VIEWPOINT

The Proposal to Lower *P* Value Thresholds to .005

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Meta-Research Innovation Center at Stanford, Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University, Stanford, California.

P values and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less.[1] However, many of the claims that these reports highlight are likely false.[2] Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published[3] a statement on *P* values in 2016. The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributors to the ASA statement also wrote 20 independent, accompanying commentaries focusing on different aspects and prioritizing different solutions. Another large coalition of 72 methodologists recently proposed[4] a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P values are misinterpreted, overtrusted, and misused. The language of the ASA statement enables the dissection of these 3 problems. Multiple misinterpretations of *P* values exist, but the most common one is that they represent the "probability that the studied hypothesis is true."[3] A *P* value of .02 (2%) is wrongly considered to mean that the null hypothesis (eg, the drug is as effective as placebo) is 2% likely to be true and the alternative (eg, the drug is more effective than placebo) is 98% likely to be correct. Overtrust ensues when it is forgotten that "proper inference requires full reporting and transparency."[3] Better-looking (smaller) *P* values alone do not guarantee full reporting and transparency. In fact, smaller *P* values may hint to selective reporting and nontransparency. The most common misuse of the *P* value is to make "scientific conclusions and business or policy decisions" based on "whether a *P* value passes a specific threshold" even though "a *P* value, or statistical significance, does not measure the size of an effect or the importance of a result," and "by itself, a *P* value does not provide a good measure of evidence."[3]

These 3 major problems mean that passing a statistical significance threshold (traditionally $P = .05$) is wrongly equated with a finding or an outcome (eg, an association or a treatment effect) being true, valid, and worth acting on. These misconceptions affect researchers, journals, readers, and users of research articles, and even media and the public who consume scientific information. Most claims supported with *P* values slightly below .05 are probably false (ie, the claimed associations and treatment effects do not exist). Even among those claims that are true, few are worth acting on in medicine and health care.

Lowering the threshold for claiming statistical significance is an old idea. Several scientific fields have carefully considered how low a *P* value should be for a research finding to have a sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analyses are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently arrived at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analyses are nonsystematic and nontransparent. For most observational exploratory research that lacks preregistered protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research and randomized trials. Even though it is now more common to have a preexisting protocol and statistical analysis plan and preregistration of the trial posted on a public database, there are still substantial degrees of freedom regarding how to analyze data and outcomes and what exactly to present. In addition, many studies in contemporary clinical investigation focus on smaller benefits or risks; therefore, the risk of various biases affecting the results increases.

Moving the *P* value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the category of just "suggestive."[1] This shift is essential for those who believe (perhaps crudely) in black and white, significant or nonsignificant categorizations. For the vast majority of past observational research, this recategorization would be welcome. For example, mendelian randomization studies show that only few past claims from observational studies with $P < .05$ represent causal relationships.[5] Thus, the proposed reduction in the level for declaring statistical significance may dismiss mostly noise with relatively little loss of valuable information. For randomized trials, the proportion of true effects that emerge with *P* values in the window from .005 to .05 will be higher, perhaps the majority in several fields. However, most findings would not represent treatment effects that are large enough for outcomes that are serious enough to make them worthy of further action. Thus, the reduction in the *P* value threshold may largely do more good than harm, despite also removing an occasional true and useful treatment effect from the coveted significance zone. Regardless, the need for also focusing on the magnitude of all treatment effects and their uncertainty (such as with confidence intervals) cannot be overstated.

Lowering the threshold of statistical significance is a temporizing measure. It would work as a dam that could

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, 1265 Welch Rd, Medical School Office Building, Room X306, Stanford, CA 94305 (jioannid@stanford.edu).

Table. Various Proposed Solutions for Improving Statistical Inference on a Large Scale

	Apply to Past Literature: Easy or Fast Solution?	Apply to Future Research and Publications: Easy or Fast Solution?
Lower *P* value thresholds	A rather simple temporizing solution	Has potential collateral harms (see text) and success depends on adoption or enforcement by stakeholders (eg, journals, funders, societies)
Abandon *P* value thresholds and instead use exact *P* value	Many published *P* values have only been reported with thresholds	Success depends on extent of adoption or enforcement by stakeholders
Abandon *P* values entirely	Not easy because often nothing or little else has been provided; many articles did not report effect sizes and most lacked confidence intervals *P* values are still a good choice for some research applications	Previous pleas have not been successful to gain traction May succeed more easily in some fields (eg, assessment of diagnostic performance or choosing of predictors for prognostic models in which use of *P* values makes little or no sense)
Use alternative inference methods (eg, Bayesian statistics)	Partly doable (eg, one may convert *P* values to Bayes factors, but needs sophisticated training)	Would be suitable for most studies; increase in use of Bayesian methods (and other inferential approaches such as false-discovery rates) has been substantial recently, but would need to accelerate in the future
Focus on effect sizes and their uncertainty	Often not reported at all, but has become more common in more recent literature, particularly in clinical trials and meta-analyses	Relevant to the vast majority of the clinical literature, should be heavily endorsed as more directly linked to decision making, and it may be easier to promote than more sophisticated solutions
Train the scientific workforce	Takes time and major commitment to achieve sufficient statistical literacy.	Can lead to a more definitive solution, choosing fit for purpose statistics and inference tools, but may require major recasting of training priorities in curricula
Address biases that lead to inflated results	Requires major training; biases are often impossible to detect from published reports	Preemptively dealing with biases is ideal, but needs concerted commitment of multiple stakeholders to promote and incentivize better research practices

help gain time and prevent drowning by a flood of statistical significance, while promoting better, more-durable solutions.[6] These solutions may involve abandoning statistical significance thresholds or *P* values entirely. If any thresholds are to continue in use, even lower thresholds are probably preferable for most observational research. Comprehensive reviews (termed *umbrella reviews*) that have evaluated multiple systematic reviews of observational studies propose a $P < 10^{-6}$ threshold.[5] In addition, falsification end-point methods (ie, using such *P* value thresholds that almost all well-established null associations will not be able to pass them) also point to very low *P* values.[7] With the advent of big data, statistical significance will increasingly mean very little because extremely low *P* values are routinely obtained for signals that are too small to be useful even if true.

Adopting lower *P* value thresholds may help promote a reformed research agenda with fewer, larger, and more carefully conceived and designed studies with sufficient power to pass these more demanding thresholds. However, collateral harms may also emerge. Bias may escalate rather than decrease if researchers and other interested parties (eg, for-profit sponsors) try to find ways to make the results have lower *P* values. Selected study end points may become even less clinically relevant because it is easier to reach lower *P* values with weak surrogate end points than with hard clinical outcomes. Moreover, results that pass a lower *P* value threshold may be limited by greater regression to the mean and new discoveries may have even more exaggerated effect sizes than before.

Because the proposed threshold of $P < .005$ is imperfect, other more difficult but more durable alternative solutions should also be contemplated (**Table**). These solutions vary based on how quickly and easily they can be adopted. They can target the use and interpretation of the past biomedical literature accumulated to date or the design and deployment of the new literature that will accumulate in the future. The situation is dire for the past literature because there is no perfect remedy after the fact. In the long-term, the scientific workforce will need to be more properly trained in using the best fit for purpose statistical inference tools and biases will need to be addressed preemptively rather than retrospectively. However, these may continue to be largely unachievable goals.

Data are becoming more complex. If time for rigorous training in methods and statistics for researchers and for research users remains limited, subpar medical statistics and concomitant misinterpretations may continue. Nevertheless, hopefully several fields will adopt better standards for *P* values, will decrease their dependence on *P* values, and enhance the adoption of other useful inferential tools (eg, Bayesian statistics) when appropriate. The rapidity and extent of these changes is unpredictable. Low adoption in the past may cause some pessimism. However, a fresh start and a rapid acceleration of adoption of better practices is always possible. Incentives from major journals and funders as well as radical changes in training curricula may be necessary to achieve more widespread and effective shifts.

ARTICLE INFORMATION

Published Online: March 22, 2018. doi:10.1001/jama.2018.1536

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Ioannidis reported being a member of the panel working on the American Statistical Association statement and an author of the article proposing decreasing the threshold of statistical significance.

Funding/Support: The Meta-Research Innovation Center at Stanford has been supported by the Laura and John Arnold Foundation. The work of Dr Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell.

Role of the Sponsors: The sponsors had no role in the preparation, review, or approval of the manuscript or decision to submit the manuscript for publication.

REFERENCES

1. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA*. 2016;315(11):1141-1148.

2. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

3. Wasserstein RL, Lazar NA. The ASA's statement on *P*-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-133.

4. Benjamin DJ, Berger JO, Johnson VE, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2:6-10.

5. Li X, Meng X, Timofeeva M, et al. Serum uric acid levels and multiple health outcomes. *BMJ*. 2017;357:j2376.

6. Resnick B. What a nerdy debate about *P* values shows about science-and how to fix it. https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005. Accessed February 1, 2018.

7. Prasad V, Jena AB. Prespecified falsification end points. *JAMA*. 2013;309(3):241-242.