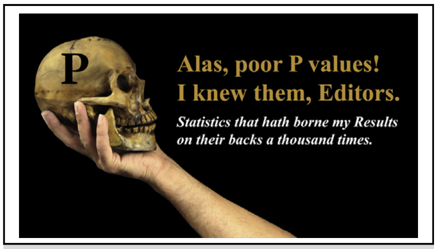
To *P* or not to *P*, that is the question: Four expert opinions on the *P* value controversy

Eugene H. Blackstone, MD


P
Alas, poor P values!
I knew them, Editors.
Statistics that hath borne my Results on their backs a thousand times.

Alas, poor *P* values! Statistics that hath borne my results on their backs a thousand times.

CENTRAL MESSAGE

The New England Journal of Medicine author instructions encourage reporting the magnitude and direction of effects and confidence intervals rather than *P* values. Statisticians for this *Journal* discuss this controversial directive.

See Articles pages 1367, 1373, 1377, and 1379.

In this issue of the *Journal*, 4 PhD statisticians weigh in on the use and misuse of *P* values. The series is stimulated by new instructions to authors for manuscripts submitted to *The New England Journal of Medicine* that include limiting the use of *P* values in favor of other metrics, with justification by eminent statisticians.[1] The controversy is not new; it has raged since the 2016 unprecedented White Paper from the American Statistical Association that cites important misuses of *P* values,[2,3] accompanied by many commentaries by statisticians with varying viewpoints. As I introduce each expert opinion, I will comment briefly from personal experience.

Adin-Cristian Andrei, PhD, begins his commentary provocatively with "Statistical Testing in Crisis."[4] Through the visionary leadership of Joseph F. Volker, DDS, the fledgling University of Alabama Medical Center (later to become The University of Alabama at Birmingham [UAB]) had successfully recruited cardiac surgical pioneer Dr John W. Kirklin from the Mayo Clinic in 1966 and, shortly thereafter, mathematician and computer scientist Josiah (Jay) Macy, Jr, from Albert Einstein College of Medicine in New York. In my interview for a position with Dr Kirklin, he recognized that I also might fit well into Dr Macy's Biophysical Sciences Division. In the course of our conversation, Dr Macy introduced me to "the light bulb." The only thing you need at the end of a clinical trial, he said, was a light bulb. If the trial were successful with $P < .05$, the light bulb would go on. That's all you needed to know! Even today, our *Journal* receives manuscripts that, particularly in the Abstract, report only *P* values, with no direction of effect, no magnitude of effect accompanied by confidence limits, no indication of clinical relevance, just statistical significance. Dr Andrei puts statistical significance into the context of conducting and reporting good science, not throwing it out with the proverbial baby in the bathwater just because of misuse. At the same time, he points to methods that provide more information than just the *P* value.

Steven Staffa, MS, and David Zurakowski, PhD, go a step further than even the American Statistical Association in presenting an exceptionally helpful and appealing quadrangle diagram of *P* value overuse and misuse accompanied by specific cardiac surgery examples.[5] An addition to the *P* value controversy by these authors is the matter of multiple testing and correcting of *P* values for this. Multiple testing—when it is needed, when it is not—is not understood by authors, readers, or many statisticians outside the field of clinical trials. That would be a good topic for a commentary!

Frank Harrell, PhD, then at Duke, and David Naftel, PhD, in our group at UAB, enjoyed walking the Appalachian Trail about twice per year. Before their departure, we would meet in the conference room for a statistical "food fight." One such topic was multiple testing. Consider candidate

From the Sydell and Arnold Miller Family Heart, Vascular, and Thoracic Institute and Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland Clinic, Cleveland, Ohio.

Received for publication Dec 10, 2020; revisions received Dec 10, 2020; accepted for publication Dec 10, 2020; available ahead of print December 28, 2020.

Address for reprints: Eugene H. Blackstone, MD, Department of Thoracic and Cardiovascular Surgery, Cleveland Clinic, 9500 Euclid Ave/Desk JJ-40, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2021;161:1365-6
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.12.091

variables for a multivariable analysis of an outcome. Every variable that you consider has a chance of being falsely related to outcome—so should one apply a progressively lower *P* value threshold for every variable considered? What if you have 100 variables? How much lower should the *P* value threshold be? What if you dream about variables in your sleep? Should the *P* values be set at an even lower threshold? These are the extremes of such statistical food fights! Of course, Staffa and Zurakowski[5] have a more attractive and informative suggestion: use standardized differences instead of *P* values when reporting baseline differences between groups, something also advocated by Seo Young Park, PhD, in her expert opinion.[6]

Dr Park also introduces the *S* statistic, a logarithmic transformation of the *P* value.[6] The *S* statistic transforms the *P* value into bits of information. This reminds me of teaching information theory to graduate students at UAB. There is a strong relationship between statistical information theory and quantities at the heart of computational information theory and even communications information theory. Dr Park also uses a term that I first heard discussed by Malcolm Turner, PhD, the late chair of biomathematics at UAB: "expressions of surprise." He and I worked for years on trying to come up with an expression of surprise that was not as strongly dependent on sample size as the *P* value. Standardized differences are helpful. But nothing irritates more than an author stating that treatment X significantly affects Y, but Z is ineffective when, looking at the data, the first has a *P* value of .045 and the magnitude of difference is 24%, and the second has a *P* value of .061 and the magnitude of difference is 29%. The second comparison is based on a smaller sample size, but the paper concludes that Z is ineffective, but Y is effective.

Paul Visintainer, PhD, in his expert opinion, raises this same issue.[7] Like *The New England Journal of Medicine*, Dr Visintainer defends the case for wide use of confidence intervals rather than *P* values. He starts with the provocative statement, "It is difficult to see what useful information a clinician might derive from knowing the *P* value." He shows that confidence intervals provide the same information as *P* values but go beyond them in showing that magnitude and direction of effects in a clinically useful format, and also provide important information about precision of that effect by width of the interval.

This brings up my recommendation to authors about what confidence intervals to use. The knee-jerk choice today is 95%. However, this shows that we need to return to lessons of the 1970s and 1980s (or before). Back in "those days of yore," we recommended 70% confidence limits for proportions and actuarial curves as scanning tools. If the upper 70% confidence limit of a proportion just touches the lower 70% confidence limit of another, then the *P* value will be just a bit above .05; 95% confidence limits under the same scenario would be associated with a *P* value around .01. The equivalent confidence interval for a ratio or a difference would be 95%. Thus, we selected confidence limits that told a consistent story.[8]

Finally, I must come back to Dr Park. At the end, she says, "I believe that many new methods will be proposed." In an invited expert opinion several years ago, Lu and Ishwaran[9] explained how all the measures used in machine learning could replace *P* values of ordinary statistical regression analyses.[10] It may be the rapid developments in machine learning that wean us from *P* values, allowing us to appreciate those situations when they are useful and appropriate. But, with apologies to William Shakespeare,

Shall we take arms against a sea of P *values,*
And by opposing end them?
But that the dread of something after P *values,*
The undiscover'd country from whose bourn
No traveller returns, puzzles the will
And makes us rather bear those ills we have
Than fly to others that we know not of?

Conflict of Interest Statement

The author 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.

References

1. 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.
2. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70:129-33.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "*p* < .05." *Am Stat*. 2019;73:1-19.
4. Andrei AC. Statistical significance: is there a way out of it? *J Thorac Cardiovasc Surg*. 2021;161:1377-8.
5. Staffa SJ, Zurakowski D. Guidelines for improving the use and presentation of *P* values. *J Thorac Cardiovasc Surg*. 2021;161:1367-72.
6. Park SY. Replacing *P* values with confidence intervals may not achieve anything. *J Thorac Cardiovasc Surg*. 2021;161:1379-80.
7. Visintainer PF. Moving beyond significance testing: confidence intervals in clinical research. *J Thorac Cardiovasc Surg*. 2021;161:1373-6.
8. Kouchoukos NT, Blackstone EH, Hanley FL, Kirklin JK. *Kirklin/Barratt-Boyes Cardiac Surgery*. 4th ed. Philadelphia: Elsevier; 2012. 295-7.
9. Lu M, Ishwaran H. A prediction-based alternative to *P* values in regression models. *J Thorac Cardiovasc Surg*. 2018;155:1130-6.e4.
10. Blackstone EH. Can we live without *P* values? The answer. *J Thorac Cardiovasc Surg*. 2018;155:1137.