No Estimation without Inference:

A Response to the International Society of Physiotherapy Journal Editors

Keith Lohse, PhD[1]

[1] Physical Therapy and Neurology, Washington University School of Medicine, Saint Louis, MO

Acknowledgments: I would like to thank Dr. Emma Johnson and Dr. Kristin Sainani for their detailed comments on earlier drafts of this commentary.

Date Submitted
2022-07-20
Keywords:
"physical therapy"; "statistical significance"; "inference"; "estimation"

Corresponding Author:
Keith Lohse, PhD, PStat; lohse@wustl.edu

Recently, Elkins et al.[1] (hereafter referred to as “the Editorial”) published an editorial on behalf of the International Society of Physiotherapy Journal Editors (ISPJE), recommending that researchers stop using null hypothesis significance tests and adopt “estimation methods”. Further, the editorial warns that this is not merely an idea to consider, but a coming policy of journals: “the [ISPJE] will be expecting manuscripts to use estimation methods ***instead*** of null hypothesis statistical tests” (emphasis added).

I commend the Editorial for encouraging researchers to think deeply about the statistical tools available to them, to consider “practical significance” as well as “statistical significance”, and for bringing important methodological discussions to the forefront of physical therapy research. However, the Editorial is also deeply flawed in its statistical reasoning. If these practices were adopted, they could damage the statistical literacy and scientific integrity of the field. I detail each of these critiques below, but in short the Editorial: (1) fails to adequately grapple with the inherent connection between hypothesis testing and estimation as methods of statistical inference, (2) presents several misleading arguments about the flaws of statistical significance tests, and (3) presents an alternative that is, in itself, a form of significance testing – the minimal effects test[2] (but this test is done implicitly and muddles two-sided and one-sided hypothesis testing). Finally, I end with a short list of more urgent problems that the ISPJE could work to address.

Hypothesis Testing and Estimation are Inescapably Intertwined

The Editorial presents hypothesis testing and estimation as two distinct methodological approaches. However, these approaches are two sides of the same coin, as illustrated by a simple example in Figure 1. When a 95% confidence interval excludes the null value then one can reject the null hypothesis at $p<.05$. This is because hypothesis tests and confidence intervals are based on the same underlying mathematics: e.g., how big is the observed effect relative to the variability we would expect due to sampling? Importantly, the null hypothesis can assume either zero or non-zero effects. So, as shown in the figure, we can ascertain the probability of observing the data we did, assuming a null value of 0 or a null value of 1.

Hypothesis testing and estimation cannot be fully disentangled: estimation (frequentist or Bayesian) asks about *plausible values* of the parameter in the population, hypothesis testing asks about the plausibility of *a specific parameter value*. These are both inferences, because we are inferring something about the population based on the data in our sample. In the frequentist paradigm, uncertainty in the inference is accounted for with long-run error control; e.g., setting the Type 1 Error rate, α=0.05. We can see this behavior Figure 1A-E: any confidence interval that does not contain zero also has p<0.05, for the null hypothesis significance test (NHST). Focusing on estimation, the 95% confidence interval shows values in the population that are *compatible* with what we observed in the sample.[3] That is, if you move outside of the confidence interval, any of those parameter values (the "true" mean differences Δ's) would be statistically different from the mean difference observed in the sample ($\overline{x_d}$) at the p<0.05 level. The p-value is the probability of observing data as extreme or more extreme, assuming that the null hypothesis is true: $p(\geq \overline{x_d}|H_0)$. Inside of the confidence interval, none of those parameter values would be statistically different (p>0.05) from the observed mean difference. Typically the null hypothesis significance test (NHST) assumes that the true value in the population is 0 (i.e., $H_0\colon \Delta = 0$). The further the sample mean difference is away from 0, the lower the probability of observing that sample mean (if the null hypothesis were true).

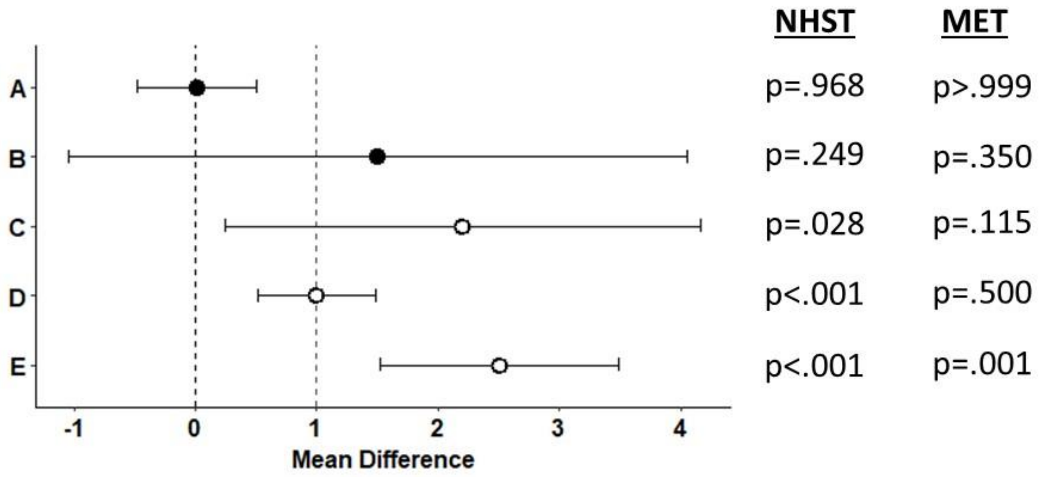

Figure 1. 95% confidence intervals and corresponding p-values for testing $H_0\colon \Delta = 0$ (NHST, null hypothesis significance testing) and $H_0\colon \Delta \leq 1$ (MET, a one-sided minimal effects test).

Importantly, the Editorial does not explicitly address the fact that we can set H_0 to be any value. For instance, rather than setting $H_0\colon \Delta = 0$ (sometimes referred to as the "nil-hypothesis")[4], we can set H_0 equal to any clinically meaningful value of interest. This is referred to as a minimal effects test (or minimum effect test, MET[2,5]). For the sake of argument, let's say this value is 1 in Figure 1. Comparing the confidence intervals to the new null value, you can see that any confidence intervals that only contain values larger than 1 also have a p<0.05 for the minimal effects test (i.e., Figure 1E).[A] Thus, we have both an inference about a specific hypothesis and an estimate in both the NHST and the MET[B], but the hypothesis test and the estimate are complementary and connected.

Misleading Arguments about flaws with Significance Tests

The Editorial bases many arguments on a previous list of perceived problems from Herbert (2019).[6] The Herbert paper is in itself an editorial that presents informed arguments, but is not an objective demonstration of any mathematical facts. So, reinforcing the Editorial's list through a citation to Herbert does not provide an evidentiary foundation: it is layering opinion on top of opinion. Second, each of the five "problems" outlined by the Editorial is either not really a problem inherent to p-values or the problem is a true but misleading statement. I address each problem from the Editorial (in quotes) below:

1. ***"A p-value is not the probability that a hypothesis is (or is not) true."*** – This is correct, but it does not follow that this makes p-values (or even statistical significance tests) unhelpful or uninformative. Knowing that the observed data are incompatible with the null hypothesis is a crucial step for many research questions.

[A] METs are typically directional, using one-sided hypothesis tests (e.g., $H_0 \leq 1$) whereas NSHTs are often non-directional, using two-sided hypothesis tests (e.g., $H_0 = 0$). Thus, although the confidence interval for Figure 1A does not contain the null value of 1, the whole of the confidence interval is below 1, thus yielding a non-significant minimal effects test.

[B] For convenience, I am referring to NHST and MET as separate tests. However, it is more accurate to think of the MET as type of NHST where you have a one-sided test of a non-zero null value. I use the different terms because readers are likely more familiar with the term NHST when referring to the specific case of $H_0 = 0$.[4]

2. *"A p-value does not constitute evidence"* – This is an oversimplification and misleading. The Editorial is correct that a single p-value cannot tell us about the probability of the null hypothesis being true, but p-values can be used as evidence against the null-hypothesis. In essence, this a question of absolute probability versus relative probabilities. For instance, if I find that eating green jelly beans reduces post-surgical recovery time by 10% relative to controls, $p<0.05$, then the most likely explanation is still that jelly beans have no effect on recovery. That is, the null hypothesis is still most likely explanation even though $p<0.05$. This is a silly example to show why a single p-value is not strong evidence against the null hypothesis by itself. However, when we are dealing with more physiologically plausible research questions, the situation gets more complicated.

The p-value is calculated assuming that the null is true, so the Editorial is correct that we cannot simply flip the question around, assume the data, and get the probability of the null: $p(\bar{x}_d|H_0) \neq p(H_0|\bar{x}_d)$. To truly determine the probability of the null hypothesis, we would need Bayesian statistics in which we formalize some *prior* probability about the null hypothesis.[7] If we have a strong enough prior probability that the null is true, then the current data in the sample may not lead us to change our beliefs in the *posterior* distribution. This was case in the jelly bean example. However, for any given prior distribution, observing larger effects, which have smaller p-values under an NHST, will also lead to lower likelihoods in the posterior distribution.[C]

It is important to note, however, that small p-values are *relatively* less likely to be observed when the null hypothesis is true compared to when an alternative hypothesis is true. Thus, in a practical sense, p-values provide a form of evidence against the null-hypothesis, assuming that we are testing at least some real effects. As shown in Figure 2A, p-values have a uniform distribution under the null hypothesis, and 5% of p-values are thus below 0.05.[8] However, if the null is not true, then we will see a shift in the distribution of p-values, with small p-values becoming more common. An example of this is shown in

[C] For a humorous demonstration see: https://xkcd.com/1132/ ; for a more quantitative visualization of the relationship between priors, p-values, and posteriors see: https://rpsychologist.com/d3/bayes/.

Figure 2B, where the null is false and 34% of p-values are below 0.05. However, correctly rejecting the null hypothesis only 34% of the time is not ideal, so consider Figure 2C, where I have now tripled the sample size and 80% of p-values are below 0.05. That is, with 64 people per group, we now have 80% statistical power to detect a $\Delta = 0.5$.

So, small p-values do provide a *relative* measure of evidence against the null hypothesis, with small p-values being less likely when the null is true. However, to make *absolute* judgments about evidence for/against a specific hypothesis, we need more than a single p-value.[8] For instance, we can make specific assumptions about the prior likelihood of hypotheses in a Bayesian framework[7,9] or we could look at the distribution of multiple p-values testing the same hypothesis.[10]

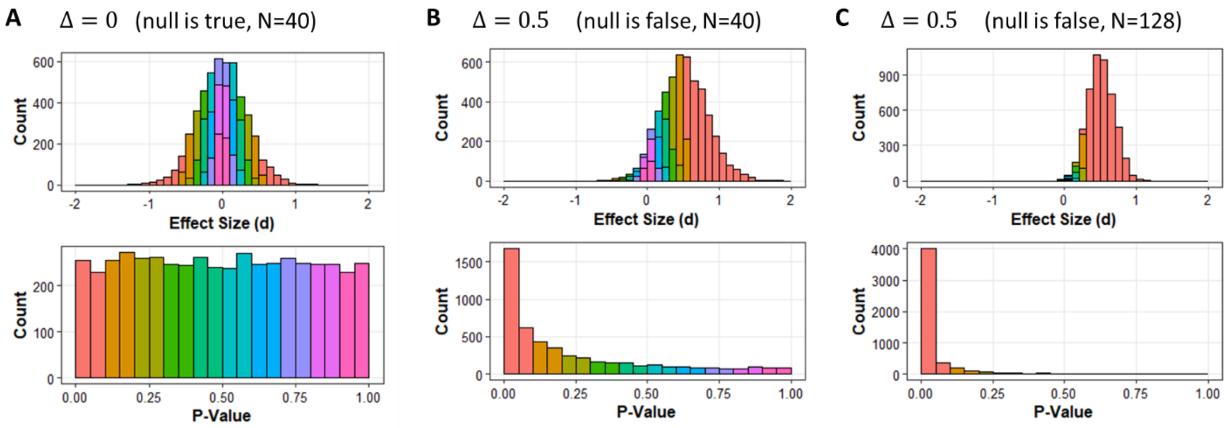

Figure 2. P-values do provide a form of evidence against the null hypothesis as p<0.05 is relatively less likely to occur when the null is true. Plots show simulated experiments (k=5,000, σ=1 for all populations) in which the means of two independent groups are compared using a t-test. In Panel A, the null hypothesis is true and the true difference between population means is 0. In Panel B, the null hypothesis is false and the true difference between population means is 0.5. In Panel C, the null-hypothesis is still false, but I have increased the sample size from 40 to 128, yielding 80% of p-values <0.05 (i.e., 80% statistical power). Quantiles are color coded with respect to their p-values and effects sizes are given as Cohen’s d.

3. ***“Statistically significant findings are not very replicable.”*** – This is misleading. First, it is difficult to precisely define replication,[11,12] but across many different definitions we would expect more statistically significant findings to “replicate” provided that hypothesis tests have adequate statistical power, researchers have not engaged in p-hacking, there is not selective reporting of results, etc. Thus, not all statistically significant findings will replicate,[13] but statistically significant findings in well-designed studies are more likely to replicate.[14–16] Second, any threats to replicability are also going to affect confidence intervals (the Editorial's proposed solution) as much as they affect p-values, because, again, the p-value is inextricably linked to the confidence interval. Thus, the Editorial is correct in a practical sense: many statistically significant findings in the current literature do not replicate. However, a lack of replication is the fault of poor study design and questionable research practices, not the use of hypothesis tests as a method of inference.

4. ***"In most clinical trials, the null hypothesis must be false."*** – This is true but very misleading. It is true that "real" treatment effects are unlikely to be precisely 0 (e.g., they might be +0.001), but it begs the question: do we really care if the true effect is 0 or 0.001? And will we ever have the statistical precision to discern that difference? Scientists are often working on the frontiers of human knowledge; this is costly work where we need to explore a lot of different ideas and many them do not pan out.[14] So, simply because a point estimate of precisely 0 is unlikely to be true does not mean that it is unhelpful to ask. ,It should be a very low bar to show that your clinical treatment has a non-zero effect! Further, the Editorial is specifically critiquing this "nil" hypothesis (i.e., $H_0 = 0$), when we could hypothesize any value.[2,5] And, if assuming $H_0 = 0$ is not desirable, we can set that null value to be anything we want (i.e., $H_0\colon \Delta \leq 0.4$ m/s for improvement in gait speed or $H_0\colon \Delta \leq 30\%$ change on a pain scale).

5. "*Researchers need information about the size of effects*." – This is a true statement, but it is not a problem with p-values nor null hypothesis significance tests. To my knowledge, no statistician has ever recommended that applied researchers ignore measures of effect size (either raw or standardized). Measures of effect size are integral to any results section. I would even take this one step further and encourage authors to share their data whenever possible[17], enabling other researchers to calculate their own effect sizes as there can be limitations with and confusion about standardized effects sizes, and there is no one-size-fits-all solution to effect sizes[18–20].

The Editorial's "Alternative" is a Hypothesis Test – The Minimal Effects Test

After detailing the potential problems with the NHST, the Editorial proposes an alternative solution in which they encourage authors to compare their 95% confidence to some minimum clinically meaningful value (which I will write as δ).[D] This is absolutely a good practice and I would encourage researchers to report 95% confidence intervals and interpret their upper and lower limits in context, when appropriate. However, what the Editorial is suggesting is effectively an MET where $H_0: \Delta \leq \delta$. That is, if the test is to see if the 95% confidence interval does not contain δ, then that is mathematically equivalent to an MET assuming $H_0: \Delta \leq \delta$ and finding $p < 0.025$. Note p<0.025, not p<0.05, because most METs are one-sided hypothesis tests whereas confidence intervals are two sided (see Figure 1 and Footnote A) After heavily critiquing hypothesis testing as a method of inference, the Editorial ends up effectively proposing a hypothesis test. This is clearly an illogical proposition.

I want to emphasize that it is absolutely valid for the Editorial to recommend that authors consider their 95% confidence interval relative to some clinically meaningful value. However, this is not an "alternative" to conducting a null hypothesis significance test, it is in fact mathematically identical to conducting a null hypothesis test with a carefully chosen null hypothesis. Both are valid.

Finally, it is important to stress that history provides us with several examples of how authors will view their data through rose-tinted glasses when quantitative statistical safeguards are removed. For instance, when *Basic and Applied Social Psychology* banned p-values, authors were found to overstate their conclusions well beyond what would have been considered if "statistical significance" had been a benchmark.[23] In sport and exercise science, "magnitude-based inference" was leveraged as a niche method that allowed authors to interpret differences as meaningful when they had very little statistical support (e.g., p's >0.25).[24–26] Statistical significance in an NHST does not necessarily need to be the

[D] I caution that it is difficult to find a single measure of δ; it changes as a function of the study population, the study context, and has its own uncertainty due to sampling error.[21,22]

benchmark nor 0.05 the default value[27–30], but it is always important to have a statistically sound framework for dealing with uncertainty.

Virtues of Hypothesis Testing

One of the great virtues of null hypothesis significance testing is Type I error control while making minimal assumptions about the nature of the data or the world at large; if we set $\alpha = 0.05$, then we can be confident we will only get data greater than or equal to what we observed 5% of the time when the null is true. Importantly, this works for a wide range of statistics and types of tests, including F- and χ^2-statistics that have multiple degrees of freedom from models asking questions about multiple effects simultaneously. For instance, in a randomized control trial with three arms, I could conduct an omnibus F-test and obtain a p-value to see if there is any evidence of a difference between groups overall, before conducting additional post-hoc tests to compare specific groups. This situation is not covered by the Editorial and the Editorial's confidence interval alternative is not easily applied here, although one could plausibly adjust the width of the confidence intervals to control for multiple comparisons.

Bigger Threats to Statistical Integrity

Misinterpretation or misuse of p-values are a threat to statistical integrity. However, questionable research practices such as p-hacking, sub-group analyses, flexible stopping rules, selective exclusion of outliers, selective reporting, or hypothesizing after results are known are much larger threats.[31–35] Furthermore, these questionable research practices have consistently negative consequences regardless of the method of inference. For instance, although the term p-hacking connotes the NHST, these questionable research practices pose an equal threat to confidence intervals because confidence intervals are p-values are based on the same underlying mathematics. Similarly, switching to a fully Bayesian method of analysis is not an antidote for poor study design, small samples, and questionable research practices. As others have argued[36,37], p-values get a disproportionate amount of attention in popular discussions of research methodology. I encourage the ISPJE to instead focus their attention on methods for improving data/code sharing, transparency, and replicability through tools like preregistration, results-blind review, registered reports, or even “data papers” whose primary function is to report a study and archive the data, without drawing inferences from limited samples.

It is entirely valid to say that p-values are often mis-used and mis-interpreted, and “statistical significance” may not ultimately be the best term for applied researchers to use.[38] However, it is incorrect to present these human errors as inherent flaws in hypothesis testing. For instance, if someone mis-interprets p>0.05 as evidence of “no difference”, then I would argue the correct action is to teach them about equivalence tests and non-inferiority designs, not ban p-values. Similarly, there are times when Bayesian inference is what authors are really interested in (e.g., what is the probability that the null is true, given the evidence?), and in those cases Bayesian inference can and should be used. However, Bayesian analysis is not a panacea and needs to be used thoughtfully like any statistical tool. So, although a simple heuristic of p<0.05 may well be overused as “the” test in physical therapy research, frequentist hypothesis tests are still valid and useful tools for physical therapy researchers. Moreover, the scientific integrity of the field has much larger concerns, and both p-values and confidence intervals will be corrupted by p-hacking, under-powered subgroup analyses, surrogate outcomes, and other questionable research practices.

In conclusion, I agree with the Editorial on the importance of reporting effect sizes and interpreting them in context. However, the Editorial makes numerous statistical faux pas that could harm the statistical literacy in our field, if readers take them at face value, and harm the scientific integrity of our field, if put into editorial practice.

References

1. Elkins, M. R. *et al.* Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. *Phys. Ther.* **102**, pzac066 (2022).
2. Murphy, K. R. & Myors, B. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *J. Appl. Psychol.* **84**, 234–248 (1999).
3. Rafi, Z. & Greenland, S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med. Res. Methodol.* **20**, 244 (2020).
4. Cohen, J. The earth is round (p < .05). *Am. Psychol.* **49**, 997–1003 (1994).
5. Lakens, D. The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspect. Psychol. Sci.* **16**, 639–648 (2021).
6. Herbert, R. Research Note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. *J. Physiother.* **65**, 178–181 (2019).
7. Goodman, S. N. & Royall, R. Evidence and scientific research. *Am. J. Public Health* **78**, 1568–1574 (1988).
8. Lakens, D. Why p-values are not measures of evidence. (2021).
9. Goodman, S. N. Why is getting rid of p-values so hard? Musings on science and statistics. *Am. Stat.* **73**, 26–30 (2019).
10. Simonsohn, U., Nelson, L. D. & Simmons, J. P. P-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534 (2014).
11. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
12. Patil, P., Peng, R. D. & Leek, J. T. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect. Psychol. Sci.* **11**, 539–544 (2016).

13. Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Adv. Methods Pract. Psychol. Sci.* **4**, 25152459211007468 (2021).

14. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

15. Anderson, S. F. & Maxwell, S. E. Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivar. Behav. Res.* **52**, 305–324 (2017).

16. Nosek, B. A. *et al.* Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).

17. Borg, D. N. *et al.* Sharing data and code: a comment on the call for the adoption of more transparent research practices in sport and exercise science. (2020).

18. Caldwell, A. & Vigotsky, A. D. A case against default effect sizes in sport and exercise science. *PeerJ* **8**, e10314 (2020).

19. McGrath, R. E. & Meyer, G. J. When effect sizes disagree: the case of r and d. *Psychol. Methods* **11**, 386 (2006).

20. Levine, T. R. & Hullett, C. R. Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Hum. Commun. Res.* **28**, 612–625 (2002).

21. Tenan, M. & Caldwell, A. A Critical Review of Phyiotherapy Editor's Comments on Statistical Practice.

22. Dabija, D. I. & Jain, N. B. Minimal Clinically Important Difference of Shoulder Outcome Measures and Diagnoses: A Systematic Review. *Am. J. Phys. Med. Rehabil.* **98**, 671–676 (2019).

23. Fricker Jr, R. D., Burke, K., Han, X. & Woodall, W. H. Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *Am. Stat.* **73**, 374–384 (2019).

24. Sainani, K. L. The Problem with" Magnitude-based Inference". *Med. Sci. Sports Exerc.* **50**, 2166–2176 (2018).

25. Sainani, K. L., Lohse, K. R., Jones, P. R. & Vickers, A. Magnitude-based inference is not Bayesian and is not a valid method of inference. *Scand. J. Med. Sci. Sports* **29**, 1428 (2019).

26. Lohse, K. R. *et al.* Systematic review of the use of "magnitude-based inference" in sports science and medicine. *PloS One* **15**, e0235318 (2020).

27. Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).

28. Lakens, D. *et al.* Justify your alpha. *Nat. Hum. Behav.* **2**, 168–171 (2018).

29. Amrhein, V. & Greenland, S. Remove, rather than redefine, statistical significance. *Nat. Hum. Behav.* **2**, 4–4 (2018).

30. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon statistical significance. *Am. Stat.* **73**, 235–245 (2019).

31. Simmons, J. P., Nelson, L. D. & Simonsohn, U. Life after p-hacking. in *Meeting of the society for personality and social psychology, New Orleans, LA* 17–19 (2013).

32. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. (2016).

33. Sun, X. *et al.* Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *Bmj* **344**, (2012).

34. Kerr, N. L. HARKing: Hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* **2**, 196–217 (1998).

35. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, (1979).

36. Borg, D. N., Lohse, K. R. & Sainani, K. L. Ten common statistical errors from all phases of research, and their fixes. *PM&R* **12**, 610–614 (2020).

37. Leek, J. T. & Peng, R. D. Statistics: P values are just the tip of the iceberg. *Nature* **520**, 612–612 (2015).

38. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a world beyond "p< 0.05". *The American Statistician* vol. 73 1–19 (2019).