Liberation From the *P* Value's Tyranny

Lem Moyé, Michelle Cohen

P-values, next to nuclear weapons, are the worst invention of the 20th century.

—Herman Rubin, Purdue University, 1981

More than a half century ago, the *P* value was selected as the measuring tool for health care research by administrators. The difficulties generated by statistical hypothesis testing in addition to the constraints it places on cardiovascular trial researchers must now be declared unacceptable by the cardiology research community.

Developed >90 years ago by statistician, Ronald Fisher, while working on a manure experiment in England,[1,2] *P* values have risen to dominate the field of cardiovascular investigation for no other reasons than those of custom and habit. The abuse of *P* values has become so rampant that the American Statistical Association in 2016 was compelled to provide a consensus statement decrying their misuse.[3]

Counterpoint, see p 1049

P Value Ascendency

Statistical hypothesis testing and *P* values were criticized by scientists from the beginning. They thought that its reverse logic—embracing a null hypothesis that one did not think, only to have the data reject it compelling the scientist to accept what one did think—was tortured and unnecessarily complex. *P* values rose to prominence in health care when grant administrators, journal editors, and Food and Drug Administration officials, overwhelmed by the post–World War II explosion in research, chose the *P* value instrument to help identify worthy research.[4] The *P* value was not chosen because it had any basis in epidemiology, biology, pathophysiology, or cardiovascular medicine. It was selected because these administrators needed a metric to help them to separate the research wheat from the chaff, and they could see no alternative.

The fledgling clinical trial research community accepted this *P* value imprimatur, not realizing that this nonmalicious selection portended pernicious consequences. Even though the *P* value's sole, small role was to manage a probability involving mathematical assumptions and sampling error (The American Statistical Association definition of a *P* value is the probability under a given statistical model that a statistical summary of the data would be equal to or more extreme than its observed value), it was used as an uber-corrective factor for nonstatistical issues. Sketchy protocols subject to data-driven changes, slovenly patient follow-up, and desultory analyses were all methodologic sins permitted to be covered by small *P* values.

The damage escalated to cardiology as the race to the 0.05 level generated public health havoc. In their haste to obtain statistically significant results, august researchers stampeded over contributory findings from their signature clinical trials, such as MRFIT (Multiple Risk Factor Intervention Trial) and LRC (Lipid Research Clinics).[5] By focusing on the *P* value rather than the plausibility, coherence, and veracity of the findings, the investigators confused the community about the value of blood pressure and lipid control. Heart failure researchers were thrown into confusion by the Vesnarinone trials, followed by the ELITE trials (Evaluation of Losartan in the Elderly) and then the PRAISE trials (Prospective Randomized Amlodipine Survival Evaluation),[5] each comprising a pair of state-of-the-art studies in which *P* value–driven results were inconsistent. The pharmacological approach to cardiovascular research was dismembered by the *P* value polemics that enveloped the US Carvedilol program controversy.[6,7] For each trial, the cardiology investigators, tightly strapped into the research train under the full control of a statistical engineer, fretted over the narrow range of selected destinations; would the train ride the shiny ($P<0.05$) rails of success, or get switched to the side rails of statistical insignificance, ending in this elephant's graveyard of ignored findings.[8]

Consequences

The *P* value is a condensate, constructed from (1) sample size, (2) effect size, (3) the precision of the estimate, and (4) a sampling error assessment (This last measure is incorporated by applying a probability assessment to the test statistic, ie, computing the probability of a value at least as large as the test statistic.). Each of these ingredients is important in the assessment of research interpretation. However, by integrating them all into a single number, the audience commonly succumbs to the temptation of accepting the summary value and not interpreting each component for itself. By doing so, a multifarious research issue is reduced to a mere 1-dimensional projection of the truth.

Curiously, the response of the quantitative science community to this well-recognized problem was not to reign in the *P* value but to constrain the investigators. The result is the current research leviathan. Highly structured, it contains a small number of type I error controlled primary end points, followed by a larger number of supportive secondary end points (exploratory end points, like wayward children, are banished to their rooms). Over time, this formulation was accepted by the cardiovascular and regulatory communities. However, cardiovascular studies are expensive and inefficient. Consuming hundreds if not thousands of patients, hundreds of thousands of person-hours, and millions in public

From the UTHealth School of Public Health, Houston, TX.

Correspondence to Lem Moyé, MD, PhD, UTHealth School of Public Health, 1200 Herman Pressler E-1009, Houston, TX 77030. E-mail LemMoye@msn.com

(*Circ Res*. 2018;122:1046-1048.
DOI: 10.1161/CIRCRESAHA.117.312227.)

***Circulation Research* is available at http://circres.ahajournals.org**
DOI: 10.1161/CIRCRESAHA.117.312227

and private funds, they focus all attention on a small number of end points, resembling an upside-down pyramid teetering on only a small number of *P* values (Figure). Meanwhile, much of the data lie languid, unanalyzed, and unpublished because the analyses are not type I error controlled, or not prospectively declared, or are simply the fruit of the poisoned tree, the so-called negative trial. This inefficiency is a dangerous state of affairs when cardiovascular trials have been criticized for being unaffordable.

Epidemiologists warned us about the difficulties with *P* value primacy,[9,10] yet by ignoring these monitories, we have allowed the *P* value to become weaponized, giving ourselves over to its tyrannical rule. And now with the new call for a reduction in the significance level from 0.05 to 0.005,[11] the impact of smaller required *P* values on trial size, study duration, and financial burden is too terrible to contemplate.

What Is the Problem?

This current, unfortunate state of affairs is because of the insistence on managing sampling error in a manner that is in contravention to clinical research principals. Statistical decision theory's implementation in clinical trials compels the cardiology researcher to answer a question about a null hypothesis, with penalties accrued for the number of questions the cardiologist addresses. However, the purpose of clinical investigation is to observe and learn all that we can from the data and responsibly report it. The type I error management concept of reduced end point selection is one of statistics—not one of nature or the scientific method. This becomes clear when one recognizes that the rules of statistical hypothesis testing are neither based in or derived from the foundation of epidemiology or cardiology but are the collection of ad hoc tools, as applicable to piston rods as they are to patients.

Consider the Bonferroni adjustment that cardiovascular investigators are commonly required to use to adjust a *P* value for the number of statistical hypothesis tests. This tool was not derived from clinical trial methodology, but rather from the 1920s financial world, and lay dormant for years. It was resuscitated in the 1950s by Dr Olive Dunn,[12] who applied it to the management of sampling error in clinical research. Although it was a helpful improvisation, little thought was given to its long-term consequences on clinical research.

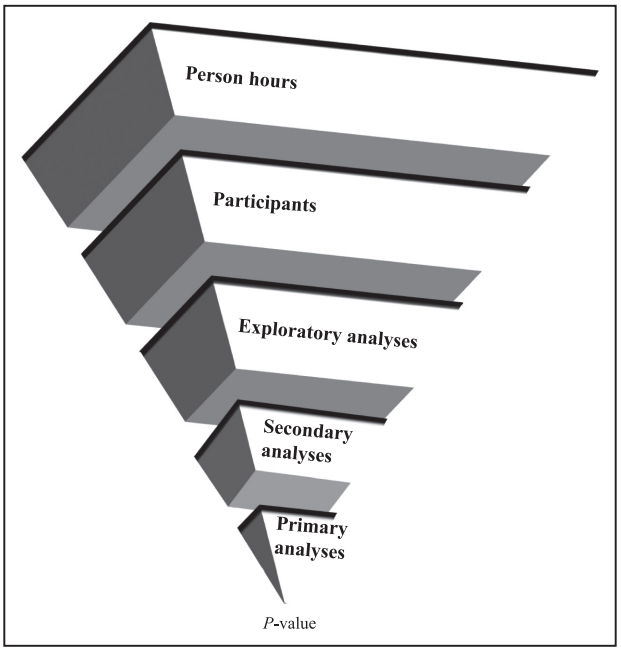

Figure. The clinical trial effort resting on a *P* value.

There is nothing in elementary clinical trial methodology that directs how sampling error must be managed, a situation that can and has generated alternative approaches that do not focus on traditional statistical hypothesis testing, for example, credibility and prediction intervals, Bayesian methods, and other modeling approaches. The father of clinical trials, Bradford Hill, thought that the presence of a contemporary control group, randomization, and a degree of blinding would make the interpretation of a clinical trial self-evident.[13] Thus, statistical hypothesis testing was injected into clinical trials to answer a question that Dr Hill thought had already been answered.

Where Do We Go From Here?

The traditional *P* value–centric approach may be admissible for the foreseeable future in the regulatory community where there is a concern about the use of a drug in the larger population that bears the cost and brunt of that drug's adverse effects profile. However, publically funded research, whose metric is not product but knowledge, is hampered by the *P* value and the confining interpretative structure that it has spawned. Many statisticians recognize this unsatisfactory state of affairs and are working on alternative solutions.

Two commonly used arguments raised in the *P* value's defense are as follows: (1) the reproducibility crisis in modern medical research, and (2) if not *P* values, then what? However, the reproducibility issue in healthcare research is not about *P* values but about the epidemiological definition of consistency, that is, the degree to which researchers can expect a common answer to the same question asked by different researchers studying differing populations with different research designs. The second argument in defense of *P* values has already been answered; the alternative to *P* values is the integration of (1) research methodology, (2) effect size, and (3) that effect size's variability into an assessment of the risk and benefit of the intervention to which each analysis contributes. Up to this point, that assessment has been cerebral, and in many instances, unfortunately pushed aside in the haste to statistical significance.

The time for this assessment's quantification is here. We do this not by beginning with the goal of the statistician which is to manage sample-to-sample variability and apply that to decision making. Instead, we begin with the goal of the investigator which is to learn all that is learnable from the research effort's data on a specific question. Publication of complete summary-level analytic data regardless of the *P* value and National Institutes of Health's commitment to data sharing is consistent with this view; trial data are collected for good reason and should be used to good effect. Quantification can then follow based on (1) reliance on mathematical and statistical theory, (2) the incorporation of sampling error that does not involve statistical hypothesis testing, and (3) the need for straightforward implementation.

Statistical hypothesis testing has served a purpose. There was a dearth of tools available in the 1950s and the administrative need for objective methods in clinical research appeared critical. But any benefit from this mid-20th century decision comes at too great a price. A new model of flexible, efficient, and adaptable cardiovascular research will require methodologic rigor yet contain intellectual freedom; this is not granted under the current statistical hypothesis-testing regime.

Administrators and biostatisticians have become comfortable with *P* values. It is the cardiovascular researcher, who is (1) forced to make a Sophie's choice from candidate primary end points because of statistical hypothesis-testing demands, and (2) then struggles to publish illuminating findings with large *P* values that suffers under them. It is, therefore, up to the cardiovascular research community to discard statistical hypothesis testing and demand better analysis tools, permitting us to learn and share all that we can from data commonly collected and funded at the public's expense. After >90 years, is it still not clear that the only way to free ourselves from the *P* value is to simply walk away from it?

Disclosures

None.

References

1. Fisher RA. *Statistical Methods for Research Workers*. Edinburg, Scotland: Oliver and Boyd; 1925.
2. Fisher RA. The arrangement of field experiments. *J Ministry Agric*. 1926; 33:503–513.
3. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Statistician*. 2016;70:129–133.
4. Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med*. 1999;130:995–1004.
5. Moye' LA. *Statistical Reasoning in Medicine—the P Value Primer*. 2nd ed. New York, NY: Springer; 2006.
6. Fisher LD, Moyé LA. Carvedilol and the Food and Drug Administration approval process: an introduction. *Control Clin Trials*. 1999;20:1–15.
7. Moyé LA. End-point interpretation in clinical trials: the case for discipline. *Control Clin Trials*. 1999;20:40–49; discussion 50.
8. Desai AS, Pfeffer MA. Beyond the P-value and the sound bite: learning from 'negative' clinical trials. *Eur Heart J*. 2017;38:2349–2351. doi: 10.1093/eurheartj/ehx395.
9. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987;77: 195–199.
10. Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology*. 1998;9:7–8.
11. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2:6–10.
12. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc*. 1961;56:52–64.
13. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.

Key Words: health services research ■ methods ■ probability ■ statistics