Do We Give Too Much Significance to Statistical Significance?

Lauren Bresee

If you think back to your introductory statistics course during university, you may recall the topic of hypothesis testing, where the null hypothesis states there is no difference between study groups, and the alternative hypothesis states there is a difference between study groups.[1] Now, as a practising pharmacist, you are conducting a randomized controlled trial (RCT) to evaluate whether an intervention you developed improves patients' adherence to their medication therapy after discharge. Study participants will be randomly assigned to receive your adherence intervention or usual care, and adherence at 90 days after discharge from hospital is your primary study outcome. Your null hypothesis is that there is no difference in adherence at 90 days after discharge between the patients who receive your intervention and the patients who receive usual care, and your alternative hypothesis states that there is a difference in adherence between your intervention and control groups. To evaluate your primary outcome, you decide to perform inferential statistical testing to determine whether you will accept or reject your null hypothesis. On the basis of what is commonly reported in the medical literature, you have decided to use a threshold probability (p) value of 0.05 to determine whether your intervention group is statistically different from your control group; that is, if the p value associated with your statistical test is less than 0.05, you will reject the null hypothesis and conclude that the difference between your intervention and control groups is statistically significant.

The concept of statistical significance and the use of a threshold p value (and corresponding 95% confidence intervals) to determine statistical significance have long been sources of controversy in the research community. If you're a statistics nerd like me, you may have noticed a pair of recent publications regarding the use of statistical significance in research. In the first article, an editorial published in March 2019 and entitled "Scientists rise up against statistical significance", Amrhein and others[2] called for no longer using statistical significance to determine whether there is a difference between groups, because the concept of significance is frequently applied dichotomously, instead of being evaluated on a continuum. The authors stated, "Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero."[2] Instead, the authors suggested using confidence intervals as "compatibility intervals", that is, your point estimate and confidence interval are the most compatible with your data, given the statistical model you have used to calculate your results.[2] Going back to the RCT described above, you conduct your inferential statistical test, and your result is a relative risk of 2.0, with a 95% confidence interval of 1.5–2.5 and a p value far below 0.001. Under traditional statistical test reporting, you would conclude that the people in your intervention group were twice as likely as your control group to be adherent to their medication therapy at 90 days after discharge, and that this difference is statistically significant because the confidence interval does not encompass the measure of equivalence of 1. However, if you were to use the proposal set out by Amrhein and others,[2] you would instead state that the values for relative risk, 95% confidence interval, and p value most compatible with your data indicate that people who received your intervention were twice as likely to be adherent to their medication 90 days after hospital discharge, and that the risk difference between your treatment and control groups ranged from 1.5 times more likely to 2.5 times more likely to be adherent, given the assumptions of the statistical testing.

In rebuttal to the editorial by Amrhein and others,[2] Ioannidis published an editorial the following month in *JAMA*, entitled "The importance of predefined rules and prespecified statistical analyses: do not abandon significance".[3] In his editorial, Ioannidis emphasized that decisions made in medicine are most often dichotomous, and that more focus is being put on inappropriate claims of finding no statistical difference than on addressing "unwarranted claims of 'difference' and unwarranted denial of refutation", particularly when prespecified rules of statistical testing are not developed or followed by researchers.[3] Instead of banning the concept of statistical significance, Ioannidis emphasized that researchers must focus on both following the rules of statistical testing and ensuring that clinical relevance is applied to decision-making.[3] It is clear that underlying both of these editorials is the concern that statistical significance testing and *p* values are often used inappropriately.

What is a *p* value, and what is it not? The American Statistical Association (ASA) defines a *p* value as "the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."[4] The key items on which to focus in this definition are the specified statistical model and the data that are used in the statistical test. If we use a *p* value threshold of 0.05 to determine statistical significance, this means that the probability a given result is due to chance is less than 5%, specific to the statistical model and the data used for the test. In an effort to reduce the likelihood of misinterpretation of *p* values, the ASA board of directors published a statement on how *p* values should and should not be used.[4] Although *p* values can be used to determine whether there is evidence against the null hypothesis, as mentioned above, such conclusions are specifically applicable to the data used and the assumptions made to calculate the *p* value.[4] The *p* value does not describe the strength of the effect size or the precision of your result (that is, a smaller *p* value does not reflect a larger effect size or a more precise estimate), nor does it represent the probability that the overall study hypothesis is true or due to chance when applied to the population of interest.[4]

There will likely always be controversy associated with statistical significance and the use of *p* values. There are, however, fundamental consistencies within the editorials by Amrhein and others[2] and Ioannidis[3] and the ASA statement[4] that, if adhered to, will help to minimize the controversy. First, we must ensure that the statistical plan for each study is prespecified, transparent, and applicable to the study data.[2-4] Second, the results of all statistical tests conducted, including point estimates, measures of precision such as confidence intervals, and *p* values, must be reported. Such reporting ensures that we are not selectively reporting certain outcomes and allows for the evaluation of the possibility of type 1 error, that is, finding a statistically significant result where none actually exists, due to a multiplicity of statistical tests.[2-4] Third, we must be realistic when interpreting the results of any study and must avoid over- or under-emphasizing the study results.[2-4] Lastly, clinical decisions should never be based on statistical results alone. We must take into consideration other factors, including the study's validity, the consistency of the study's results with other available information, and the generalizability of the results to the overall population under consideration.[2-4] Following these recommendations will help to ensure the appropriate use of statistical testing in research, so that we can make the best possible clinical decisions for our patients.

References

1. Gaddis GM, Gaddis ML. Introduction to biostatistics: part 3, sensitivity, specificity, predictive value, and hypothesis testing. *Ann Emerg Med.* 1990; 19(5):591-7.
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance [editorial]. *Nature*. 2019;567(7748):305-7.
3. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: do not abandon significance [editorial]. *JAMA*. 2019;321(21): 2067-8.
4. Wasserstein RL. ASA statement on statistical significance and *P*-values. *Am Stat.* 2016;70(2):131-3.

Lauren Bresee, BScPharm, ACPR, MSc, PhD, is a Scientific Advisor with the Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, Ontario; an Adjunct Assistant Professor with the Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta; and a member of the O'Brien Institute for Public Health, University of Calgary. She is also an Associate Editor with the *Canadian Journal of Hospital Pharmacy*.

Competing interests: None declared.

Address correspondence to:
Dr Lauren Bresee
Canadian Agency for Drugs and Technologies in Health
865 Carling Avenue, Suite 600
Ottawa ON K1S 5S8

e-mail: LaurenB@cadth.ca