




To adjust or not to adjust: it is not the tests performed that count, but how they are reported and interpreted

Anne-Laure Boulesteix,¹ Sabine Hoffmann ²

For numbered affiliations see end of article.

Correspondence to: Anne-Laure Boulesteix, Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich, Bayern, Germany; boulesteix@ibe.med.uni-muenchen.de

Cite this as: *BMJMED* 2024;3:e000783. doi:10.1136/bmjmed-2023-000783

Received: 12 October 2023

Accepted: 18 March 2024

There has been growing realisation that the reliability and credibility of research findings in medicine and in many other disciplines is severely impaired by the selective reporting of the most notable and significant results arising from a multiplicity of analyses. Indeed, performing a large number of tests will in many cases lead to findings that reflect nothing but lucky fluctuations. If you perform enough tests, you will obtain a P value that is smaller than the magical 0.05 threshold even if there are actually no real effects in your data, just in the same way as you will find someone who is taller than the 95th centile if you measure enough individuals from a representative sample.

A common solution to make valid statistical inference when multiple tests are performed is to use an adjustment procedure, such as Bonferroni, which aims to control the probability of obtaining at least one false positive result. Much has been said on how to adjust for multiple testing, but guidance on when to adjust is comparably scarce and often contradictory, which can lead to considerable uncertainty and confusion among researchers. In a few simple cases, there is general agreement on whether one should adjust for multiple testing or not. In many other cases, however, there are almost as many opinions given as there are tests to be performed. Researchers can find a vast spectrum of recommendations in the literature that range from clear theoretical arguments in favour of multiple testing¹ to forceful assertions against it.² For some, adjustment is "at best, unnecessary and, at worst, deleterious to sound statistical inference."³ Between these two extremes, you can find authors urging for adjustment only in explanatory settings but never in confirmatory settings, while you might find just as many authors urging for adjustment only in confirmatory settings but never in exploratory settings.

Contradictory recommendations and uncertainty on how to interpret them make the question of when to adjust for multiple testing a source of frustration in the communication between medical researchers and consulting statisticians. Both parties seem to be either convinced that they are more qualified to decide on whether and which analyses should be adjusted for or they are convinced that it is entirely the responsibility of the other party (or of the reviewers) to decide on this question. This uncertainty opens the door to questionable research practices, ranging from the omission of analyses that did not yield significant results to the salami slicing of research projects into multiple papers of low quality. Multiple testing issues are thus hidden behind fancy

terms such as "fishing for significance" and "cherry picking" that contribute to a distortion of the medical literature, undermining the paradigm of evidence based decision making.

Here, we suggest a unique approach that has the advantage of being easily understandable for both medical researchers and statisticians, thereby enabling efficient communication concerning the question of whether to adjust for multiple testing. It is not a complex set of rules, but one approach that is valid in all situations and includes many previously proposed rules as special cases:

Multiple testing should be adjusted for only where authors use the significance of statistical tests to weight the reporting, discussion, and interpretation of their findings.

The decision to adjust thus does not depend on a subjective label classifying the study as exploratory or confirmatory, nor does it rely on theoretical mathematical concepts. Instead, the decision depends on how the results of the multiplicity of performed tests are reported and interpreted. If significant results are used to hide the multiplicity of non-significant results in the interpretation of a study, adjusting for multiple testing would then be necessary. If, on the other hand, the results of all performed tests are transparently reported with equal emphasis, regardless of the significance of the findings, adjustment would not be necessary.

For example, if a clinical trial is declared successful when only one among three co-primary endpoints shows a significant difference between treatment and control group, this significant result is given more weight in the interpretation. According to our approach, the three corresponding tests form a set of tests over which adjustment is needed, which matches standard practice. On the other hand, the results of tests for secondary efficacy endpoints do not influence the reporting of primary endpoints and are usually reported on equal footing, no matter whether the respective null hypothesis was rejected or not. As a consequence, according to our approach, the tests on secondary efficacy endpoints do not form a set over which adjustment is necessary. In more general situations, the sets among which selective reporting dependent on significance occurs do not arise naturally from the research question but they are a deliberate decision in the reporting of results. Our approach can also support the choice for or against adjustment in these more complex situations. Finally, the approach is meaningful regardless which adjustment procedure (eg, Bonferroni or a more complex method) is used, if adjustment is chosen.

However, humans are not infallible.⁴ After results are known, it is often easy for authors to convince themselves that they were planning to report only the results of some of the performed tests in the first place (and these often happen to be significant). In this sense, the approach we propose is not a universal remedy, and will be most useful in combination with pre-registration. But confusion is much more likely if authors find that the rules to be applied are unreasonably strict or that the recommendations on whether to adjust for multiple testing or not are contradictory.

In this article, we have proposed a simple approach that aims to dispel this impression and to create a common understanding for medical researchers and statisticians alike, ultimately making the decision to adjust for multiple testing more transparent and less prone to bias. While this approach deals with misunderstandings in the decision on whether to adjust for multiple testing, it does not resolve other issues that arise when research conclusions are based on significance.^{5,6} In this sense, we do not mean to favour the sustained focus on P values in the medical literature and, in principle, the approach can also apply to other statistical inference approaches such as confidence intervals and Bayes factors, which have been increasingly put forward as meaningful alternatives to P values.⁵

AUTHOR AFFILIATIONS

¹Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich, Germany

²Institute for Statistics, LMU Munich, Munich, Germany

Acknowledgements We thank Anna Jacob and Savanna Ratky for language corrections.

Contributors ALB developed the concept of the approach and SH refined it. SH and ALB drafted the manuscript.

Funding This work was partly funded by individual grant BO3139/7 from the German Research Foundation to ALB.

Competing interests None declared.

Provenance and peer review Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Sabine Hoffmann <http://orcid.org/0000-0001-6197-8801>

REFERENCES

- 1 Sedgwick P. Multiple hypothesis testing and Bonferroni's correction. *BMJ* 2014;349:g6284. [10.1136/bmj.g6284](https://doi.org/10.1136/bmj.g6284)
- 2 Rubin M. When to adjust alpha during multiple testing: a consideration of Disjunction, conjunction, and individual testing. *Synthese* 2021;199:10969–1000. [10.1007/s11229-021-03276-4](https://doi.org/10.1007/s11229-021-03276-4)
- 3 Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;316:1236–8. [10.1136/bmj.316.7139.1236](https://doi.org/10.1136/bmj.316.7139.1236)
- 4 Nuzzo R. Fooling ourselves. *Nature* 2015;526:182–5. [10.1038/526182a](https://doi.org/10.1038/526182a)
- 5 Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond P < 0.05. *Am Stat* 2019;73:1–19.
- 6 Amrhein VS, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305–7. [10.1038/d41586-019-00857-9](https://doi.org/10.1038/d41586-019-00857-9)