COMMENT

ILLUSTRATION BY DAVID PARKINS

Retire statistical significance

Valentin Amrhein, **Sander Greenland**, **Blake McShane** and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

PERVASIVE PROBLEM

Let's be clear about what must stop: 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. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.

For example, consider a series of analyses of unintended effects of anti-inflammatory drugs[2]. Because their results were statistically non-significant, one set of researchers concluded that exposure to the drugs was "not associated" with new-onset atrial fibrillation (the most common disturbance to heart rhythm) and that the results stood in contrast to those from an earlier study with a statistically significant outcome.

Now, let's look at the actual data. The researchers describing their statistically non-significant results found a risk ratio of 1.2 (that is, a 20% greater risk in exposed patients relative to unexposed ones). They also found a 95% confidence interval that spanned everything from a trifling risk decrease of 3% to a considerable risk increase of 48% ($P = 0.091$; our calculation). The researchers from the earlier, statistically significant, study found the exact same risk ratio of 1.2. That study was simply more precise, with an interval spanning from 9% to 33% greater risk ($P = 0.0003$; our calculation).

It is ludicrous to conclude that the statistically non-significant results showed "no association", when the interval estimate included serious risk increases; it is equally absurd to claim these results were in contrast with the earlier results showing an identical observed effect. Yet these common practices show how reliance on thresholds of statistical significance can mislead us (see 'Beware false conclusions').

These and similar errors are widespread. Surveys of hundreds of articles have found that statistically non-significant results are interpreted as indicating 'no difference' or 'no effect' in around half (see 'Wrong interpretations' and Supplementary Information).

In 2016, the American Statistical Association released a statement in *The American Statistician* warning against the misuse of statistical significance and *P* values. The issue also included many commentaries on the subject. This month, a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond $P < 0.05$'. The editors introduce the collection with the caution "don't say 'statistically significant'"[3]. Another article[4] with dozens of signatories also calls on authors and journal editors to disavow those terms.

We agree, and call for the entire concept of statistical significance to be abandoned.

> *"Eradicating categorization will help to halt overconfident claims, unwarranted declarations of 'no difference' and absurd statements about 'replication failure'."*

We are far from alone. When we invited others to read a draft of this comment and sign their names if they concurred with our message, 250 did so within the first 24 hours. A week later, we had more than 800 signatories — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling (see the list and final count of signatories in the Supplementary Information). These include statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries and across all continents except Antarctica. One advocate called it a "surgical strike against thoughtless testing of statistical significance" and "an opportunity to register your voice in favour of better scientific practices".

We are not calling for a ban on *P* values. Nor are we saying they cannot be used as a decision criterion in certain specialized applications (such as determining whether a manufacturing process meets some quality-control standard). And we are also not advocating for an anything-goes situation, in which weak evidence suddenly becomes credible. Rather, and in line with many others over the decades, we are calling for a stop to the use of *P* values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis[5].

QUIT CATEGORIZING
The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different[6–8]. The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.

Unfortunately, the false belief that crossing the threshold of statistical significance is enough to show that a result is 'real' has led scientists and journal editors to privilege such results, thereby distorting the literature. Statistically significant estimates are biased upwards in magnitude and potentially to a large degree, whereas statistically non-significant estimates are biased downwards in magnitude. Consequently, any discussion that focuses on estimates chosen for their significance will be biased. On top of this, the rigid focus on statistical significance encourages researchers to choose data and methods that yield statistical significance for some desired (or simply publishable) result, or that yield statistical non-significance for an undesired result, such as potential side effects of drugs — thereby invalidating conclusions.
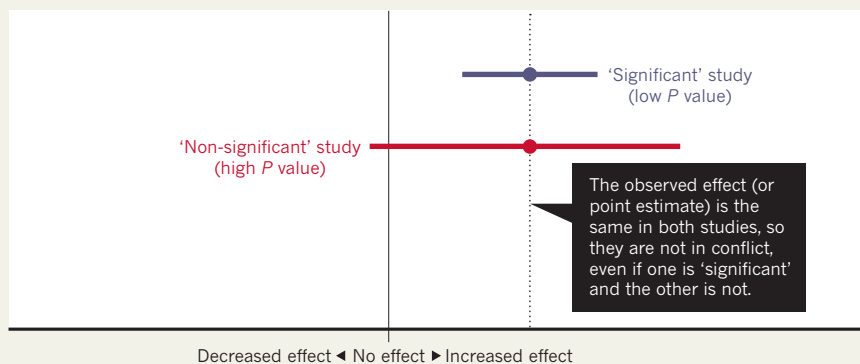
The pre-registration of studies and a commitment to publish all results of all analyses can do much to mitigate these issues. However, even results from pre-registered studies can be biased by decisions invariably left open in the analysis plan[9]. This occurs even with the best of intentions.

Again, we are not advocating a ban on *P* values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors.

One reason to avoid such 'dichotomania' is that all statistics, including *P* values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree. In fact, random variation alone can easily lead to large disparities in *P* values, far beyond falling just to either side of the 0.05 threshold. For example, even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $P < 0.05$, it would not be very surprising for one to obtain $P < 0.01$ and the other $P > 0.30$.

BEWARE FALSE CONCLUSIONS
Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

'Significant' study (low *P* value)

'Non-significant' study (high *P* value)

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is 'significant' and the other is not.

Decreased effect ◀ No effect ▶ Increased effect

Whether a *P* value is small or large, caution is warranted.

We must learn to embrace uncertainty. One practical way to do so is to rename confidence intervals as 'compatibility intervals' and interpret them in a way that avoids overconfidence. Specifically, we recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits. In doing so, they should remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval[7,10]. Therefore, singling out one particular value (such as the null value) in the interval as 'shown' makes no sense.

We're frankly sick of seeing such nonsensical 'proofs of the null' and claims of non-association in presentations, research articles, reviews and instructional materials. An interval that contains the null value will often also contain non-null values of high practical importance. That said, if you deem all of the values inside the interval to be practically unimportant, you might then be able to say something like 'our results are most compatible with no important effect'.
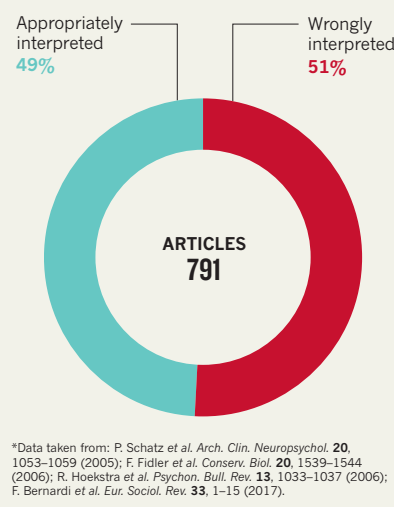
When talking about compatibility intervals, bear in mind four things. First, just because the interval gives the values most compatible with the data, given the assumptions, it doesn't mean values outside it are incompatible; they are just less compatible. In fact, values just outside the interval do not differ substantively from those just inside the interval. It is thus wrong to claim that an interval shows all possible values.

Second, not all values inside are equally compatible with the data, given the assumptions. The point estimate is the most compatible, and values near it are more compatible than those near the limits. This is why we urge authors to discuss the point estimate, even when they have a large *P* value or a wide interval, as well as discussing the limits of that interval. For example, the authors above could have written: 'Like a previous study, our results suggest a 20% increase in risk of new-onset atrial fibrillation in patients given the anti-inflammatory drugs. Nonetheless, a risk difference ranging from a 3% decrease, a small negative association, to a 48% increase, a substantial positive association, is also reasonably compatible with our data, given our assumptions.' Interpreting the point estimate, while acknowledging its uncertainty, will keep you from making false declarations of 'no difference', and from making overconfident claims.

Third, like the 0.05 threshold from which it came, the default 95% used to compute intervals is itself an arbitrary convention. It is based on the false idea that there is a 95% chance that the computed interval itself contains the true value, coupled with the vague

WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

Appropriately interpreted
49%

Wrongly interpreted
51%

ARTICLES
791

*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

feeling that this is a basis for a confident decision. A different level can be justified, depending on the application. And, as in the anti-inflammatory-drugs example, interval estimates can perpetuate the problems of statistical significance when the dichotomization they impose is treated as a scientific standard.

Last, and most important of all, be humble: compatibility assessments hinge on the correctness of the statistical assumptions used to compute the interval. In practice, these assumptions are at best subject to considerable uncertainty[7,8,10]. Make these assumptions as clear as possible and test the ones you can, for example by plotting your data and by fitting alternative models, and then reporting all results.

Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones. Inferences should be scientific, and that goes far beyond the merely statistical. Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as *P* values or intervals.

The objection we hear most against retiring statistical significance is that it is needed to make yes-or-no decisions. But for the choices often required in regulatory, policy and business environments, decisions based on the costs, benefits and likelihoods of all potential consequences always beat those made based solely on statistical significance. Moreover, for decisions about whether to pursue a research idea further, there is no simple connection between a *P* value and the probable results of subsequent studies.

What will retiring statistical significance look like? We hope that methods sections and data tabulation will be more detailed and nuanced. Authors will emphasize their estimates and the uncertainty in them — for example, by explicitly discussing the lower and upper limits of their intervals. They will not rely on significance tests. When *P* values are reported, they will be given with sensible precision (for example, $P = 0.021$ or $P = 0.13$) — without adornments such as stars or letters to denote statistical significance and not as binary inequalities ($P < 0.05$ or $P > 0.05$). Decisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking.

Our call to retire statistical significance and to use confidence intervals as compatibility intervals is not a panacea. Although it will eliminate many bad practices, it could well introduce new ones. Thus, monitoring the literature for statistical abuses should be an ongoing priority for the scientific community. But eradicating categorization will help to halt overconfident claims, unwarranted declarations of 'no difference' and absurd statements about 'replication failure' when the results from the original and replication studies are highly compatible. The misuse of statistical significance has done much harm to the scientific community and those who rely on scientific advice. *P* values, intervals and other statistical measures all have their place, but it's time for statistical significance to go. ∎

Valentin Amrhein *is a professor of zoology at the University of Basel, Switzerland.* **Sander Greenland** *is a professor of epidemiology and statistics at the University of California, Los Angeles.* **Blake McShane** *is a statistical methodologist and professor of marketing at Northwestern University in Evanston, Illinois. For a full list of co-signatories, see Supplementary Information.*
e-mail: v.amrhein@unibas.ch

1. Fisher, R. A. *Nature* **136**, 474 (1935).
2. Schmidt, M. & Rothman, K. J. *Int. J. Cardiol.* **177**, 1089–1090 (2014).
3. Wasserstein, R. L., Schirm, A. & Lazar, N. A. *Am. Stat.* https://doi.org/10.1080/00031305.2019.1583913 (2019).
4. Hurlbert, S. H., Levine, R. A. & Utts, J. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1543616 (2019).
5. Lehmann, E. L. *Testing Statistical Hypotheses* 2nd edn 70–71 (Springer, 1986).
6. Gigerenzer, G. *Adv. Meth. Pract. Psychol. Sci.* **1**, 198–218 (2018).
7. Greenland, S. *Am. J. Epidemiol.* **186**, 639–645 (2017).
8. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1527253 (2019).
9. Gelman, A. & Loken, E. *Am. Sci.* **102**, 460–465 (2014).
10. Amrhein, V., Trafimow, D. & Greenland, S. *Am. Stat.* https://doi.org/10.1080/00031305.2018.1543137 (2019).