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>

margin of error (ie, the confidence interval [CI]) should be reported for the treatment or exposure.

In the April 2021 issue of the *Journal*, Seo Young Park⁸ highlights that the duality between *P* values and CIs does not eradicate the problem of multiplicity. As Dr Park points out, searching for a CI that excludes 0 or 1 is no different than repeated null hypothesis significance testing for a *P* value $< .05$ (*P* hacking). The changed NEJM statistical policy could be subject to the same potential abuse that the new policy attempts to address. As Dr Park insightfully notes, we must consider alternatives to *P* values and CIs.

To this end, Dr Park begins by describing the alternative use of *S* values or bits, a concept born from information theory that is interesting from a historical perspective.^{9,10} More important is the suggestion to use standardized mean differences (SMDs) rather than *P* values. Using SMDs is a very practical recommendation for presentation of the typical baseline characteristics in first table of an observational study. The SMD shows the magnitude of the difference regardless of the sample size, whereas the *P* value is a function of both the difference and the sample size. Large datasets will inevitably have numerous miniscule *P* values of uncertain clinical significance, whereas small studies may be underpowered; differences of fairly large magnitude can have large *P* values yet be clinically important.

The use of SMDs is common before and after propensity score–matched observational studies but is valuable in other contexts as well.¹¹ In social sciences, where outcomes are typically measured in scales, the exact same calculation for SMD is commonly used to measure differences in outcomes. This is termed Cohen's *d*, and meaningful effects can be interpreted as small (>0.2), medium (>0.5), or large (>0.8).¹² The other related suggestion is considering testing

against a clinically important difference rather than the null, or no effect.

We thank Dr Park for bringing to the forefront these important considerations for investigators when planning their statistical analyses. As investigators, we have the responsibility to reduce the type I error when reporting experimental results, but need to be cognizant that the substitution of a CI for *P* values might not be a sufficient condition to accomplish the stated goal.

References

1. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond " $p < 0.05$." *Am Stat*. 2019;73:1-19.
2. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-33.
3. 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;31:337-50.
4. Khan MS, Khan MS, Ansari ZN, Siddiqi TJ, Khan SU, Bin Riaz I, et al. Prevalence of multiplicity and appropriate adjustments among cardiovascular randomized clinical trials published in major medical journals. *JAMA Netw Open*. 2020;3:e203082.
5. Michel MC, Murphy TJ, Motulsky HJ. New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. *J Pharmacol Exp Ther*. 2020;372:136-47.
6. Trafimow D, Amrhein V, Areshenkoff CN, Barrera-Causil CJ, Beh EJ, Bilgic YK, et al. Manipulating the alpha level cannot cure significance testing. *Front Psychol*. 2018;9:699.
7. Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand SLT, et al. New guidelines for statistical reporting in the *Journal*. *N Engl J Med*. 2019;381:285-6.
8. Park S. Replacing *P* values with confidence intervals may not achieve anything. *J Thorac Cardiovasc Surg*. 2021;161:1379-80.
9. Greenland S. Valid *p*-values behave exactly as they should: some misleading criticisms of *p*-values and their resolution with *s*-values. *Am Stat*. 2019;73(suppl):106-14.
10. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379-423.
11. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28:3083-107.
12. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Front Psychol*. 2013;4:863.