OPEN ACCESS

On reporting and interpreting statistical significance and p values in medical research

Herman Aguinis,[1] Matt Vassar,[2] Cole Wayant [2]

10.1136/bmjebm-2019-111264

[1]Management, The George Washington University, Washington, District of Columbia, USA
[2]Psychiatry and Behavioral Sciences, Oklahoma State University Center for Health Sciences, Tulsa, Oklahoma, USA

Correspondence to:
Cole Wayant, Oklahoma State University Center for Health Sciences, Tulsa, OK 74107, USA; cole.wayant@okstate.edu

To cite: Aguinis H, Vassar M, Wayant C. *BMJ Evidence-Based Medicine* 2021;**26**:39–42.

Recent proposals to change the p value threshold from 0.05 to 0.005 or to retire statistical significance altogether have garnered much criticism and debate.[1 2] As of the writing of our manuscript, the proposal to eliminate statistical significance testing, backed by over 800 signatories, achieved record-breaking status on Altmetrics, with an attention score exceeding 13 000 derived from 19 000 Twitter comments and 35 news stories. We appreciate the renewed enthusiasm for tackling important issues related to the analysis, reporting and interpretation of scientific research results. Our perspective, however, focuses on the current use and reporting of statistical significance and where we should go from here.

1. We begin by saying that p values themselves are not flawed. Rather, the use, misuse or abuse of p values in ways antithetical to rigorous scientific pursuits is the flaw. If p values are a hammer, scientists are the hammer wielders. One would not discard the hammer if the wielder, when using the hammer, repeatedly missed the nail. Similarly, one would not discard the hammer if the wielder used the hammer in a way not suited to the hammer's purpose, such as in an attempt to drive a screw. Rather, one would expect that the fault lies with the hammer-wielder and recommend ways to refine the hammer's use. Thus, a focus on education and reform may be more helpful than the abandonment of statistical significance testing, which is a tool that can be used well, or misused and even abused.
2. Similarly, in this perspective, we argue that abandoning statistical significance because scientists misuse p values does not address the underlying problems of statistical negligence. Similarly, it does not address the incorrect belief that statistical significance equates to clinical significance.[3]

The a priori level (ie, alpha or type I error rate) and the precisely observed probability values (ie, p) should be explicitly stated and justified in protocols and published reports of medical studies. We have examined current guidance on p value reporting in influential sources in medicine (table 1). Generally, this guidance supports reporting exact p values but fails to issue direction on specifying the a priori significance level. The 'conventional' a priori significance (ie, type I error) level in many scientific disciplines is 0.05—an arbitrary choice. Two issues arise when scientists arbitrarily default to an a priori significance level: results become misleading and the relative seriousness of making a type I ('false-positive') or type II error ('false-negative') is ignored.

First, misleading results may fall on either side of the conventional 0.05 threshold, with scientists either rejecting or accepting the null hypothesis blindly—failing to consider sample size, measurement error and other factors that affect observed p values but are unrelated to the size of the effect in the population. Also, when considering the dichotomous interpretation of a truly continuous probability, Rosnow and Rosenthal[4] sarcastically lamented that 'Surely, God loves the 0.06 nearly as much as the 0.05'. Second, the choice of an a priori significance level should be made in the context of the potential for type II error. When researchers arbitrarily default to a type I error rate of 0.05, it has been calculated that the corresponding type II error is approximately 60%, because statistical power (ie, probability to correctly reject a null hypothesis) is usually insufficient given small sample sizes and the pervasive and unavoidable use of less-than-perfectly reliable measures.[5 6] In other words, while authors focus on whether their results show an acceptably small type I error rate, type II error—the probability of accepting the null hypothesis erroneously and incorrectly concluding that an effect is absent—looms large. Do authors, peer reviewers, editors and readers of studies that fail to reach statistical significance consider the probability that the results are falsely 'negative'?

A second limitation in the current guidance is the inconsistency in mandating effect size reporting that describes the strength of the relationship and/or the effect found. The only information to be gleaned from p values is whether the observed data are likely where the null hypothesis (that no effect exists) true. Therefore, a p value without an effect size is like peering into a pool of murky water: one cannot determine the depth, just say that it is likely that a pool exists. Consider interventions for improving medication adherence for patients with hypertension. A recent systematic review of medication adherence interventions found that the overall standardised mean difference for systolic blood pressure was 0.235—a 3 mm Hg difference.[7] Translating mean differences to clinical differences assists in determining the practical value of the intervention. In this example, the clinician must consider whether a 3 mm Hg reduction in systolic blood pressure is clinically meaningful and weigh this reduction against the factors associated with enacting the intervention as well as whether other interventions might yield a more clinically meaningful improvement. Some of the influential guidance (or omission thereof) provided to authors in medicine (table 1) may serve to promote the poor

Table 1 Guidance on p value, alpha prespecification and effect size reporting from influential sources in medicine

Source	Verbatim statement on p value reporting	Verbatim statement on alpha specification	Verbatim statement on effect size reporting
New England Journal of Medicine[8]	Unless one-sided tests are required by study design, such as in non-inferiority clinical trials, all reported p values should be two-sided. In general, p values larger than 0.01 should be reported to two decimal places, and those between 0.01 and 0.001 to three decimal places; p values smaller than 0.001 should be reported as p<0.001. Notable exceptions to this policy include p values arising from tests associated with stopping rules in clinical trials or from genome-wide association studies. When comparing outcomes in two or more groups in confirmatory analyses, investigators should use the testing procedures specified in the protocol and SAP to control the overall type I error—for example, Bonferroni adjustments or prespecified hierarchical procedures. P values adjusted for multiplicity should be reported when appropriate and labelled as such in the manuscript. In hierarchical testing procedures, p values should be reported only until the last comparison for which the p value was statistically significant. P values for the first non-significant comparison and for all comparisons thereafter should not be reported. For prespecified exploratory analyses, investigators should use methods for controlling the false discovery rate described in the SAP—for example, Benjamini-Hochberg procedures. When no method to adjust for multiplicity of inferences or controlling false discovery rate was specified in the protocol or SAP of a clinical trial, the report of all secondary and exploratory endpoints should be limited to point estimates of treatment effects with 95% CIs. In such cases, the Methods section should note that the widths of the intervals have not been adjusted for multiplicity and that the inferences drawn may not be reproducible. No p values should be reported for these analyses. Therefore, in most cases, no p values for interaction should be provided in the forest plots. If significance tests of safety outcomes (when not primary outcomes) are reported along with the treatment-specific estimates, no adjustment for multiplicity is necessary. Because information contained in the safety endpoints may signal problems within specific organ classes, the editors believe that the type I error rates larger than 0.05 are acceptable. Editors may request that p values be reported for comparisons of the frequency of adverse events among treatment groups, regardless of whether such comparisons were prespecified in the SAP. When appropriate, observational studies should use prespecified accepted methods for controlling family-wise error rate or false discovery rate when multiple tests are conducted. In manuscripts reporting observational studies without a prespecified method for error control, summary statistics should be limited to point estimates and 95% CIs. In such cases, the Methods section should note that the widths of the intervals have not been adjusted for multiplicity and that the inferences drawn from the inferences may not be reproducible. No p values should be reported for these analyses.	When comparing outcomes in two or more groups in confirmatory analyses, investigators should use the testing procedures specified in the protocol and SAP to control the overall type I error—for example, Bonferroni adjustments or prespecified hierarchical procedures. Because information contained in the safety endpoints may signal problems within specific organ classes, the editors believe that the type I error rates larger than 0.05 are acceptable.	Significance tests should be accompanied by CIs for estimated effect sizes, measures of association or other parameters of interest. The CIs should be adjusted to match any adjustment made to significance levels in the corresponding test.
Journal of the American Medical Association[9]	Avoid solely reporting the results of statistical hypothesis testing, such as p values, which fail to convey important quantitative information. For most studies, p values should follow the reporting of comparisons of absolute numbers or rates and measures of uncertainty (eg, 0.8%, 95% CI −0.2% to 1.8%; p=0.13). P values should never be presented alone without the data that are being compared. If p values are reported, follow standard conventions for decimal places: for p values less than 0.001, report as 'p<0.001'; for p values between 0.001 and 0.01, report the value to the nearest thousandth; for p values greater than or equal to 0.01, report the value to the nearest hundredth; and for p values greater than 0.99, report as 'p>0.99'. For studies with exponentially small p values (eg, genetic association studies), p values may be reported with exponents (eg, $p=1\times10^{-5}$). In general, there is no need to present the values of test statistics (eg, F statistics or χ^2 results) and df when reporting results.	No guidance	Meta-analyses should state the major outcomes that were pooled and include ORs or effect sizes.
The Lancet[10]	P values should be given to two significant figures, unless p<0.0001.	No guidance	No guidance
BMJ	No guidance; refers readers to SAMPL[11]	No guidance	No guidance; refers readers to SAMPL
Annals of Internal Medicine[12]	For p values between 0.001 and 0.20, please report the value to the nearest thousandth. For p values greater than 0.20, please report the value to the nearest hundredth. For p values less than 0.001, report as 'p<0.001'.	No guidance	Authors should report results for meaningful metrics rather than reporting raw results. For example, rather than reporting the log OR from a logistic regression, authors should transform coefficients into the appropriate measure of effect size, OR, relative risk or risk difference.

Continued

Table 1 Continued

Source	Verbatim statement on p value reporting	Verbatim statement on alpha specification	Verbatim statement on effect size reporting
ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9[13]	When reporting the results of significance tests, precise p values (eg, p=0.034) should be reported rather than making exclusive reference to critical values.	Conventionally, the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. Alternative values to the conventional levels of type I and type II errors may be acceptable or even preferable in some cases.	No guidance
SAMPL guideline[11]	Although not preferred to CIs, if desired, p values should be reported as equalities when possible and to one or two decimal places (eg, p=0.03 or 0.22 not as inequalities: eg, p<0.05). Do NOT report 'NS'; give the actual p value. The smallest p value that needs to be reported is p<0.001, save in studies of genetic associations.	Report the alpha level (eg, 0.05) that defines statistical significance.	Likewise, p values are not sufficient for re-analysis. Needed instead are descriptive statistics for the variables being compared, including sample size of the groups involved, the estimate (or 'effect size') associated with the p value and a measure of precision for the estimate, usually a 95% CI.
CONSORT statement[14]	Actual p values (eg, p=0.003) are strongly preferable to imprecise threshold reports, such as p<0.05.	No guidance	n—For each outcome, study results should be reported as a summary of the outcome in each group (eg, the number of participants with or without the event and the denominators, or the mean and SD of measurements), together with the contrast between the groups, known as the effect size.

SAP: statistical analysis plan; SAMPL: Statistical Analyses and Methods in the Published Literature; ICH: International Council for Harmonisation; CONSORT: CONsolidated Standards for Reporting Of Trials

statistical practices that readers work to mitigate. Therefore, it is our perspective that not only should all guidance emphasise reporting effect sizes, but that all guidance to interpret and report effect sizes in a meaningful way should be included as well. For example, one may report the absolute difference between groups and the number needed to treat for a medical intervention. Readers may be incapable of determining the meaningfulness of a p value but are well-equipped to interpret an absolute difference in effectiveness.

Taken together, reporting (1) precise observed p values (rather than whether it is larger or smaller than arbitrary cutoffs), (2) effect sizes and (3) the practical importance of effect sizes (ie, their interpretation for clinical practice) would improve our understanding of the meaning of study findings. Let us not throw out the baby with the bathwater.

Twitter Cole Wayant @ColeWayant_OK

Contributors HA and MV conceptualised the paper. CW extracted all the data. HA, MV and CW wrote the manuscript and approve of it in its final form.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Cole Wayant http://orcid.org/0000-0001-8829-8179

References

1 **Benjamin DJ**, Berger JO, Johannesson M, *et al*. Redefine statistical significance. *Nat Hum Behav* 2018;**2**:6–10.
2 **Amrhein V**, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;**567**:305–7.
3 **Aguinis H**, Werner S, Lanza Abbott J, *et al*. Customer-centric science: reporting significant research results with rigor, relevance, and practical impact in mind. *Organ Res Methods* 2010;**13**:515–39.
4 **Rosnow RL**, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 1989;**44**:1276–84.
5 **Sedlmeier P**, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? *Psychol Bull* 1989;**105**:309–16.
6 **Aguinis H**, Stone-Romero EF. Methodological artifacts in moderated multiple regression and their effects on statistical power. *J Appl Psychol* 1997;**82**:192–206.
7 **Conn VS**, Ruppar TM, Chase J-AD. Blood pressure outcomes of medication adherence interventions: systematic review and meta-analysis. *J Behav Med* 2016;**39**:1065–75.
8 **New England Journal of Medicine**. Submitting to NEJM, 2019. Available: https://www.nejm.org/author-center/new-manuscripts [Accessed 1 Oct 2019].
9 **Journal of the American Medical Association**. Instructions for authors: statistical methods and data presentation, 2019. Available: https://jamanetwork.com/journals/jama/pages/instructions-for-authors#SecStatisticalMethodsandDataPresentation [Accessed 1 Oct 2019].
10 **The Lancet**. Information for authors, 2019. Available: https://els-jbs-prod-cdn.literatumonline.com/pb/assets/raw/Lancet/authors/tl-info-for-authors-1568214645933.pdf [Accessed 1 Oct 2019].
11 **Lang TA**, Altman DG. Basic statistical reporting for articles published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *Int J Nurs Stud* 2015;**52**:5–9.
12 **Annals of Internal Medicine**. Information for authors - general statistical guidance, 2019. Available: https://annals.org/aim/pages/AuthorInformationStatisticsOnly [Accessed 1 Oct 2019].
13 **International Conference on Harmonisation**. Ich Harmonised tripartite guideline statistical principles for clinical trials E9, 1998. Available: https://database.ich.org/sites/default/files/E9_Guideline.pdf [Accessed 1 Oct 2019].
14 **Schulz KF**, Altman DG, Moher D, *et al*. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c332.