PROF. BEN VAN CALSTER (Orcid ID : 0000-0003-1613-7450)

PROF. EWOUT STEYERBERG (Orcid ID : 0000-0002-7787-0122)

Article type : Methods

Consequences of relying on statistical significance: some illustrations

Ben Van Calster[1,2], PhD, Ewout W Steyerberg[2], PhD, Gary S Collins[3], PhD, Tim Smits[4], PhD

1 KU Leuven, Department of Development and Regeneration, Herestraat 49 box 805, 3000 Leuven, Belgium

2 Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, the Netherlands

3 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

4 KU Leuven, Institute for Media Studies, Parkstraat 45 box 3603, 3000 Leuven, Belgium

Correspondence and requests for reprints to:

Ben Van Calster

KU Leuven, Department of Development and Regeneration

Herestraat 49 box 805

3000 Leuven, Belgium

ben.vancalster@med.kuleuven.be, +3216377788.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/eci.12912

Abstract

Background – Despite regular criticisms of null hypothesis significance testing (NHST), a focus on testing persists, sometimes in the belief to get published and sometimes encouraged by journal reviewers. This paper aims to demonstrate known key limitations of NHST using simple nontechnical illustrations.

Design – The first illustration is based on simulated data of 20,000 studies that compare two groups for an outcome event. The true effect size (difference in event rates) and sample size (20 to 100 per group) were varied. The second illustration used real data from a meta-analysis on alpha blockers for the treatment of ureteric stones.

Results – The simulations demonstrated the large between-study variability of p-values (range between <0.0001 and 1 for most simulation conditions). A focus on statistically significant effects ($p<0.05$), notably in small to moderate samples, led to strongly overestimated effect sizes (up to 240%) and many false positive conclusions, i.e. statistically significant effects that were in fact true null effects. Effect sizes also exerted strong between-study variability, but confidence intervals accounted for this: the interval width decreased with larger sample size, and the percentage of intervals that contained the true effect size was accurate across simulation conditions. Reducing alpha level, as recently suggested, reduced false positive conclusions but strongly increased the overestimation of significant effects (up to 320%).

Conclusions - Researchers and journals should abandon statistical significance as a pivotal element in most scientific publications. Confidence intervals around effect sizes are more informative, but should not merely be reported to comply with journal requirements.

Introduction

P-values are routinely used by applied researchers to analyze and interpret results from empirical studies [1]. Although p-values are quantitative, the key interest typically is whether a p-value is lower than the widely adopted threshold of 0.05. This creates an artificial dichotomy in order to label a result as statistically significant or not. Often, statistical significance is used to conclude whether or not there is an effect (or association). However, the interpretation of p-values and in particular null hypothesis significance testing (NHST) have limitations that have long been known [2-11]. Recently, the American Statistical Association (ASA) has fueled the discussion by releasing a statement including six principles on how to deal with p-values (Table 1) [12]. The key question is the extent to which the criticisms reach applied researchers and journal reviewers, and if so, the extent to which technical expositions are understandable by non-statisticians. Our experiences are that the problematic consequences of NHST are ignored, often in the belief to increase chances of publication, or not well understood. The present article aims to contribute to the debate about p-values and NHST by presenting clear-cut illustrations of the most important limitations, rather than a general technical description of these limitations.

The first obstacle is that the p-value is a complicated concept. P-values are used to evaluate the compatibility of the observed data with a pre-specified null hypothesis, which usually states that there is no effect. A common misinterpretation is that the p-value is the

probability that the null hypothesis is true. Instead, it is the probability that, if the null hypothesis is true, data can be observed that are at least as extreme from the null hypothesis as the actually observed data. Thus, p-values make a statement about the data, and assume that the null hypothesis is true in doing so.

Strong reliance on statistical significance has erratic consequences, mainly when sample size is limited [6,9]. P-values vary when the same study is repeated in exactly the same way [9,13]. Generally, this variability increases as sample size decreases, although it depends on the underlying true effect size. As a result, the declaration of statistical significance relies considerably on chance in samples of limited size. Crucial consequences are that the size of a statistically significant effect is on average overestimated, and that statistically significant results are surprisingly often incorrect ('false positive') [5,6,14-16]. Recently, it has been suggested that the alpha level should be lowered to 0.005 for declaring statistical significance [17,18]. The key argument is that this should enhance the reproducibility of claimed effects, in terms of a reduction of false positive findings, although the authors indicate that lack of reproducibility has a complex multitude of causes.

Other critical papers about p-values state that we should shift focus to calculating the observed effect size together with a confidence interval [1,3,7,9,11-13,19]. The wider the confidence interval, the more uncertain the estimated effect size. Importantly, confidence intervals deal with precision of the effect size, and should not be used to determine statistical significance.

Despite the existing literature on the topic, the reporting of p-values in abstracts and papers has increased over time [1]. A recent review of published studies in four leading general medical journals in 2016 (Annals of Internal Medicine, BMJ, JAMA, New England Journal of

Medicine) found that 82% (122/149) of articles that analyzed original numerical data reported at least 1 p-value [20]. For articles that report p-values, on average 36 p-values were given (range 1 to 141). Reporting confidence intervals without interpretation as statistical tests increases in leading medical journals, but the evolution is modest [21].

This suggests that the message does not sufficiently reach or convince applied researchers. Furthermore, we recognize that current incentives for research (i.e. pressure to publish papers on novel findings) are another important determinant of scientific practice [22].

In what follows, we aim to illustrate the abovementioned NHST limitations based on simulated data, followed by a real life illustration based on a recent meta-analysis. We also illustrate how lowering the alpha level affects overestimation and the number of false positives.

Methods

Illustration based on simulated data

An enormous number of studies are published every year, many of which quantitatively compare groups of individuals. Let us imagine 20,000 independent studies have compared two groups of individuals with respect to the presence of an outcome event. We assume that for 80% of the studies (16,000) there is no true difference in the outcome. This may seem enormous, but is in fact not exaggerated for many research domains [5,6,16,23-25]. For simplicity, we assume for each of these 16,000 studies that the event rate is 0.40 in each group. Furthermore, for 2,000 (10%) studies we assume the true effect size is small (event

rate 0.40 vs 0.50), for 1500 (7.5%) studies it is moderate (0.40 vs 0.60), and for 500 (2.5%) studies it is large (0.40 vs 0.70).

For each of the 20,000 studies we simulate study participants using the event rates given above. The groups are compared using a likelihood ratio chi-square test, and a confidence interval on the risk difference is constructed using the standard asymptotic method. We adopt an alpha level of 5% to declare statistical significance, but afterwards we compare results with newly suggested lower alpha levels of 0.5% and 0.1%. We carry out the 20,000 studies four times, each time with a different sample size per group: 100, 50, 30, and 20. Such sample sizes are common in the scientific literature [26].

Illustration based on real data from a recent meta-analysis

Recently, Hollingsworth and colleagues published a systematic review and meta-analysis of randomized controlled trials to assess the effect of alpha blockers on the treatment of ureteric stones [27]. The authors included 55 studies. The outcome was the proportion of patients with passage of the stones. Study results were presented as risk ratios: the outcome proportion for interventional vs control patients. We calculated the p-value based on the reported risk ratio and 95% confidence interval [28]. The overall risk ratio was estimated as 1.49 (95% CI 1.39 to 1.61), suggesting a clear positive effect of alpha blockers.

Results

Simulation of studies

We observed considerable variability in the p-values, even when stratifying for sample size and true effect size (Table 2; Figure S1). When the true effect size was between zero and moderate (40% vs 60% event rates), the observed p-values ranged between <0.0001 to 1 even with 100 patients per group (p-values range from around 0.0001 to 1). Only when the sample size and the true effect size were large, the variability was small. The standard deviation of p-values increased with decreasing sample size.

When relying on statistical significance at the classical 5% level ($p<0.05$), the consequences of the p-value variability are clear (Table 2). First, the effect size of statistically significant effects was strongly overestimated. The overestimation was more pronounced in smaller studies and in studies where the true effect size was small. When an effect that is in reality small is statistically significant in a study with 20 patients per group, the overestimation was on average no less than 241%. Overestimation decreased with the true effect size. Yet, with 20 patients per group, truly large effects that were statistically significant were still 43% overestimated. In a large study (100 patients per arm), a truly small effect that was statistically significant was 81% overestimated. Only with truly large effects in large datasets, the overestimation was small (1%).

Second, the significance criterion often resulted in false positive inferences (Table 2). With 20 patients per group, a staggering 53% of the statistically significant results dealt with effects that were true null effects. When the sample size was 100 per group, still 26% of the statistically significant effects were false positives.

Not only p-values were variable, effect sizes were too (Table 3). Yet, we observed that the average estimated effect size did not change with sample size, whereas p-values strongly decreased with increasing sample size. We also observed that confidence intervals around the effect size nicely accounted for the uncertainty caused by the variability. The width of the interval increased with decreasing sample size, such that in all settings around 95% of the intervals included the true effect size. Further, the width of the confidence interval was hardly influenced by the true effect size.

Meta-analysis of the effect of alpha blockers on the treatment of ureteric stones

We observed a large variability of the p-value between the studies (range <0.0001 to 0.96) (Figure 1a). In addition, the observed effect size was clearly larger for studies that were statistically significant (Figure 1b). The confidence interval around the effect size captured the uncertainty of the estimates, because the width of the interval gradually increased with decreasing sample size (Figure S2). This example hence illustrates the enormous variability in p-values, and the overestimation of effects if we condition on statistical significance. It cannot illustrate the fact that statistically significant effects are often false positives because we do not know the truth, and because the example deals with a single underlying effect that is either present or absent.

Effect of reducing alpha level to 0.005

When using an alpha of 0.005 for the simulated studies, the amount of false positive findings decreased markedly, although it was still 31% for the studies with 20 patients per arm (Table 4). The percentage of false negative effects, i.e. the percentage of non-significant

results that dealt with true non-null effects, increased only modestly. However, the focus on statistically significant effects led to a very strong increase in effect overestimation when lowering alpha to 0.005. True small effects, when detected in a study with 20 patients per arm, were now overestimated by 321% on average (compared to 241% with a 'traditional' alpha level at 0.05). True small effects were 128% overestimated when using 100 patients per arm (compared to 81% with alpha at 0.05).

For the meta-analysis data, a simple analysis (using the mean log risk ratio weighted by the inverse variance) resulted in a summary risk ratio of 1.46 for all studies, 1.70 for studies with $p<0.05$, and 1.84 for studies with $p<0.005$.

Discussion

We illustrated important limitations in the interpretation of p-values and statistical significance. The central finding is the inconveniently high variability of p-values. Key consequences are that the focus on statistically significant effects leads to overestimation of effect sizes and surprisingly to many false positive conclusions, mainly in small to moderately sized studies [5,6,14-16]. Overestimation is lower when the true effect is large since the result will more often be statistically significant.

The large amount of false positives is counterintuitive at first sight: if the alpha level is fixed at 5% irrespective of sample size, how can there be so many false positives, and how can there be more false positives in small samples? The alpha level refers to the likelihood of a significant result if the null hypothesis is true, whereas the false positive rate refers to the likelihood that the null hypothesis is true despite the observation of a significant effect.

Often, researchers may erroneously think the opposite: because only truly large effects have reasonable power in small samples, observing a statistically significant effect in a small sample indicates that the true effect is large. In reality, such effects surprisingly often involve true null effects.

A first limitation of our illustration is that the results depend on the settings of the simulation, but the observed tendencies hold generally. The likelihood of a false positive depends on the percentage of studies that deal with true null effects. Research fields with more true null effects suffer more from false positive results than fields with less true null effects [6,18]. The percentage of true null effects is unknown, but it is realistic to assume that many or even the majority of statistical tests deal with null effects, in particular when p-values are overused [5,6]. A second limitation is that results of the systematic review can be affected by various additional issues. Specific patient populations and inclusion and exclusion criteria may lead to differences in patient case-mix between studies. In addition, the authors of the meta-analysis found support for publication bias.

We argue that the strong focus on the search for statistical significance irrespective of sample size or study design, perhaps to increase chances of publication or in the belief that it is a proof that the effect is real, is problematic. We, as do others, believe that estimation of effect sizes and confidence intervals are more informative: these measures give more insight into the observed effect (or association), and its uncertainty [1,3,7,9,11-13,19]. Admittedly, confidence intervals around the effect size are neither the only nor the ultimate solution. A valuable alternative, that we do not address here, focuses on Bayesian methods such as credibility intervals [29]. Yet, in combination with strongly diminished emphasis on statistical significance, confidence intervals around the effect size are a sensible step

towards a more responsible scientific practice [11]. All of this implies that, for many studies, the strong focus on making clear dichotomous conclusions regarding the presence or absence of an effect should be reduced.

We stress that p-values as such are not wrong. When used in isolation, however, we do feel that they are of limited interest given that their magnitude is determined by a mix of effect size and uncertainty [30]. We emphasize the importance of separating estimation (observed effect size) and uncertainty (confidence interval around the observed effect size). Furthermore, many researchers have insufficient knowledge of p-values. Together with controversial incentives for researchers [22], this feeds the exaggerated focus on p-values, the reduction of study results to statistical significance, and the common ignorance of the effect of sample size. This practice jeopardizes robust interpretation and causes inflation of research findings. The suitability of testing for statistical significance depends on the type of study. The use of statistical significance is more warranted to evaluate a clearly defined pre-specified hypothesis in a dedicated and adequately powered study. Hence we argue to confine testing for statistical significance largely to confirmatory studies for a clearly defined primary endpoint with a realistic power calculation. In reality, many studies are exploratory, suffer from too low sample size, or have many non-primary endpoints.

Mathematically, reducing the alpha level for declaring statistical significance should improve reproducibility of findings by decreasing the amount of false positive results [18]. Our illustrations confirmed this pattern. However, the downsides are that it keeps the focus on statistical significance, which only moves the target that researchers will aim at, and that it leads to more overestimated effect sizes. As a result, follow up studies aiming to replicate claimed effects will a) probably suffer from too low power due to the initially overestimated

effect size and b) lead to stronger downward corrections of initially reported effects than is currently the case [14]. Unfortunately, the amount of false positives is also increased by common phenomena such as lack of a (pre-registered) research protocol, use of p-values without a pre-specified hypothesis, p-hacking, incorrect statistical analyses, publication bias, and data dredging [20,30-39]. All of these phenomena are common in science, and may even be more important sources of irreproducible research than the use of a 5% alpha level. We argue that working towards better research practices, and hence also towards more reasonable incentives for researchers [22], is more important than the introduction of a more stringent alpha level. We summarise a number of recommendations to improve more responsible scientific practices in text box 1.

To conclude, it is long overdue that we need to change emphasis from p-values and dichotomous statistical significance interpretations to more gradual and informative approaches such as effect sizes and confidence intervals. Journals and reviewers can play an important role in changing research and reporting practices. Reviewers should not demand p-values when authors did not provide them. It is not sufficient merely to require that authors provide confidence intervals, because researchers may only report them to comply with journal requirements [40]. In addition, journals and reviewers should pay attention to how authors obtained, described, and interpreted the study results.

Acknowledgements

We have no specific funding to report for this study.

References

1. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA 2016;315:1141-8.
2. Rozeboom WW. The fallacy of the null hypothesis significance test. Psychol Bull 1960;57:416-28.
3. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986;292:746-50.
4. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Int Med 1999;130:995-1004.
5. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226-31.
6. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124.
7. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010;25:225-30.
8. Nuzzo R. Scientific method: statistical errors. Nature 2014;506:150-2.
9. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods 2015;12:179-85.
10. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337-50.
11. Rothman KJ. Disengaging from statistical significance. Eur J Epidemiol 2016;31:443-4.
12. Wasserstein RL, Lazar NA. The ASA’s statement on p-values: context, process, and purpose. Am Stat 2016;70:129-33.

13. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 2008;3:286-300.
14. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology 2008;19:640-8.
15. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14:365-76.
16. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96:434-42.
17. Johnson VE. Revised standards for statistical evidence. PNAS 2013;110:19313-7.
18. Benjamin DJ, Berger J, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Human Behav 2018;2:6-10.
19. Poole C. Low p-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12:291-4.
20. Perneger TV, Combescure C. The distribution of p-values in medical research articles suggested selective reporting associated with statistical significance. J Clin Epidemiol 2017;87:70-7.
21. Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstract of major medical and epidemiology journals 1975-2014: a systematic review. Eur J Epidemiol 2017;32:21-9.
22. Higginson AD, Munafò MR. Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLoS Biol 2016;14:e2000995.

23. Begley CG, Ioannidis JPA. Reproducibility in science. Improving the standard for basic and preclinical research. Circ Res 2015;116:116-26.
24. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. Proc Natl Acad Sci 2015;112:15343-7.
25. Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the reproducibility of psychological science. J Am Stat Assoc 2017;112:1-10.
26. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ 2010;340:c723.
27. Hollingsworth JM, Canales BK, Rogers MAM, Sukumar S, Yan P, Kuntz GM, et al. Alpha blockers for treatment of ureteric stones: systematic review and meta-analysis. BMJ 2016;355:i6112.
28. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ 2011;343:d2304.
29. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 2016;23:103-23.
30. Lang JM, Rothman KJ, Cann CI. That confounded p-value. Epidemiology 1998;9:7-8.
31. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359-66.
32. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383:166-75.

33. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KGM, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med 2014;11:e1001671.
34. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. JAMA 2004;291:2457-65.
35. Odutayo A, Altman DG, Hopewell S, Shakir M, Hsiao AJ, Emdin CA. Reporting of a Publicly Accessible Protocol and Its Association With Positive Study Findings in Cardiovascular Trials (from the Epidemiological Study of Randomized Trials [ESORT]). Am J Cardiol 2015;116:1280-3.
36. Von Elm E, Egger M. The scandal of poor epidemiological research. BMJ 2004;329:868-9.
37. Szucs D. A tutorial on hunting statistical significance by chasing N. Front Psychol 2016;7:1444.
38. Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. Elife 2016;5:e21451.
39. Kagereki E, Gakonyo J, Simila H. Significance bias: an evaluation of the oral health literature. BMC Oral Health 2016;16:53.
40. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think - Statistical reform lessons from medicine. Psychol Sci 2004;15:119-26.

Figure legends

Figure 1. Scatter plot of p-value (panel A) and effect size (panel B) by sample size for the studies included in the meta-analysis by Hollingsworth et al. Panel B is stratified by statistical significance of the individual result.

Figure S1. Box plots of p-values by true effect size of the 20,000 simulated studies for each sample size setting.

Figure S2. Scatter plot of the width of the confidence interval around the effect size by sample size for the studies included in the meta-analysis by Hollingsworth et al.

Table 1. Summary of the six principles on p-values and statistical significance given by the American Statistical Association [12].

1. P-values can indicate how compatible the data are with a specified statistical model
2. P-values do not measure the probability that the studies hypothesis is true or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decision should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Table 2. Results about p-values and statistical significance from the simulated studies that compare two groups for the presence of an outcome event.

		Sample size per arm			
Result	**True effect size**	**100**	**50**	**30**	**20**
Mean p-value (SD)	None (0)	0.50 (0.29)	0.50 (0.29)	0.50 (0.30)	0.50 (0.30)
	Small (0.1)	0.26 (0.28)	0.35 (0.30)	0.41 (0.30)	0.45 (0.31)
	Moderate (0.2)	0.04 (0.10)	0.14 (0.22)	0.23 (0.26)	0.28 (0.29)
	Large (0.3)	0.001 (0.01)	0.03 (0.08)	0.07 (0.14)	0.15 (0.22)
p-value range	None (0)	<0.0001 – 1	0.0001 – 1	<0.0001 – 1	<0.0001 – 1
	Small (0.1)	<0.0001 – 1	<0.0001 – 1	0.0001 – 1	<0.0001 – 1
	Moderate (0.2)	<0.0001 – 1	<0.0001 – 1	<0.0001 – 1	<0.0001 – 1
	Large (0.3)	<0.0001 – 0.06	<0.0001 – 0.69	<0.0001 – 1	<0.0001 – 1
P<0.05 (%)	None (0)	5%	5%	5%	5%
	Small (0.1)	33%	20%	11%	7%
	Moderate (0.2)	83%	56%	35%	24%
	Large (0.3)	>99%	90%	71%	47%
Mean effect size if	None (0)	0.00[b]	-0.02[b]	0.00[b]	-0.03[b]
p<0.05 (bias[a])	Small (0.1)	0.18 (+81%)	0.24 (+140%)	0.28 (+184%)	0.34 (+241%)
	Moderate (0.2)	0.22 (+11%)	0.27 (+36%)	0.33 (+65%)	0.40 (+101%)
	Large (0.3)	0.30 (+1%)	0.33 (+9%)	0.36 (+21%)	0.43 (+43%)
False positives[c] (%)	NA	26%	33%	42%	53%

For each sample size, 20,000 studies were simulated, of which 16,000 with a true null effect (event rate of 0.4 in both groups), 2,000 with a small effect (event rates of 0.4 vs 0.5), 1,500 with a moderate effect (event rates of 0.4 vs 0.6), and 500 with a large effect (event rates of 0.4 vs 0.7).

SD, standard deviation.

[a] bias is the percentage difference relative to the true effect size, calculated as 100*(mean effect size – true effect size)/true effect size

[b] The mean observed effect size of significant effects is 0 because about half of them are negative and about half of them are positive.

[c] percentage of statistically significant effects (p<0.05) where the true effect size is 0 (i.e. no difference between groups).

Table 3. Results about effect sizes and confidence intervals from the simulated studies that compare two groups for the presence of an outcome event.

		Sample size per arm			
Result	**True effect size**	**100**	**50**	**30**	**20**
Mean effect size	None (0)	0.00 (0.07)	0.00 (0.10)	0.00 (0.12)	0.00 (0.15)
(SD)	Small (0.1)	0.10 (0.07)	0.10 (0.10)	0.09 (0.13)	0.09 (0.15)
	Moderate (0.2)	0.20 (0.07)	0.20 (0.10)	0.20 (0.13)	0.21 (0.15)
	Large (0.3)	0.30 (0.07)	0.31 (0.10)	0.31 (0.13)	0.30 (0.15)
Effect size range	None (0)	-0.30 – 0.27	-0.34 – 0.38	-0.53 – 0.47	-0.55 – 0.65
	Small (0.1)	-0.14 – 0.33	-0.26 – 0.44	-0.30 – 0.50	-0.45 – 0.65
	Moderate (0.2)	-0.04 – 0.43	-0.06 – 0.48	-0.33 – 0.60	-0.35 – 0.70
	Large (0.3)	0.13 – 0.52	0.04 – 0.60	-0.13 – 0.70	-0.20 – 0.70
Mean CI width (SD)	None (0)	0.27 (0.004)	0.38 (0.01)	0.49 (0.02)	0.59 (0.03)
	Small (0.1)	0.27 (0.003)	0.38 (0.01)	0.49 (0.01)	0.60 (0.02)
	Moderate (0.2)	0.27 (0.004)	0.38 (0.01)	0.49 (0.01)	0.59 (0.03)
	Large (0.3)	0.26 (0.006)	0.37 (0.01)	0.47 (0.02)	0.57 (0.03)
CIs that contain the	None (0)	95%	95%	95%	92%
true effect size (%)	Small (0.1)	94%	95%	94%	94%
	Moderate (0.2)	95%	94%	94%	95%
	Large (0.3)	95%	93%	96%	95%

For each sample size, 20,000 studies were simulated, of which 16,000 with a true null effect (event rate of 0.4 in both groups), 2,000 with a small effect (event rates of 0.4 vs 0.5), 1,500 with a moderate effect (event rates of 0.4 vs 0.6), and 500 with a large effect (event rates of 0.4 vs 0.7).
SD, standard deviation; CI, confidence interval.

Table 4. Results about statistical significance from the simulated studies that compare two groups for the presence of an outcome event using an alpha level of 0.005 instead of 0.05.

		Sample size per arm			
Result	**True effect size**	**100**	**50**	**30**	**20**
P<0.005 (%)	None (0)	0.4%	0.5%	0.4%	0.7%
	Small (0.1)	11%	5%	1%	2%
	Moderate (0.2)	53%	24%	11%	7%
	Large (0.3)	95%	66%	37%	20%
Mean effect size if	None (0)	-0.01	-0.09	-0.04	-0.06
p<0.005 (bias[a])	Small (0.1)	0.23 (+128%)	0.31 (+206%)	0.40 (+300%)	0.42 (+321%)
	Moderate (0.2)	0.25 (+26%)	0.33 (+64%)	0.41 (+103%)	0.50 (+149%)
	Large (0.3)	0.31 (+4%)	0.36 (+20%)	0.42 (+41%)	0.51 (+70%)
False positives[c] (%)	NA	4%	9%	15%	31%

For each sample size, 20,000 studies were simulated, of which 16,000 with a true null effect (event rate of 0.4 in both groups), 2,000 with a small effect (event rates of 0.4 vs 0.5), 1,500 with a moderate effect (event rates of 0.4 vs 0.6), and 500 with a large effect (event rates of 0.4 vs 0.7).
SD, standard deviation.

[a] bias is the percentage difference relative to the true effect size, calculated as 100*(mean effect size – true effect size)/true effect size

[b] The mean observed effect size of significant effects is 0 because about half of them are negative and about half of them are positive.

[c] percentage of statistically significant effects ($p<0.05$) where the true effect size is 0 (i.e. no difference between groups).

Summary text box 1: recommendations for responsible scientific practice

* Conduct preregistered protocol-driven studies

* Distinguish between confirmatory and exploratory research questions

* Distinguish between pre-specified and post-hoc analyses

* Avoid small samples where possible, and base confirmatory studies on realistic power calculations

* Be reasonable and modest when it comes to statistical analysis

* Reduce focus on dichotomous conclusions, enhance focus on estimation

* Use measures of effect size with confidence intervals

* Avoid exaggerated conclusions of research findings ('spin') as well as publication bias