Critical Evaluation of Data Requires Rigorous but Broadly Based Statistical Inference

Nancy J. Cox, Jennifer E. Below

The rampant misuse of the *P* value and of its stated meaning lead the American Statistical Association to comment, researchers often wish to turn a *P* value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *P* value is neither. It is a statement about data in relation to a specified hypothetical explanation and is not a statement about the explanation itself.[1] The American Statistical Association is not alone in its concern: a host of recent literature provides context for the desire to use statistical inference to reinforce rigor and reproducibility in scientific research.[2–4] A central focus of this literature is the widely acknowledged and severe limitations we impose on ourselves with a blind and naive adherence to the exclusive use of *P* values for understanding significance of research findings.

Counterpoint, see p 1046
Response by Moyé and Cohen, see p 1051

The inadequacy of *P* values as a singular and standard criteria for significance in the analysis of clinical trials has been known and communicated by statisticians for >35 years.[5] Given the limitations of significance testing, can we simply, as suggested by Moye and Cohen, resolve to walk away from *P* values? Many have argued the use of methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals that can be used in place of or in addition to *P* values.[6–8] Others have lobbied for alternative measures of evidence, such as likelihood ratios or Bayesian methodologies, which leverage a priori information about the probability of different magnitudes of effect that result from the treatment and modify the estimation of the effects based on the observed data.[9–11] Still, others propose approaches such as decision-theoretical modeling and improved use of false discovery rates.[1] Importantly, Mark et al[12] remind us that the "clinical or scientific importance of study results is a judgment integrating multiple elements, including effect size (expected and observed), precision of estimate of effect size, and knowledge of prior relevant research. At best, *P* value has a minor role in shaping this judgment."

Each of these proposed solutions to effectively evaluate research results strives to improve statistical inference and significance testing—to broaden our understanding of the data and what it might be trying to convey to us—rather than abandon it. In fact, far from viewing the common requirement of *P* value <0.05 as tyrannical, some in the statistical community are deeply concerned about reproducibility and replicability of scientific conclusions; a compelling argument was recently made by a persuasive cohort of statisticians to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries to improve reproducibility.[13] Although it is clear that choosing a more stringent, but still arbitrary, threshold for establishing significance cannot solve the pernicious problem of the false dichotomy of labeling findings as "significant" or "insignificant," scientists loathe irreproducibility of established results.

Clinical trials face a particularly difficult set of challenges. The human and resource costs associated with clinical trials result in studies of modest sample size that are limited in power. Any success of previous drug development inevitably makes future trials more expensive because only drugs that perform better than what is already in use will be of interest. Thus, the more success has been achieved, the harder it is to surpass. Such trials generate reams of high-quality clinical data, but because of the limited number of subjects, they are often considered to be sufficiently powered to explore only a limited number of primary end points. Moyé and Cohen argue that one of the reasons *P* values specifically, and significance testing generally, should be abandoned is because these approaches define the number of hypotheses that can be tested powerfully in any given data set, preventing us from learning all that is learnable from the data. It is, indeed, unconscionable for such high-quality data collected at such expense to go largely unexamined.

The history of the use of statistics in genetic and genomic studies of common human diseases offers some insights that may be of use to those frustrated by the tyranny of the *P* value in clinical trials research. A key challenge in any field is coming to some understanding of what the effect sizes are likely to be; the oomph we are looking for, as characterized by Mark et al.[12] The challenges of multiple testing and detection of modest effects in clinical trials is mirrored in the decades-long history of genetic linkage mapping of common disease. This work served largely to reinforce the distinction between rare Mendelian diseases—in which effect sizes were invariably sufficiently large to yield strong evidence for linkage when sufficient data had been collected—and common diseases with more complex transmission patterns—in which linkage mapping failed because the effect sizes were sufficiently small to preclude robust detection. Genetics then faced the

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN.

Correspondence to Jennifer E. Below, PhD, Vanderbilt University Medical Center, The Vanderbilt Genetics Institute, Light Hall #519B, 2215 Garland Ave, Nashville, TN 37232. E-mail jennifer.e.below@vanderbilt.edu

(*Circ Res*. 2018;122:1049-1051.
DOI: 10.1161/CIRCRESAHA.118.312530.)

Circulation Research is available at http://circres.ahajournals.org
DOI: 10.1161/CIRCRESAHA.118.312530

same problem clinical trials face now—how to design studies sufficiently powered to detect modest effects under multiple hypotheses? In the case of single variant genome-wide testing, this challenge is daunting: we need to be able to robustly test *millions* of independent hypotheses.

Abandoning significance testing wasn't the solution. The first step was to improve technology and reduce the cost of collecting the needed data. These advancements permitted genome-wide association studies, a design that allows detection of more modest signals. Yet many of the initial applications of genome-wide association studies to common disease still missed the mark: studies were too small to confidently estimate effect sizes for much of the genome variation contributing to common disease (often, odds ratio<1.2). Today, discovery at scale and real understanding of the genetic component to common disease is coming only with *massive* sample sizes (today studies typically include tens of thousands to millions of samples) and deep genome interrogations that seemed downright impossible before improvements in technology and more efficient study designs allowed us to create new kinds of studies.[14] Although much can be taken from these studies, one lesson that resonates with the commentary from Moye and Cohen is the importance of looking at all of the data in as many ways as can be imagined, but, and this is a big caveat, with an abundance of data. We are beginning to understand how naive we had been: more data than we could have imagined would be needed to make robust and accurate estimates of genomic effects on common diseases, regardless of statistical approach.[15,16]

Although that seems like a tall order for studies as expensive as clinical trials, there are a host of design innovations that are likely to create opportunities for more cost-effective trials,[17,18] and new ways of understanding the variability among subjects (genomics and other omics are obvious ways from our perspective, but disease imaging and new biomarkers are rapidly advancing as well). General progress is occurring through the renewed focus on valid and reliable measurements, as well as the publication of complete summary-level analytic data even for data sets that do not reach stringent levels of significance.[19,20] PLOS's relatively new collection: The Missing Pieces: A Collection of Negative, Null and Inconclusive Results, the National Institutes of Health's commitment to data and results sharing, and data and full results availability requirements of many top journals are all steps in the right direction. Moreover, although we think of clinical trials as being utterly unique, many elements of any trial are common to other trials and also to large publicly available data sets such as large repositories of electronic health records and biobanks. We have the opportunity to explore many secondary data characteristics and become more knowledgeable about the nature of clinical trial data and its variability if we adequately exploit these commonalities even in our prospectively declared analyses. By looking beyond the silo of our own research and aggregating all available and appropriate data sources, we should be able to powerfully explore many end points and additional facets of the data without abandoning the rigorous statistical inference that shepherds discovery.

Disclosures

None.

References

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133.
2. Hunter P. The reproducibility "crisis": reaction to replication crisis should not stifle innovation. *EMBO Rep*. 2017;18:1493–1496. doi: 10.15252/embr.201744876.
3. Peng R. The reproducibility crisis in science: a statistical counterattack. *Significance*. 2015;12:30–32.
4. Scannell JW, Bosley J. When quality beats quantity: decision theory, drug discovery, and the reproducibility crisis. *PLoS One*. 2016;11:e0147215. doi: 10.1371/journal.pone.0147215.
5. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med*. 1983;98:385–394.
6. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA*. 2016;315:1141–1148. doi: 10.1001/jama.2016.1952.
7. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J (Clin Res Ed)*. 1983;286:1489–1493.
8. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292:746–750.
9. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
10. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA*. 1995;273:871–875.
11. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *J Epidemiol Community Health*. 1998;52:318–323.
12. Mark DB, Lee KL, Harrell FE Jr. Understanding the role of P values and hypothesis tests in clinical research. *JAMA Cardiol*. 2016;1:1048–1054. doi: 10.1001/jamacardio.2016.3312.
13. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B, Brown L, Camerer C. Redefine statistical significance. *Nat Hum Behav*. 2018;2:6–10. doi: 10.1038/s41562-017-0189-z.
14. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–678. doi: 10.1038/nature05911.
15. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90:7–24. doi: 10.1016/j.ajhg.2011.11.029.
16. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101:5–22. doi: 10.1016/j.ajhg.2017.06.005.
17. Sertkaya A, Birkenbach A, Berlind A, Eyraud J. *Examination of Clinical Trial Costs and Barriers for Drug Development*. Washington, DC: Report, US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation; 2014:1–92.
18. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? *Nat Rev Drug Discov*. 2017;16:381–382. doi: 10.1038/nrd.2017.70.
19. Ioannidis JP. Journals should publish all "null" results and should sparingly publish "positive" results. *Cancer Epidemiol Biomarkers Prev*. 2006;15:186. doi: 10.1158/1055-9965.EPI-05-0921.
20. van Assen MA, van Aert RC, Nuijten MB, Wicherts JM. Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One*. 2014;9:e84896. doi: 10.1371/journal.pone.0084896.

Key Words: clinical trials ■ sample size ■ statistics

Response by Lem Moyé and Michelle Cohen

We thank Drs Below and Cox for their response to our Viewpoint. They have elegantly differentiated the needs of the clinical investigators from the work product of statisticians.

The issue of reproducibility is not just the purview of the statisticians; scientists who design research efforts and generate their data have important concerns about the consistency of findings across studies. However, the solution to reduce the *P*-value threshold from 0.05 to 0.005 is unacceptable, given the current state of the *P*-value's overreliance and abuse. Result reproducibility would be more helpfully mitigated and improved by (1) a prospective definition by the scientists in the field as to what similarity of research results constitutes, (2) a common set of clinical trial design parameters, for example, inclusion/exclusion criteria, treatment doses studied, duration of follow-up, endpoint similarity, and (3) common measures of effect. The effort to create this research climate would be substantial but would offer superior service to the research community then merely adjusting the threshold of the *P* value. Fortunately, this decision is not in the hands of any one group, for example, biostatisticians. The scientists who the biostatisticians serve must shape the final solution to the reproducibility criteria to which they will be held.

The response authors state that the goal of choosing a small number of end points is to provide some assurance that the analyses conducted would be adequately powered. Yet, there are many prospectively declared analyses that have adequate power but play only a minimal role in the evaluation of the effect of the exposure being evaluated. They are specifically excluded because of the customary and limited way in which sampling error is currently managed in the clinical trial paradigm. This limitation is a consequence of a statistical analysis system that has been retrofit onto healthcare research.

The tension in the matter at hand is the difference between the raison d'etre of statistical hypothesis testing on the one hand, and the determination of the clinical investigators on the other to draw a global conclusion about the complicated effects of the treatment being studied. The investigators require a joint consideration of multiple study outcomes. This ensemble of effects is essential in medicine, yet is poorly handled by *P*-value based evaluations. Statistical inference insists on parsing the original complex clinical questions into several (and sometimes many) statistical hypothesis tests, then computes and accumulates type I error for each test until the overall type I error is calculated. From the investigators' perspective, the individual tests were never the goal and errors associated with them are irrelevant. It is this statistical dehiscence of the larger clinical question into a number of smaller questions with clinically inconsequential type I error penalties that the Viewpoint's authors agree should be abandoned.