VIEWPOINT

The Importance of Predefined Rules and Prespecified Statistical Analyses
Do Not Abandon Significance

John P. A. Ioannidis, MD, DSc
Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California; and Meta-Research Innovation Center–Berlin (METRIC-B), Berlin, Germany.

For decades, statisticians and clinicians have debated the meaning of statistical and clinical significance. In general, most journals remain married to the frequentist approach to statistical testing and using the term *statistical significance*. A recent proposal to ban statistical significance gained campaign-level momentum in a commentary with 854 recruited signatories.[1] The petition proposes retaining *P* values but abandoning dichotomous statements (significant/nonsignificant), suggests discussing "compatible" effect sizes, denounces "proofs of the null," and points out that "crucial effects" are dismissed on discovery or refuted on replication because of nonsignificance. The proposal also indicates that "we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes zero,"[1] and that categorization based on other statistical measures (eg, Bayes factors) should be discouraged. Other recent articles have also addressed similar topics, with an entire supplemental issue of a statistics journal devoted to issues related to *P* values.[2]

Changing the approach to defining statistical and clinical significance has some merits; for example, embracing uncertainty, avoiding hyped claims with weak statistical support, and recognizing that "statistical significance" is often poorly understood. However, technical matters of abandoning statistical methods may require further thought and debate. Behind the so-called war on significance lie fundamental issues about the conduct and interpretation of research that extend beyond (mis) interpretation of statistical significance. These issues include what effect sizes should be of interest, how to replicate or refute research findings, and how to decide and act based on evidence. Inferences are unavoidably dichotomous—yes or no—in many scientific fields ranging from particle physics to agnostic omics analyses (ie, massive testing of millions of biological features without any a priori preference that one feature is likely to be more important than others) and to medicine. Dichotomous decisions are the rule in medicine and public health interventions. An intervention, such as a new drug, will either be licensed or not and will either be used or not.

> Significance (not just statistical) is essential both for science and for science-based action, and some filtering process is useful to avoid drowning in noise.

Some fields of investigation are richer than others in effects remaining to be discovered. Moreover, some fields have several effects large enough to discover and act on, whereas others struggle with mostly tiny effect sizes. The latter scenario is becoming more common. For example, tens of thousands of genome-wide significant associations have emerged for hundreds of different phenotypes, but the vast majority explain less than 0.05% of the variance of the trait of interest.[3] Some fields that claim to work with large, actionable effects (eg, nutritional epidemiology) may simply have larger, uncontrolled biases.

Sometimes there are different perspectives about the presence, frequency, and magnitude of non-null effects in the same field. For example, what percentage of nutrients affect cancer risk? Some skeptics dismiss results even if they have small *P* values or large Bayes factors. Conversely, for some enthusiasts of nutritional carcinogenesis, the weakest signals would seem strong and worthy of global action.

Some skeptics maintain that there are few actionable effects and remain reluctant to endorse belabored policies and useless (or even harmful) interventions without very strong evidence. Conversely, some enthusiasts express concern about inaction, advocate for more policy, or think that new medications are not licensed quickly enough. Some scientists may be skeptical about some research questions and enthusiastic about others. The suggestion to abandon statistical significance[1] espouses the perspective of enthusiasts: it raises concerns about unwarranted statements of "no difference" and unwarranted claims of refutation but does not address unwarranted claims of "difference" and unwarranted denial of refutation.

Interpretations go beyond statistics. They also vary depending on what other (eg, mechanistic) evidence is considered relevant. However, determination of the relevance of qualitative or triangulating types of evidence can be substantially subjective. The statistical data analysis is often the only piece of evidence processing that has a chance of being objectively assessed before experts, professional societies, and governmental agencies begin to review the data and make recommendations. This means that, ideally, the statistical analysis should use carefully prethought, rigorous probes. Similarly, a replication study ("reproducibility check")[4] may be carefully prespecified, conducting rigorous tests of success or failure (replication or refutation). When the analyses are preplanned, clear, and followed carefully, such tests are useful. Interpretation of any result is far more complicated than just significance testing, but it is a starting point.

Absent prespecified rules, most research designs and analyses have enough leeway to manipulate the data and

Corresponding Author: John P. A. Ioannidis, MD, DSc, Department of Medicine and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, Medical School Office Bldg, Room X306, Stanford, CA 94305 (jioannid@stanford.edu).

hack the results to claim important signals. Passing the threshold of "statistical significance" has been a traditional goal in this regard. A low barrier such as $P < .05$ is typically too easy to pass. Hence, one option is making the barrier more demanding[5]; many fields (eg, molecular and genetic epidemiology) have already done this by using genome-wide significance levels ($P < 10^{-9}$) or very strict false discovery rate thresholds. The proposal to entirely remove the barrier does not mean that scientists will not often still wish to interpret their results as showing important signals and fit preconceived notions and biases.[6] With the gatekeeper of statistical significance, eager investigators whose analyses yield, for example, $P = .09$ have to either manipulate their statistics to get to $P < .05$ or add spin to their interpretation to suggest that results point to an important signal through an observed "trend." When that gatekeeper is removed, any result may be directly claimed to reflect an important signal or fit to a preexisting narrative. Moreover, refutation of an early study by a subsequent replication effort can always be denied.

Many fields of investigation (ranging from bench studies and animal experiments to observational population studies and even clinical trials) have major gaps in the ways they conduct, analyze, and report studies and lack protection from bias. Instead of trying to fix what is lacking and set better and clearer rules, one reaction is to overturn the tables and abolish any gatekeeping rules (such as removing the term *statistical significance*). However, potential for falsification is a prerequisite for science. Fields that obstinately resist refutation can hide behind the abolition of statistical significance but risk becoming self-ostracized from the remit of science.

Significance (not just statistical) is essential both for science and for science-based action, and some filtering process is useful to avoid drowning in noise. Statistical significance with $P < .05$ is a weak, easily abused filter. Better and less gameable filters and more appropriate fit-for-purpose statistical methods are needed. Whatever these filters are (frequentist, Bayesian, or false discovery rates), they should be carefully considered in advance of a study. The rules of the analysis should be carefully predefined whenever possible. Statistical analysis plans are rarely specified in sufficient detail, even for study designs such as randomized trials, for which protocols are otherwise preregistered.[7] In a recent survey completed by 390 consulting statisticians, a large percentage perceived that they had received inappropriate requests from investigators to analyze data in ways that obtain desirable results.[8] Studies have shown that unless an analysis is prespecified, analytical choice (eg, different adjustments for covariates in nonrandomized studies) may allow obtaining a wide range of results.[9] With current big data, this huge "vibration of effects" is the norm. Whenever the objectives and prespecified end points of a study are known, statistical analyses can be largely predetermined and registered, and the rules of how results will be read should also be judiciously preset and transparent. Deviations may be justified because of unexpected circumstances (eg, if unexpected amounts of missing data emerge), but these should be documented, with choices explained and robustness of conclusions to different sensitivity analyses assessed. Making raw data available could further enhance trust.

Clinical, monetary, and other considerations may often have more importance than statistical findings. However, these issues are often well known in advance. If so, they should be carefully addressed in designing the best, most informative studies by preemptively accounting for these considerations. The statistical analysis and rules of statistical interpretation (including potential thresholds) can be specified in advance, incorporating these considerations. More thought should go into research before it is conducted, not after the data have been inspected.

Much research will remain highly exploratory, and this should be declared as such when results are presented.[10] However, even for exploratory research, there is an advantage in having some agreement about default statistical analysis and interpretation. Deviations from the default would then be easier to spot and questioned as to their appropriateness. For most research questions, post hoc analytical manipulation is unlikely to lead closer to the truth than a default analysis with a basic set of rules. All studies in the same field can follow the default options first before venturing into creative data dredging.

The statistical numeracy of the scientific workforce requires improvement. Banning statistical significance while retaining *P* values (or confidence intervals) will not improve numeracy and may foster statistical confusion and create problematic issues with study interpretation, a state of statistical anarchy. Uniformity in statistical rules and processes makes it easier to compare like with like and avoid having some associations and effects be more privileged than others in unwarranted ways. Without clear rules for the analyses, science and policy may rely less on data and evidence and more on subjective opinions and interpretations.

ARTICLE INFORMATION

Published Online: April 4, 2019. doi:10.1001/jama.2019.4582

Conflict of Interest Disclosures: None reported.

Funding/Support: METRICS has been funded by a grant from the Laura and John Arnold Foundation. METRIC-B has been funded by an Einstein fellowship from the Stiftung Charité and the Einstein Stiftung.

Role of the Funder/Sponsor: The funders listed above had no role in the preparation, review, or approval of the manuscript or decision to submit the manuscript for publication.

REFERENCES

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9

2. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond *p*<0.05. *Am Stat*. 2019;73:1-19. doi:10.1080/00031305.2019.1583913

3. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Genet*. 2018;50(11):1593-1599. doi:10.1038/s41588-018-0248-z

4. Nosek BA, Errington TM. Making sense of replications. *Elife*. 2017;6:e23383. doi:10.7554/eLife.23383

5. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA*. 2018;319(14):1429-1430. doi:10.1001/jama.2018.1536

6. Ioannidis JPA. Retiring statistical significance would give bias a free pass. *Nature*. 2019;567(7749):461. doi:10.1038/d41586-019-00969-2

7. Localio AR, Stack CB, Meibohm AR, et al. Inappropriate statistical analysis and reporting in medical research: perverse incentives and institutional solutions. *Ann Intern Med*. 2018;169(8):577-578. doi:10.7326/M18-2516

8. Wang MQ, Yan AF, Katz RV; Survey of Consulting Biostatisticians. Researcher requests for inappropriate analysis and reporting: a US survey of consulting biostatisticians. *Ann Intern Med*. 2018;169(8):554-558. doi:10.7326/M18-1230

9. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68(9):1046-1058. doi:10.1016/j.jclinepi.2015.05.029

10. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12(5):e1001863. doi:10.1371/journal.pbio.1001863