J Extra Corpor Technol. 2020;52:88–9
The Journal of ExtraCorporeal Technology

Editorial

p-Values and Their Unintended Consequences

We all do it or have done it.

We're busy and have competing demands on our time.

It's a quick and easy way to figure out if we should bother reading any more of the article than we need to.

What are we talking about?

The column titled "*p*-value."

It's easy and convenient to scan down that column provided by the authors (Table 1), and demanded by reviewers, to support or refute our position and/or that of the authors.

It is important to note that a *p*-value does not tell us anything about truth (1). Its role is simply to accept or reject a null hypothesis within the confines of a sample of subjects amassed by the investigators. A null hypothesis is one that is proposed by the authors, such as "del Nido cardioplegia is not associated with in-hospital mortality for adult patients undergoing non-emergent coronary artery bypass grafting." If the *p*-value is less than .05 for that association (within a study sample), we will reject the null hypothesis. An alternative hypothesis might be stated as "del Nido cardioplegia is associated with in-hospital mortality for adult patients undergoing non-emergent coronary artery bypass grafting." Purists would lead a reader to dismiss any finding where $p > .05$. In what circumstance might that be a problem? Consider the situation in which an investigator embarks on a study without actually conducting a statistical power calculation—that is, they have not calculated how many patients are needed to avoid incorrectly failing to reject the null hypothesis. The investigator is at risk of a Type II error if the sample size is not large enough and a statistical test finds that the *p*-value is greater than .05. That is, the finding of non-significance is in fact in error. In this case, blame the investigator, not the finding.

On the flip side, a *p*-value less than .05 (i.e., statistical significance) in a very large sample provides very little information regarding the relevance of the finding to one's own clinical practice absent some information related to the magnitude of the treatment effect (see the following for more information). And a very large sample size tends to result in smaller *p*-values, with an increased likelihood for a Type I error.

What are we to do?

We provide a suggested course of action, for both journal editors and readers, to address this issue:

1) Journal editors and peer reviewers should ensure the following:
   a) Authors should correctly interpret results from their analyses, including but not limited to *p*-values. Importantly, conclusions should not be made solely based on *p*-values (and threshold such as .05) but instead based on clinically meaningful effect sizes, sample size considerations, and other design considerations.
   b) Authors explicitly state whether they have conducted sample size and statistical power calculations for their primary outcome for research investigations.
      i) For studies of small samples sizes, the authors should acknowledge the issue of a Type II error.
   c) Authors explicitly report treatment effect sizes for their main comparisons. A treatment effect size is, for example, a measure of the magnitude of effect that an intervention has on a primary outcome. In the case of del Nido mentioned previously, a journal editor would require that the authors report the in-hospital mortality rates for both those receiving and not receiving del Nido cardioplegia.
   d) Authors explicitly report 95% confidence intervals for their main comparisons. A 95% confidence interval means that there is 95% certainty that the true population values fall within this interval. Wide intervals, for instance, suggest a lack of precision in the estimates (due in many instances to an insufficient sample size).
   e) Authors report exact *p*-values rather than "$p < .05$" or "$p > .05$." Statistical programs are now well-equipped to provide exact *p*-values (2).
2) Readers should consider the following:
   a) Developing their own internal estimate of what a meaningful treatment effect would be. For example, would a .05% absolute difference in-hospital mortality attributed to del Nido be meaningful (assuming a base rate of 2.0% of in-hospital mortality for patients undergoing coronary artery bypass grafting)?

Address correspondence to: Donald S. Likosky, PhD, Section of Health Services Research and Quality, Department of Cardiac Surgery (5346 CVC), Michigan Medicine, Ann Arbor, MI 48109-5864. E-mail: likosky@umich.edu

Disclaimer: Although Blue Cross Blue Shield of Michigan and MSTCVS-QC work collaboratively, the opinions, beliefs, and viewpoints expressed by the authors do not necessarily reflect the opinions, beliefs, and viewpoints of BCBSM or any of its employees, nor do they not reflect the official position of the AHRQ, the NHLBI, or the U.S. Department of Health and Human Services.

Table 1. Unintended consequences of *p*-value thresholds.

p-Value	Statement	Reader's Action	Accurate Interpretation of *p*-Value
<.05	Significant	Read on because of finding of significance	There is less than a 5% probability of finding a value equal to or as extreme as what is reported, assuming null hypothesis is true.
.05	Marginally significant	Not sure what to do	There is a 5% probability of finding a value equal to or as extreme as what is reported, assuming null hypothesis is true.
>.05	Not significant	Findings are not of interest	There is greater than a 5% probability of finding a value equal to or as extreme as what is reported, assuming null hypothesis is true.

b) First reviewing the reported treatment effect and only thereafter reviewing the confidence intervals and *p*-values.

Readers wishing more detail on *p*-values may be interested in articles by Greenland (3) and Amrhein (4).

Xiaoting Wu, PhD
Donald S. Likosky, PhD
Department of Cardiac Surgery, Michigan Medicine, University of Michigan, Ann Arbor, Michigan

ACKNOWLEDGMENT

Dr. Likosky receives funding from the Agency for Healthcare Research and Quality (R01HS026003 AHRQ) and the NHLBI (HL146619-01A1). Support for the MSTCVS Quality Collaborative was provided by the Blue Cross and Blue Shield of Michigan and Blue Care Network as part of the BCBSM Value Partnerships program.

REFERENCES

1. Altman N, Krzywinski M. Interpreting P values. Nat Methods. 2017; 213:213–4.
2. Dahiru T. P - value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med. 2008;6:21–6.
3. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.
4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–7.