ARTICLE

Is it time to stop using statistical significance?

Oliver Frank
Specialist general practitioner, Oakden Medical Centre, Hillcrest, South Australia

Senior research fellow, Discipline of General Practice, Adelaide Medical School, University of Adelaide

CW Michael Tam
Staff specialist, Primary and Integrated Care Unit, South Western Sydney Local Health District

Conjoint senior lecturer, School of Population Health, UNSW Sydney

Joel Rhee
Associate professor of General Practice, School of Medicine, University of Wollongong, New South Wales

Keywords
bias, confidence interval, statistical data analysis, statistics

Aust Prescr 2021;44:16–8
https://doi.org/10.18773/austprescr.2020.074

SUMMARY

The important first step in the critical appraisal of a randomised trial is not an evaluation of the statistical analyses. The most important aspect to consider when reviewing a study of a new drug is the appropriateness and quality of the trial design and methods.

The next most important aspect is the effect size of different treatments and its clinical significance. Rather than reporting statistical significance, studies should report the difference between treatments and its precision.

Over-reliance on statistical significance and p values may lead to incorrect conclusions. Trial reports about drugs should therefore avoid the term statistical significance and quote p values with caution.

Introduction

Criticisms of the misuse and misinterpretations of statistical significance testing (and of p values) were made throughout the last century.[1] William Rozeboom, an eminent philosopher of science, once asserted that it was 'surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students'.[2] This criticism reached a zenith in 2019, when the American Statistical Association, an international peak body of professional statisticians, formally recommended against statistical significance testing – both its use and in the reporting of results.[3]

There are many examples of how the term 'significant' can be open to interpretation. A review of fremanezumab for migraine in *Australian Prescriber*[4] stated:

> 'At the end of the trial, monthly injections had reduced the number of headache days by 4.6 days and the number of migraine days by 5.0 days. With quarterly injection the reductions were 4.3 days for headache and 4.9 days for migraine. Both regimens were significantly better than the reductions of 2.5 days and 3.2 days seen in the placebo group.'

For most readers of *Australian Prescriber*, that statement might seem eminently reasonable. However, the routine use of the word 'significantly' is misleading.[3]

Statistical significance

To understand why the term statistical significance is problematic, it is necessary to consider the context in which statistical significance testing occurs. Empirical research is about discovering and constructing knowledge about the world, for instance, whether a new drug works from the perspective of causation and predicting patient outcomes. This research often involves describing the empirical world using numbers (quantitative methods). Statistical inferential testing can be a useful tool whose results can inform us about the real world. However, discomfort with uncertainty promotes overconfidence in statistical rituals,[5] and contributes to the belief that statistical testing is always necessary.

Clinicians commonly misinterpret statistical significance and its conceptual twin, the p value.[6] This potentially results in gross overestimation of the strength of evidence.[7] Importantly, neither the validity of the study nor the truth of its findings can be inferred from p values and statistical significance alone.

Two simple heuristics to reduce misinterpretation of p values and statistical significance are:[6]

- They do not numerically refer to the probability of a phenomenon or event occurring in the real world. For instance, the claims that one or both show the likelihood of the experimental result being true, or due to chance, are incorrect.[8]
- They should not be interpreted using thresholds. Any cut-off value (such as p=0.05) is arbitrary. Making binary empirical conclusions based on which side of the threshold the test statistic falls is unsound reasoning.

Null hypothesis

Statistical significance is fundamentally a mathematical concept that should be understood only in the context of null hypothesis statistical testing. This involves creating a statistical model, a simplified and artificial 'mathematical world' where the researcher can define all the rules. In this model,

one of the rules is that drugs or procedures have zero effectiveness – hence the term null hypothesis.

Seen from within the mathematical world, using the assumptions of this 'zero-effectiveness' statistical model, the unusualness of the real-world data collected in the study can be calculated. The p value can be considered a measure of how compatible the data are with this statistical model. Larger p values are more compatible with the null hypothesis and small p values less so.

Statistical significance only means that the data reached an arbitrarily defined level of incompatibility with the statistical model. However, this zero-effectiveness statistical model might be incompatible with the data for many reasons. For instance, the data collected might have been biased, or one or more assumptions used in the statistical model were unsound or violated. Statistical significance does not indicate on its own that the result is true or that the null hypothesis is false. Moreover, statistical significance does not indicate or imply that a result is clinically important.

Clinical significance

Clinical significance pertains to patient care. Deciding whether or not a study result is clinically significant cannot be determined by an algorithm. Rather it requires judgement, clinical expertise and a respect for context.

The important first step in the critical appraisal of a clinical trial is not an evaluation of the statistical analyses. Analysing the patients, intervention, comparison and outcomes in the methods section of the report, and being satisfied with the reasonableness of the question asked by the researchers, is important in deciding whether or not to read more of the report.

Next is an appraisal of the internal validity of the trial, which can be framed as a series of questions. For a randomised trial:[9]

- was the assignment of patients to treatments randomised?
- were the groups similar at the start of the trial?
- aside from the allocated treatment, were the groups treated equally?
- were all patients who entered the trial accounted for?
- were measures objective and were the patients and clinicians kept blind to which treatment was being received?

Threats to the internal validity of a study's methodology reduce the confidence that the results usefully represent what the study sought to investigate. Simply, if the study has major methodological biases, the results will need to be taken with a grain of salt. The results might even be uninterpretable.

Effect size

When looking at trial results, the focus should be on the primary outcome, its effect size, and the precision with which that effect has been able to be estimated. This precision is often described as a confidence interval. If the differences in outcomes between groups are small, there is likely to be little clinical benefit from using a trial treatment instead of a comparator. However, it is important to remember that the reported effect size is the average for the sample of people in the study and it is likely that many participants (half of the sample, assuming normal distribution) benefited more while others benefited less (again half, assuming normal distribution). Whether an effect size is clinically significant depends on the nature of the condition, the effect and the context. Synthesising these together requires clinical judgement. Fortunately, investigators often include a discussion of clinical significance when describing the power and sample size calculations in the methods section of their reports.

A useful concept to consider is the minimum clinically important difference, especially when there may not be a good intuitive grasp of the outcome measure. For example, the six-item headache impact test (HIT-6) has a range from 36 (no impact) to 78 (very severe). The minimum clinically important difference is considered to be 2.5 points.[10] In the trial described in *Australian Prescriber*, fremanezumab reduced the HIT-6 score compared with placebo by 1.9 when given quarterly and by 2.4 when given monthly.[11] Both changes are statistically significant, but are less than the minimum clinically important difference. It is important to note that only about 20% of participants in the trial were using any migraine-preventing medicine. When balancing the modest average therapeutic effect of fremanezumab with the need for it to be injected and its high cost compared to established drugs for migraine prophylaxis, it seems hard to justify it as a first-line treatment.

Confidence intervals

The confidence interval, typically reported at 95%, can be interpreted as the (im)precision of the effect-size estimate. This is the range of values that are mathematically compatible with the effect-size estimate.

If the confidence interval is wide, the lower and upper limits indicate very different clinical effects ranging from a tiny effect size to a substantial effect. The effect-size estimate is therefore imprecise and

it would be misleading for it to be quoted without caution and appropriate context.

If the confidence interval is subjectively narrow, the lower and upper limits would give roughly the same clinical interpretation. It could then be claimed that the estimate of effect size is precise.

Judgement and care are required regardless of the confidence interval. A large drug trial undertaken in men could conceivably yield a very precise effect-size estimate, that would be incorrect in women.

It is time to stop using statistical significance

As an exercise to develop insight, try replacing instances of the term statistically significant with the synonym 'mathematically unusual'. Paraphrasing the original quoted *Australian Prescriber* new drug comment as 'both regimens were [statistically] significantly better than the placebo group' becomes 'both regimens were mathematically unusually better than the placebo group'. The apparent meaninglessness of the second sentence is what is meant by the first.

The hidden absurdity of commonly seen statements in reports such as 'the results approached [statistical] significance' is revealed when they are transformed into 'the results approached mathematical unusualness'.

Conclusion

Significance is still a useful word that should not be abandoned. However, for too long statistical significance has co-opted the use of the word. The medical literature commonly conflates statistical significance with the everyday meaning of significance. In line with the recommendation of the American Statistical Association, it is time to move on. Its executive director wrote in unambiguous terms 'statistically significant – don't say it and don't use it'.[3] Rather, we should focus on the effect-size estimate and its precision and interpret these through the lens of clinical significance. <

Conflicts of interest: none declared

REFERENCES

1. Wasserstein RL, Lazar NA. The ASA's Statement on p-values: Context, process, and purpose. Am Stat 2016;70:129-33. https://doi.org/10.1080/00031305.2016.1154108
2. Siegfried T. P value ban: small step for a journal, giant leap for science: ScienceNews. 2015 Mar 17. https://www.sciencenews.org/blog/context/p-value-ban-small-step-journal-giant-leap-science [cited 2021 Jan 4]
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p<0.05". Am Stat 2019;73 supp1:1-19. https://doi.org/10.1080/00031305.2019.1583913
4. Fremanezumab for migraine. Aust Prescr 2020;43:68-9. https://doi.org/10.18773/austprescr.2020.016
5. Gigerenzer G. Statistical rituals: the replication delusion and how we got there. Adv Methods Pract Psychol Sci 2018;1:198-218. https://doi.org/10.1177/2515245918771329
6. Tam CW, Khan AH, Knight A, Rhee J, Price K, McLean K. How doctors conceptualise P values: a mixed methods study. Aust J Gen Pract 2018;47:705-10. https://doi.org/10.31128/AJGP-02-18-4502
7. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337-50. https://doi.org/10.1007/s10654-016-0149-3
8. Tam CW. How we (mis)conceptualise p-values (and what we can do about it). 2018 Oct 15. https://www.youtube.com/watch?v=kK6I1KCb7jQ [cited 2021 Jan 4]
9. Centre for Evidence-Based Medicine. Critical appraisal tools. Oxford, UK: CEBM; 2014. http://www.cebm.ox.ac.uk/resources/ebm-tools/critical-appraisal-tools [cited 2021 Jan 4]
10. Smelt AF, Assendelft WJ, Terwee CB, Ferrari MD, Blom JW. What is a clinically relevant change on the HIT-6 questionnaire? An estimation in a primary-care population of migraine patients. Cephalalgia 2014;34:29-36. https://doi.org/10.1177/0333102413497599
11. Silberstein SD, Dodick DW, Bigal ME, Yeung PP, Goadsby PJ, Blankenbiller T, et al. Fremanezumab for the preventive treatment of chronic migraine. N Engl J Med 2017;377:2113-22. https://doi.org/10.1056/NEJMoa1709038

FURTHER READING

Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305-7. https://doi.org/10.1038/d41586-019-00857-9

Greenhalgh T. How to read a paper: the basics of evidence-based medicine and healthcare. 6th ed: Wiley; 2019.

Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124. https://doi.org/10.1371/journal.pmed.0020124

Nuzzo R. Scientific method: statistical errors. Nature 2014;506:150-2. https://doi.org/10.1038/506150a