and comparing the nominal statistical significance, is not a valid method to look for sex differences. Yet, this method was used in nearly half of articles that used a factorial design and concluded a sex-specific effect.

the authors of the original studies concluded that a sex difference existed. This approach, which is sometimes called 'differences in nominal significance', or 'DINS' error (***George et al., 2016***), is invalid and has been found to occur for decades among several disciplines, including neuroscience (***Nieuwenhuis et al., 2011***), obesity and nutrition (***Bland and Altman, 2015***; ***George et al., 2016***; ***Vorland et al., 2021***), and more general areas (***Gelman and Stern, 2006***; ***Makin, 2019***; ***Matthews and Altman, 1996***; ***Sainani, 2010***; ***Figure 1C***).

This approach is invalid because testing within each sex separately inflates the probability of falsely concluding that a sex-specific effect is present compared to testing between them directly. Other inappropriate analyses that were identified in the survey included testing sex within treatment and ignoring control animals; not reporting results after claiming to do an appropriate analysis; or claiming an effect when the appropriate analysis was not statistically significant despite subscribing to 'null hypothesis significance' testing. Finally, when articles pooled the data of males and females together in their analysis, about half of them did not first test for a sex difference, potentially masking important differences.

The results of Garcia-Sifuentes and Maney highlight the need for thoughtful planning of study design, analysis, and communication to maximize our understanding and use of biological sex differences in practice. Although the survey does not quantify what proportion of this research comes to incorrect conclusions from using inappropriate statistical methods, which would require estimation procedures or reanalyzing the data, many of these studies' conclusions may change if they were analyzed correctly. Misleading results divert our attention and resources, contributing to the larger problem of 'waste' in biomedical research, that is, the avoidable costs of research that does not contribute to our understanding of what is true because it is flawed, methodologically weak, or not clearly communicated (***Glasziou and Chalmers, 2018***).

What can the scientific enterprise do about this problem? The survey suggests that there may be

a large variability in discipline-specific practices in the design, reporting, and analysis strategies to examine sex differences. Although larger surveys are needed to assess these more comprehensively, they may imply that education and support efforts could be targeted where they are most needed. Compelling scientists to publicly share their data can facilitate reanalysis when statistical errors are discovered – though the burden on researchers performing the reanalysis is not trivial. Partnering with statisticians in the design, analysis, and interpretation of research is perhaps the most effective means of prevention.

Scientific research often does not reflect the diversity of those who benefit from it. Even when it does, using methods that are inappropriate fails to support the progress toward equity. Surely this is nothing less than a scandal.

Colby J Vorland is at the Department of Applied Health Science, Indiana University School of Public Health, Bloomington, United States
cvorland@iu.edu
http://orcid.org/0000-0003-4225-372X

Competing interests: The author declares that no competing interests exist.

Published 02 November 2021

References

Altman DG. 1994. The scandal of poor medical research. *BMJ* **308**: 283–284. DOI: https://doi.org/10.1136/bmj.308.6924.283

Beltz AM, Beery AK, Becker JB. 2019. Analysis of sex differences in pre-clinical and clinical data sets. *Neuropsychopharmacology* **44**: 2155–2158. DOI: https://doi.org/10.1038/s41386-019-0524-3, PMID: 31527863

Bland JM, Altman DG. 2015. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *The American Journal of Clinical Nutrition* **102**: 991–994. DOI: https://doi.org/10.3945/ajcn.115.119768, PMID: 26354536

Garcia-Sifuentes Y, Maney DL. 2021. Reporting and misreporting of sex differences in the biological sciences. *eLife* **10**: e70817. DOI: https://doi.org/10.7554/eLife.70817

Gelman A, Stern H. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* **60**: 328–331. DOI: https://doi.org/10.1198/000313006X152649

George BJ, Beasley TM, Brown AW, Dawson J, Dimova R, Divers J, Goldsby TU, Heo M, Kaiser KA, Keith SW, Kim MY, Li P, Mehta T, Oakes JM, Skinner A, Stuart E, Allison DB. 2016. Common scientific and statistical errors in obesity research. *Obesity* **24**: 781–790. DOI: https://doi.org/10.1002/oby.21449, PMID: 27028280

Glasziou P, Chalmers I. 2018. Research waste is still a scandal – an essay by Paul Glasziou and Iain Chalmers. *BMJ* **363**: k4645. DOI: https://doi.org/10.1136/bmj.k4645

Makin TR. 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* **8**: e48175. DOI: https://doi.org/10.7554/eLife.48175

Matthews JN, Altman DG. 1996. Statistics notes. Interaction 2: Compare effect sizes not P values. *BMJ* **313**: 808. DOI: https://doi.org/10.1136/bmj.313.7060.808, PMID: 8842080

Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience* **14**: 1105–1107. DOI: https://doi.org/10.1038/nn.2886, PMID: 21878926

NIH. 2015. Consideration of Sex as a Biological Variable in NIH-funded Research. https://grants.nih.gov/grants/guide/notice-files/not-od-15-102.html [Accessed October 13, 2021].

Sainani K. 2010. Misleading comparisons: the fallacy of comparing statistical significance. *PM&R* **2**: 559–562. DOI: https://doi.org/10.1016/j.pmrj.2010.04.016

Vorland CJ, Brown AW, Dawson JA, Dickinson SL, Golzarri-Arroyo L, Hannon BA, Kahathuduwa CN. 2021. Errors in the implementation, analysis, and reporting of randomization within obesity and nutrition research: a guide to their avoidance. *Ternational Journal of Obesity* **45**: 2335–2346. DOI: https://doi.org/10.1038/s41366-021-00909-z

Woitowich NC, Beery A, Woodruff T. 2020. A 10-year follow-up study of sex inclusion in the biological sciences. *eLife* **9**: e56344. DOI: https://doi.org/10.7554/eLife.56344