results will show whether our prior beliefs are in concordance with what the data show or whether the data are strong enough in the other direction to overwhelm our prior beliefs. It is, philosophically, moving more toward thinking about whether a result fits with our clinical thinking and helps incorporate clinical thinking into the analysis.

One final point pertains to the concerns among researchers and clinicians alike that wholesale abandonment of *P* values will result in a chaotic situation wherein the end users of evidence are left not knowing who or what to trust, akin to a lawless frontier town where the sheriffs have abandoned their post.^{4,5} This concern can be mitigated by, instead, focusing on the effect estimate and its uncertainty, an output that is obtained in conjunction with the *P* value but often not reported. In fact, effect estimates are more directly related to clinical decision-making,⁴ as they help us understand what degree of effect an intervention will have on our patient and what the expected range of response is.⁵ This suggestion is precisely in line with the suggestion by Visintainer¹ to adopt confidence intervals and does not require more sophisticated statistical understanding or analysis.

As statistical understanding has advanced, and statistical education has been increasingly incorporated into surgical training in academic programs, we have the opportunity to adopt more appropriate methodologic techniques to answer our clinical questions, communicate the answers we find, and, ultimately, bring improvements to patient care. As surgeons, we are experienced in innovating and adopting new technologies, and, similarly, we must be prepared to move beyond our historical reliance on *P* values to more nuanced analyses of surgical data.

References

1. Visintainer P. Moving beyond significance testing: confidence intervals in clinical research. *J Thorac Cardiovasc Surg*. 2021;161:1373-6.
2. Harrington D, D'Agostino RB Sr, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, et al. New guidelines for statistical reporting in the journal. *N Engl J Med*. 2019;381:285-6.
3. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129-33.
4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond " $p < 0.05$ ". *Am Stat*. 2019;73(suppl 1):1-19.
5. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA*. 2018;319:1429-30.

See Article page 1367 in the April 2021 issue.

Commentary: Can I trust the *P* value?

Aaron J. Weiss, MD,^a and
Anelechi C. Anyanwu, MD, FRCS^b

It has been 15 years since physician-scientist and "scourge of sloppy science"¹ Dr John Ioannidis published his seminal paper titled "Why Most Published Research Findings are False."² In it, he proclaims that, "Several methodologists have pointed out that the high rate of nonreplication (lack

Aaron J. Weiss, MD, and Anelechi C. Anyanwu, MD, FRCS

CENTRAL MESSAGE

Our proposed checklist, "can I trust the *P* value?," can be used as a roadmap for individual readers to critically evaluate the appropriateness of published *P* values.

From the ^aDepartment of Thoracic and Cardiovascular Surgery, Heart, Vascular and Thoracic Institute, Cleveland Clinic, Cleveland, Ohio; and ^bDepartment of Cardiovascular Surgery, Mount Sinai Medical Center, New York, NY.

Disclosures: The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Received for publication April 25, 2020; revisions received April 25, 2020; accepted for publication April 26, 2020; available ahead of print May 11, 2020.

Address for reprints: Aaron J. Weiss, MD, Department of Thoracic and Cardiovascular Surgery Heart, Vascular, and Thoracic Institute, Cleveland Clinic, 9500 Euclid Ave, Desk J4-1, Cleveland, OH 44115 (E-mail: weissa2@ccf.org).

J Thorac Cardiovasc Surg 2022;164:e42-4

0022-5223/\$36.00

Copyright © 2020 by The American Association for Thoracic Surgery

<https://doi.org/10.1016/j.jtcvs.2020.04.125>

of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a P -value less than 0.05. Research is not most appropriately represented and summarized by P -values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on P -values.”²

The obsession with P values and statistical significance has persisted over the years despite many evidence-based pleas to the contrary published in pan-specialty journals such as *Nature*,³ *New England Journal of Medicine*,⁴ and *The American Statistician*.⁵ In the April 2021 issue of the *Journal*, Steven Staffa and David Zurakowski once again comment on the widespread troubles that result from P value misuse and overuse.⁶ They provide examples of the inappropriate use of P values for 3 commonly performed statistical analysis designs in clinical-level cardiothoracic surgery research, as well as guidance for appropriate practice. Although the authors provide valuable direction, further systematic efforts to curtail P value misuse that were not covered might include incorporating standardized education for resident training in statistical analysis and research design. Such concepts as uncertainty and reproducibility, and the value of involving trained statisticians in the early stages of study design to appropriately pair statistical design with the research question and study workflow, are also important.

Unfortunately, despite the best intentions of Staffa and Zurakowski, as well as the additional suggestions provided above, P value and statistical significance abuse will likely continue due to the overexpansion of journals and the academic culture of quantity, rather than quality, inherently tied to career promotion. What additional efforts can then be made at the reader level to help provide solutions to these problems? How can the average cardiothoracic surgeon reading a manuscript make a critical appraisal

of the study, the measurement of effect, and the attributed P value?

To tackle this problem at the reader level, we propose a checklist of questions titled “Can I trust the P value?” (Table 1) that readers should consider when appraising published studies. This list is intended to be a roadmap for the individual cardiac surgeon to follow to decide for themselves whether the study results reflect clinically relevant data and whether reported differences are translatable to their everyday practice. This is by no means an exhaustive list, and we invite individuals to go beyond the proposed checklist in their understanding of research study design and statistical analysis and how it may uniquely apply to each publication.

Our checklist is meant to stimulate the individual reader to systematically evaluate the nature of the P values being reported and then consider the context in which these are parlayed into inference. However, it must be stressed that rejecting the null hypothesis based on a P value does not mean that the null hypothesis is indeed false and that any effect seen is causal; for example, increased use of mobile phones may be associated with increased prevalence of obesity, but this does not mean that mobile phone use causes obesity. Sometimes the study methodology may result in increased chances that the null hypothesis is inappropriately rejected. Likewise, when the null hypothesis is not rejected (traditionally if $P > .05$), this does not necessarily mean the null hypothesis is true.

The P value itself is a probability based on a null hypothesis, as opposed to an absolute truth, and requires interpretation in the context of the study methodology. The more “no” answers in our checklist (Table 1), the less robust the interpretation of the P value. Readers should also be aware that the clinical effect being assessed might not be appropriately reflected in a null hypothesis. For example, there is a distinction between a null hypothesis that states there is no difference in the myocardial infarction rate between group A and group B versus a null hypothesis that states there is no difference in the troponin level between

TABLE 1. Can I trust the P value?

Question	Yes/no
1. Is there a research hypothesis that includes the effect under study?	
2. Is the study design appropriate to answer the question or measure the effect being qualified by the P value?	
3. Does the study methodology exclude or adjust for major sources of bias and confounding?	
4. Are the study subjects representative of those in a wider population?	
5. Are the treatments and surgical techniques reproducible and generalizable?	
6. Has the effect tested by the P value been measured and defined in an appropriate and standardized fashion?	
7. Is the effect, magnitude of effect, and direction of effect, measured by the P value clinically relevant?	
8. Does the measurement of effect have acceptable precision (confidence boundaries)?	
9. Does the P value suggest a high probability that the null hypothesis is false?	
If the answer is “yes” for all, then it is reasonable to reject the null hypothesis and assume that the statistical test identifies an association that is most probably real.	

group A and group B. Even if myocardial infarction is defined based solely on troponin, the latter hypothesis using a more sensitive numerical scale is more likely to be rejected ($P < .05$) than the former, because the former uses a more clinically relevant categorical scale (either you have a myocardial infarction or you do not). Thus, the same set of data can produce different results on influence of treatment type on myocardial infarction depending on how the null hypothesis was framed.

A sustained and complementary effort must be made by individual readers, researchers, and journals alike to hold ourselves to a higher standard and prevent the rampant P value misuse and overuse plaguing our field. We must refrain from the motivations that prioritize searching for statistical significance in the absence of clinical relevancy and must strive for scientifically robust and reproducible

underlying relationships that can be translated to the bedside to help care for our patients. For far too long, the “tool has become the tyrant,”⁵ and it is time for us to be the sharper tools in the shed.

References

1. Ioannidis J. John Ioannidis: uncompromising gentle maniac. *BMJ*. 2015;351:h4992.
2. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
3. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567:305-7.
4. Harrington D, D'Agostino RB Sr, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, et al. New guidelines for statistical reporting in the *Journal*. *N Engl J Med*. 2019;381:285-6.
5. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “ $p < 0.05$.” *Am Stat*. 2019;73:1-19.
6. Staffa SJ, Zurakowski D. Expert opinion: guidelines for improving the utilization and presentation of P values. *J Thorac Cardiovasc Surg*. 2021;161:1367-72.

See Article page 1379 in the April 2021 issue.

Commentary: Let's not trade one problem for another: Moving beyond P values and confidence intervals

Derrick Y. Tam, MD, and
Stephen E. Fremes, MD, MSc

The reliance on the P value as the threshold for statistical significance in biomedical studies has become controversial recently. Its pitfalls have been well described by Greenland and colleagues in a series of statements from the American Statistical Association.¹⁻³ Issues of multiplicity are rampant in cardiovascular medicine,⁴ and also exist in experimental

Stephen E. Fremes, MD, MSc, and Derrick Y. Tam, MD

CENTRAL MESSAGE

The duality between P values and confidence intervals does not eradicate the problem of multiplicity; other statistical tests and concepts must be considered.

science and the social sciences.^{5,6} The publication of this series in *The American Statistician* was the genesis for the statistical editorial policy change in *The New England Journal of Medicine* (NEJM), aimed at minimizing the risk of a type 1 error.⁷ In essence, investigators should describe a plan a priori and remain true to that plan. P values should not be reported for secondary outcomes or post hoc analyses. Instead, the presentation of both point estimates and

From the Division of Cardiac Surgery, Department of Surgery, Schulich Heart Centre, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada.

Disclosures: The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Received for publication April 29, 2020; revisions received April 29, 2020; accepted for publication April 30, 2020; available ahead of print May 30, 2020.

Address for reprints: Stephen E. Fremes, MSc, MD, Dr Bernard S. Goldman Chair in Cardiovascular Surgery, Schulich Heart Centre, Sunnybrook Health Sciences Centre, 2075 Bayview Ave, Room H4 05, Toronto, Ontario, Canada M4N 3M5 (E-mail: stephen.fremes@sunnybrook.ca).

J Thorac Cardiovasc Surg 2022;164:e44-5

0022-5223/\$36.00

Copyright © 2020 by The American Association for Thoracic Surgery

<https://doi.org/10.1016/j.jtcvs.2020.04.159>