P-values

From: Key Concepts for assessing claims about treatment effects and making well-informed treatment choices (Version 2022)

2.4c Be cautious of p-values.

Explanation

The observed difference in outcomes is the best estimate of how relatively effective and safe treatments are (or would be, if the comparison were made in many more people). However, because of the play of chance, the true difference may be larger or smaller than this. The confidence interval is the range within which the true difference is likely to lie, after considering the play of chance. Although a confidence interval (margin of error) is more informative than a p-value, often only the latter is reported. P-values are often misinterpreted to mean that treatments have or do not have important effects.

For example, George Siontis and John Ioannidis reviewed 51 articles that reported "statistically significant tiny effects" published in four high profile journals *[Siontis 2011 (SR)]*. Even minimal bias in those studies could explain the observed "effects". Yet, more than half (28) of the articles did not express any concern about the size or uncertainty of the estimate of the observed effect. Despite the low p-values reported in these articles, the results often excluded effects that would be large enough to be important. Interpretation of small effects based on p-values alone is likely to be misleading.

Basis for this concept

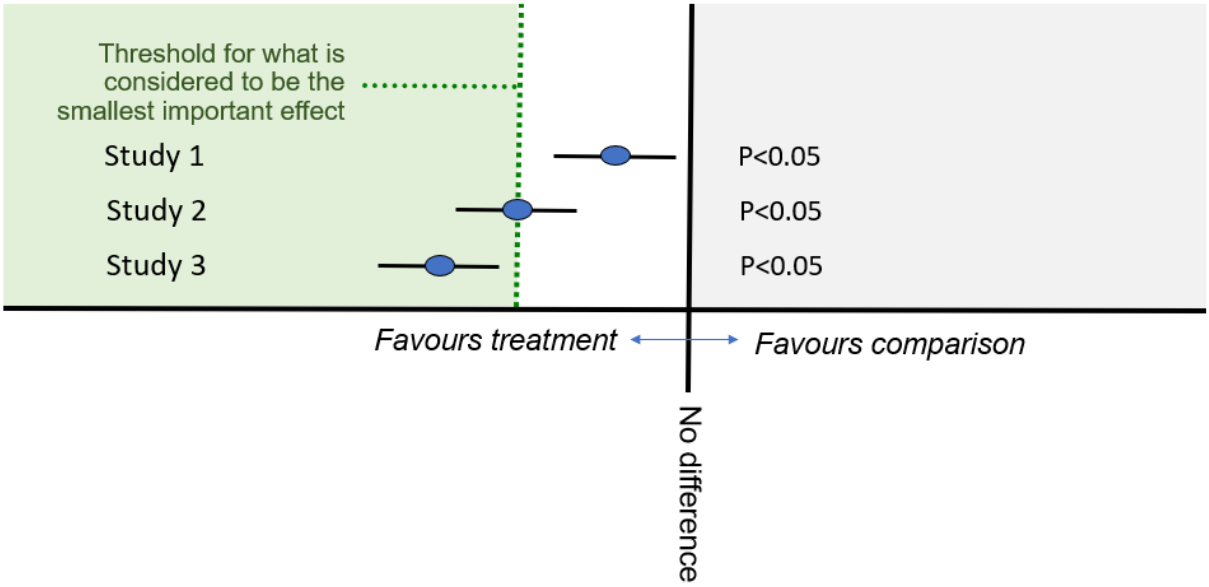
P-values, or "significance" levels, measure the probability of observing a result as extreme or more extreme than the actual result, simply by chance, if, in reality, there is no treatment difference. The smaller the p-value the less likely it is that there is no difference. Hundreds of warnings of the limitations of p-values and significance testing have been published *[Stang 2017 (SR)]*. From the 1970s to 2014, the proportion of abstracts (summaries of studies) with significance testing without any confidence intervals decreased from close to 100% to below 25%. However, the proportion of abstracts reporting only confidence intervals (and not p-values) in the top medical journals was only 22%. Another systematic review of abstracts indexed in MEDLINE found that more abstracts and articles reported p-values over time between 1990 and 2015 *[Chavalarias 2016 (SR)]*. Almost all abstracts and articles with p-values reported "statistically significant" results. Confidence intervals were only reported in about 2% of abstracts.

Despite all the warnings about p-values and significance testing, use and misinterpretation of p-values continues to be a problem. In 2016, the American Statistical Association released a policy statement on statistical significance and p-values, which included this warning *[Wasserstein 2016]*: ''The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.''

A systematic review of abstracts describing the results of cancer randomized trials with p-values between 0.01 and 0.10 found that trials commonly failed to convey uncertainty when describing results of "marginal statistical significance" *[Rubinstein 2019 (SR)]*. The results were often conveyed as definitively demonstrating that the null hypothesis (no difference) was false. This is likely associated with a discrete threshold for "statistical significance" (generally 0.05).

Another systematic review of surgical randomized trials found that outcomes reported in the abstract had three times the odds of being "statistically significant" compared to the corresponding full text *[Assem 2017 (SR)]*. Biased reporting of outcomes in abstracts based on p-values being below an arbitrary threshold has been found in other studies *[Boutron 2010 (SR), Chavalarias 2016 (SR), Ginsel 2015 (SR), Gøtzsche 2006 (SR)]*. This problem is like problems with publication bias and selective outcome reporting (see Concept 2.2b).

P-values can be misinterpreted in several ways *[Goodman 2008 , Greenland 2016]*. Perhaps most importantly, "statistical significance" may be confused with importance, and the cut-off for considering a result as statistically significant (generally $p \leq 0.05$) is arbitrary (see Concept 2.4d). People often assume that a low p-value indicates an important effect. However, a low p-value may or may not indicate an important effect, as illustrated in the figure below. All three studies have a p-value less than 0.05, indicating that it is unlikely that the observed treatment difference could have occurred simply by chance. But Study 1 indicates that it is unlikely that the difference was important, Study 2 indicates it is uncertain whether there was an important difference, and Study 3 indicates it is likely there was an important difference.



The blue dots in the figure above indicate the observed treatment effect and the horizontal lines indicate the confidence interval for each effect estimate. The figure illustrates why confidence intervals are more informative than p-values, as well as why the results of treatment comparisons should be interpreted in relation to thresholds for what is considered to be an important effect, not in relation to no difference.

Another problem with p-values is that people may assume that a p-value is the probability that there is no treatment difference and that a high p-value indicates a high probability that there is not a difference *[Sterne 2001]*. However, p-values indicate the probability of a "type I error" (assuming there is a difference when in fact there is not). They do not indicate the probability of a "type II error" (assuming there is not a difference when in fact there is). Many studies are too small to rule out an important difference (see Concept 2.3d).

Furthermore, people may assume that a low p-value indicates the likelihood that the observed treatment effect is the "true" effect. However, p-values only indicate the probability of wrongly assuming there is a difference when the observed difference could have occurred simply by chance

("random error"). It does not indicate anything about the risk of bias (systematic errors) because of how studies are designed, analysed, or reported (see Concepts 2.1a-2.1g).

P-values are used for testing the "null hypothesis" (that there is not a difference). A low p-value indicates that the null hypothesis can be rejected, with respect to random error. But hypothesis testing is unhelpful for people deciding whether to use a treatment. Hypothesis testing implies that there is a simple yes or no answer (there is or is not an effect) and it does not convey any information about the size of the effect *[Gardner 1986]*. Estimation of the size of the effect – for example, how big the difference is – and the confidence interval for that estimate is much more informative and less likely to mislead people.

However, it should be noted that confidence intervals are also sometimes misinterpreted *[Greenland 2016]*. In addition, 95 % confidence intervals correspond to a 0.05 cut-off for p-values. Thus, they have some of the same shortcomings as p-values. Nonetheless, confidence intervals are preferable to "significance" tests and p-values because they shift the focus away from the null hypothesis, toward the range of effect estimates compatible with the data. Provided they are interpreted carefully, they can also shift the focus from any difference greater than zero to effects that are large enough to be important *[Zeng 2021]*.

Implications

Understanding a confidence interval may be necessary to understand the reliability of estimates of treatment effects. Whenever possible, consider confidence intervals when assessing estimates of treatment effects. Do not be misled by p-values.

References

Systematic reviews

Assem Y, Adie S, Tang J, Harris IA. The over-representation of significant p values in abstracts compared to corresponding full texts: A systematic review of surgical randomized trials. Contemp Clin Trials Commun. 2017;7:194-9. https://doi.org/10.1016/j.conctc.2017.07.007

Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303(20):2058-64. https://doi.org/10.1001/jama.2010.651

Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016;315(11):1141-8. https://doi.org/10.1001/jama.2016.1952

Ginsel B, Aggarwal A, Xuan W, Harris I. The distribution of probability values in medical abstracts: an observational study. BMC Res Notes. 2015;8:721. https://doi.org/10.1186/s13104-015-1691-x

Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ. 2006;333(7561):231-4. https://doi.org/10.1136/bmj.38895.410451.79

Rubinstein SM, Sigworth EA, Etemad S, Martin RL, Chen Q, Warner JL. Indication of measures of uncertainty for statistical significance in abstracts of published oncology trials: a systematic review and meta-analysis. JAMA Netw Open. 2019;2(12):e1917530. https://doi.org/10.1001/jamanetworkopen.2019.17530

Siontis GC, Ioannidis JP. Risk factors and interventions with statistically significant tiny effects. Int J Epidemiol. 2011;40(5):1292-307. https://doi.org/10.1093/ije/dyr099

Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. Eur J Epidemiol. 2017;32(1):21-9. https://doi.org/10.1007/s10654-016-0211-1

Other references

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;292(6522):746-50. https://doi.org/10.1136/bmj.292.6522.746

Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45(3):135-40. https://doi.org/10.1053/j.seminhematol.2008.04.003

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(4):337-50. https://doi.org/10.1007/s10654-016-0149-3

Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ. 2001;322(7280):226-31. https://doi.org/10.1093/ptj/81.8.1464

Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129-33. https://doi.org/10.1080/00031305.2016.1154108

Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol. 2021;137:163-75. https://doi.org/10.1016/j.jclinepi.2021.03.026