Received: 8 October 2019 | Accepted: 8 October 2019

DOI: 10.1111/eci.13176

COMMENTARY

WILEY

Statistical significance gives bias a free pass

Valentin Amrhein[1] | **Sander Greenland**[2] | **Blakeley B. McShane**[3]

[1]Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland

[2]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA

[3]Kellogg School of Management, Northwestern University, Evanston, IL, USA

Correspondence
Valentin Amrhein, Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland.
Email: v.amrhein@unibas.ch

Whether or not "the foundations and the practice of statistics are in turmoil",[1] it is wise to question methods whose misuse has been lamented for over a century.[2-4] Perhaps the most widespread misuse of statistics is taking the crossing of some threshold as license for declaring "statistical significance" and for generalizing from a single study. Such generalized conclusions are often taken up by science communicators, media and political stakeholders without recognition of their uncertainty. A major consequence is flip-flopping headlines such as "chocolate is good for you" followed by "chocolate is bad for you".[5] No wonder, only about a third of over 2000 respondents in a survey on the British public said they would trust data from medical trials.[6]

Often, it is better to simply describe observed associations and their uncertainties (eg by giving point and interval estimates and plotting raw data). If inference to some target population is required, it typically suffices to suggest a range of values that are highly compatible with the data and modelling assumptions—for example, by explicitly interpreting both endpoints of interval estimates and noting that such intervals likely understate the degree of uncertainty.[7]

A call to describe observed associations does *not* grant a "free pass" to report results from single studies as revealing some general truth. Instead, it encourages honest description of all results and humility about conclusions, thereby reducing selection and publication biases. The aim of single studies should be to report uncensored information that can later be used to make more general conclusions based on cumulative evidence from multiple studies.

In contrast to Ioannidis,[8] we and others[9-15] hold that it is using—not retiring—statistical significance as a "filtering process" or "gatekeeper"[16] that "gives bias a free pass".[8] As has been known for decades, statistically significant estimates are biased away from the null and statistically nonsignificant estimates are biased towards the null. Therefore, any discussion that focuses on estimates chosen for their statistical significance or nonsignificance will be biased.

Not only does statistical (non)significance introduce bias, but also it fails to address various biases that can afflict studies. As any survey research textbook will confirm, those who choose to respond to a survey typically differ from those who choose not to—whether, for example, the British survey respondents[6] discussed above or those of Hardwicke and Ioannidis.[17] Raw results from such surveys are biased and can mislead about the target population. Statistical significance cannot detect or adjust for those or other biases and thus relying on it gives bias a free pass.

The biases produced by selecting results for their statistical significance or nonsignificance arise at all steps in scientific research, including decisions about what to include in models, discuss in papers, accept for publication and emphasize in editorials, reviews and popular reports. Such biases arise not only from the use of *P*-value thresholds but also from the use of Bayes factor (or any other) thresholds, as well as from focusing on whether or not interval estimates include some null value.

Statistics from single studies are often better reported as compact summaries of relations in the data, not as inferences about some (often ill-defined) target population—in other words, inferential statistics should be treated as descriptive statistics.[13] Authors should write sentences like "we found a risk ratio of 1.20 (95% CI: [0.80, 1.80]; $P = .38$)" without being criticized for overstating the evidence—as long as they do not claim general conclusions; and they should be criticized for misrepresenting their results, for example, as "our study shows there is an increased risk" or "our study shows there is no association".[7]

We also disagree that "abandoning the concept of statistical significance would make claims of 'irreproducibility'

difficult if not impossible to make".[17] In reality, it is difficult if not impossible to make claims of 'irreproducibility' based on statistical significance vs nonsignificance. For example, Ioannidis and Lau[18] summarized 32 studies on antibiotic prophylaxis in colon surgery. Although only about half the studies attained statistical significance, this does not mean the effect was irreproducible: the cumulative evidence across the 32 studies strongly suggested that antibiotic prophylaxis was effective. This is one of many examples of meta-analysis illustrating a key point of the excellent paper by Goodman, Fanelli and Ioannidis,[19] who noted that after an initial statistically significant result, "the failure to observe a significant result in a second experiment of similar design is to be expected and cannot be used as a criterion to undermine the credibility of the first experiment," and that "a preferred way to assess the evidential meaning of two or more results with substantive stochastic variability is to evaluate the cumulative evidence they provide vis-á-vis a hypothesis of interest and not whether one contradicts or discredits the other through the lens of statistical significance".

REFERENCES

1. Gelman A. When we make recommendations for scientific practice, we are (at best) acting as social scientists. *Eur J Clin Invest*. 2019;49:e13165.
2. Pearson K. Note on the significant or non-significant character of a sub-sample drawn from a sample. *Biometrika*. 1906;5:181-183.
3. Boring EG. Mathematical vs. scientific significance. *Psychol Bull*. 1919;16:335-338.
4. Fisher RA. Statistical tests. *Nature*. 1935;136:474.
5. Oxman AD, Aronson JK, Barends E, et al. Key concepts for making informed choices. *Nature*. 2019;572:303-306.
6. ComRes. Academy of medical sciences: medical information survey. 2016; http://www.acmedsci.ac.uk/evidence/survey. Accessed October 24, 2019.
7. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature*. 2019;567:305-307.
8. Ioannidis JPA. Retiring statistical significance would give bias a free pass. *Nature*. 2019;567:461.
9. Lane DM, Dunlap WP. Estimating effect size: bias resulting from the significance criterion in editorial decisions. *Br J Math Stat Psychol*. 1978;31:107-112.
10. Gelman A, Carlin J. Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. *Pers Psychol Sci*. 2014;9:641-651.
11. McShane BB, Gal D. Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Manage Sci*. 2016;62:1707-1718.
12. Greenland S. Invited commentary: the need for cognitive science in methodology. *Am J Epidemiol*. 2017;186:639-645.
13. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *Am Stat*. 2019;73(sup1):262-270.
14. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat*. 2019;73(sup1):235-245.
15. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p<0.05". *Am Stat*. 2019;73(sup1):1-19.
16. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA*. 2019;321:2067-2068.
17. Hardwicke TE, Ioannidis JPA. Petitions in scientific argumentation: dissecting the request to retire statistical significance. *Eur J Clin Invest*. 2019;49:e13162.
18. Ioannidis JPA, Lau J. State of the evidence: current status and prospects of meta-analysis in infectious diseases. *Clin Infect Dis*. 1999;29:1178-1185.
19. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med*. 2016;8:341.