A review of high impact journals found that misinterpretation of non-statistically significant results from randomized trials was common

Hemming, Karla; Javid, Iqra; Taljaard, Monica

DOI:
10.1016/j.jclinepi.2022.01.014

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Hemming, K, Javid, I & Taljaard, M 2022, 'A review of high impact journals found that misinterpretation of non-statistically significant results from randomized trials was common', *Journal of Clinical Epidemiology*, vol. 145, pp. 112-120. https://doi.org/10.1016/j.jclinepi.2022.01.014

Journal of Clinical Epidemiology 145 (2022) 112–120

REVIEW

A review of high impact journals found that misinterpretation of non-statistically significant results from randomized trials was common

Karla Hemming[a,*], Iqra Javid[b], Monica Taljaard[c]

[a] *Institute of Applied Health Research, University of Birmingham, Birmingham, UK*
[b] *Biomedical Science, University of Birmingham, Birmingham, UK*
[c] *Clinical Epidemiology Program, and School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

Accepted 18 January 2022; Available online 23 January 2022

Abstract

Objectives: To determine the prevalence of poor interpretation practices, such as conflating evidence of absence with absence of evidence and over-emphasis of statistical non-significance in abstract conclusions, in a sample of randomized controlled trials (RCTs) with non-statistically significant primary outcomes published after the 2016 American Statistical Association statement on the interpretation of P-values.

Design and setting: Review of 50 two-arm individually randomized superiority trials with non-statistically significant results in four high impact journals published between 2017 and 2020, to determine the proportion that conclude evidence of no impact (thus, likely conflating evidence of absence with absence of evidence) or place emphasis on statistical non-significance (technically correct but arguably uninformative) in the abstract conclusion.

Results: Of the 50 RCTs with non-statistically significant results for primary outcomes, 28 (56%) of abstract were classified as concluding there was no difference between the two treatments; 19 (38%) placed an over-emphasis on statistical significance; only one acknowledged any uncertainty and the remaining 2 (4%) concluded that one treatment was more effective. Only four studies provided any justification for a finding of no difference, for example that the confidence interval gave no support to values of importance.

Conclusions: RCTs with non-statistically significant primary outcomes almost always present their conclusion in the abstract as evidence of no impact or ambiguously as "not statistically significant" without giving due attention to values supported by the confidence interval.

Keywords: Non-statistically significant; Conflating absence of evidence; Misinterpretation; Randomised controlled trials; Reporting; Treatment effects; Confidence intervals; P-values; Clinical importance

What is new

Despite the publication of the 2016 American Statistical Association statement on the interpretation of p-values, RCTs with non-statistically significant findings almost always misinterpret the primary outcome result.

1. Background

Randomised controlled trials (RCTs) are the most robust method of assessing the effectiveness of treatments [14]. A key requirement is to pre-specify a primary outcome [25]. This focuses assessment of effect on outcomes that are a-priori considered to be clinically important and reduces the

Conflict of interests: All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organization for the submitted work no financial relationships with any organizations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Author contributions: KH and MT led the development of the idea and undertook the independent assessment of study conclusions. IJ undertook the search and data abstraction. All authors made an intellectual contribution to the development of the ideas and commented on draft versions of the paper.

Funding: This research was partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative. Karla Hemming is funded by a NIHR Senior Research Fellowship SRF-2017-10-002. This research is independent to the funder.

* Corresponding author.
E-mail address: k.hemming@bham.ac.uk (K. Hemming).

https://doi.org/10.1016/j.jclinepi.2022.01.014

likelihood of chance findings. In a frequentist approach, the primary outcome analysis is conventionally deemed to be statistically significant if the p-value is less than 0.05 and non-statistically significant if the p-value is greater than or equal to 0.05 [24]. This binary classification system has long been accepted to be sub-optimal: of importance is not whether (an arbitrary) threshold of statistical significance has been crossed but rather the likely size or magnitude of effect, which can be determined by values supported by the confidence interval [31].

Consequently, so that readers can infer the range of treatment effects supported by the study, CONSORT guidelines specify that trials should report treatment effects together with confidence intervals [32]. This advice is mostly adhered to, at least in the higher impact journals [12,17,34]. However, reporting guidelines also urge the interpretation of results to be consistent with the study findings [20,32,40]. Whilst more difficult to assess, methodological reviews suggest that many RCTs report confidence intervals, but do not interpret them appropriately [10,11]. Although confidence intervals are now commonly reported, when it comes to interpretation it appears that most authors simply focus on whether the confidence interval includes the null, and then revert back to interpretation based on statistical significance. The persistence of this highly problematic situation led to the American Statistical Association publishing, in 2016, what became a highly cited statement on the correct use of *P*-values [38], considerably raising the profile of the problem [3].

2. Misinterpretation of confidence intervals

Interpretation of study findings consistent with the study results is nuanced, because conclusions should be concordant not only with the findings for the primary outcome, but must also consider the harms, costs and other evidence. However, inconsistency between the study results and overall conclusions mostly stem from a misinterpretation of the statistical results for the primary outcome [10,11].

The first and most problematic issue is when inconclusive trials are interpreted as providing definitive evidence that the treatment under evaluation is ineffective [10]. This is referred to as conflating no evidence of a difference with evidence of no difference (i.e., conflating absence of evidence with evidence of absence) [1]. Whilst mostly considered as a feature of misinterpretation of *P*-values, this can arise even when investigators report point estimates and confidence intervals (in accordance with CONSORT reporting guidelines), but fail to interpret them appropriately. This misinterpretation can be particularly problematic in small trials, as small trials are more likely to be inconclusive, especially those for which investigators might have been over optimistic about the target effect size. Take for example, the ORBITA trial - a comparison of percutaneous coronary intervention (PCI) vs. placebo for angina relief in 230 patients for which the primary outcome (exercise capacity in seconds) was non-statistically significant (PCI minus placebo 16.6 seconds, 95% CI −8.9 to 42.0, $P = 0.20$) and which interprets the overall finding in the abstract as "*PCI did not increase exercise time by more than the effect of a placebo*" [2]. Yet, this conclusion is inconsistent with the confidence interval for the primary outcome, which includes the target effect size (+ 30 seconds) and so presumably was an effect considered to be of clinical importance. An appropriate interpretation here would either acknowledge the uncertainty by reflecting on the treatment effects supported by the confidence interval (up to a 42 second increase or reduction of up to 9 seconds), or an explicit acknowledgement that the confidence interval ruled out effect sizes considered clinically important (if this were the case).

The second issue is when trials interpret their primary outcome results with an over-emphasis on statistical significance rather than considering the confidence intervals [10]. This issue is more contentious as tight control of statistical significance levels prevents type-1 error inflation [16]. As an example of a trial that has handled this issue well, the RIGHT-2 trial (N=1,149) assessed the effectiveness of Glyceryl Trinitrate (GTN) when delivered to patients very early after a presumed stroke in a two-arm randomised blinded trial [4]. The primary outcome was poor function at 90 days. The study reported the odds ratio for poor outcome as 1.25 (95% CI 0.97-1.60, $P = 0.08$), interpreting this finding in the abstract as "*Prehospital treatment with GTN worsened outcomes in patients with intracerebral hemorrhage. Since these results could relate to the play of chance, confounding, or a true effect of GTN, further randomized evidence on the use of vasodilators in ultra-acute intracerebral hemorrhage is needed.*" To justify this interpretation, values supported by the confidence interval are interpreted in the main text: "95% CI covering a range from a clinically insignificant benefit (OR 0.97) to a clinically significant hazard (OR 1.60)". The authors used language in their interpretation which recognised the possibility of chance, and in the main text recognised that values supported by the confidence interval mostly suggested a harmful effect rather than benefit. Had the authors simply concluded the finding was "not-statistically significant" readers might not have gleaned the extent of the finding (i.e., that the confidence interval mostly supported harm rather than benefit).

These two case studies suggest that, despite some improvements in reporting practices, the error in interpretation made by conflating no evidence of a difference with evidence of no difference, and an over-emphasis on statistical significance, may be persisting. Moreover, though widely cited, some have questioned whether the American Statistical Associations statement, has had any real impact [26,27]. In this study, we describe the interpretation of non-statistically significant primary outcomes in trials published in four high impact journals after the publication of the American Statistical Association statement.

3. Objectives

Our specific objectives were to review a contemporary sample of abstracts from RCTs with non-statistically significant primary outcomes and describe the frequency with which (i) the findings are interpreted to imply evidence of no difference between treatments; and (ii) findings are interpreted with an over-emphasis on statistical significance. We also aimed to assess whether the primary outcome result is reported in accordance with the CONSORT guideline, namely as point estimate, confidence interval and both relative and absolute effects for binary outcomes. We hypothesized that despite high adherence to reporting of primary outcome results, authors are still not interpreting the findings appropriately. Of note, assessment of reporting of absolute effects for binary outcomes is considered a key component as reporting as it facilitates the interpretation of the magnitude of the clinical impact. Finally, we also aimed to document the key language used in the conclusion of the abstract and any supporting justifications for the conclusion of no difference. Our rationale for focusing on the abstract was that, whilst correct interpretation is ultimately important throughout the manuscript, it is particularly important to ensure correct interpretation in the study abstract [13].

4. Methods

4.1. Search strategy

We included two-arm superiority parallel design randomized trials from four high ranking general medical journals that all endorse CONSORT guidelines: the New England Journal of Medicine, Annals of Internal Medicine, the BMJ, and JAMA (The Journal of American Medical Association). Studies had to be published in English between 2017 and 2020, and have a non-statistically significant primary outcome. We a priori decided to review 50 studies to make the task manageable. We excluded studies which were not individually randomized (e.g., cluster randomized studies); studies which were not designed as superiority studies (e.g., non-inferiority and equivalence studies); studies that had more than two arms; and studies with an unclear or undefined primary outcome. We only included primary study reports, and so excluded protocols or secondary analyses. High impact general medical journals were chosen because journals that are high ranking are often expected to have better reporting standards [36,37]. Consequently, if evidence of misinterpretation is found in high impact journals, then it is likely that misinterpretation is present in lower impact journals. Finally, an additional inclusion requirement was that the study had to have an associated editorial – this requirement was included because of a related project assessing reporting practices in editorials using the same sample.

The search was conducted in the Ovid platform and searched both the Medline and Embase databases on the 3rd of February 2020 (Supplementary Table S1). Abstracts identified by the search were exported to Excel where duplicates were excluded. The remaining abstracts underwent initial eligibility screening (IJ). Full text reports were then obtained for those studies identified as eligible and the full text was reviewed to confirm eligibility. Reports were screened in chronological order in batches of five until the sample size of 50 was achieved. When there was any uncertainty about inclusion, the final decision was made by a second independent reviewer (KH).

4.2. Data abstraction and analysis

All data were abstracted from study abstracts only and focused on the primary outcome. Firstly, we determined whether results were reported in a way which allowed readers to determine if the findings supported clinically meaningful effects as per the CONSORT guidance on reporting results. To this end, we determined if the abstract results section: reported results as a summary of each arm; reported the effect size and associated confidence interval; and (for binary outcomes) reported both relative and absolute differences (where absolute differences are known to be more clinically interpretable). These extractions were completed by a single reviewer (IJ) since it was not the primary objective of our review and was not considered a subjective assessment.

We then reviewed the conclusions section of the abstract. Two reviewers (MT and KH) independently classified the overall abstract conclusion into one of four mutually exclusive categories, as either (i) stating evidence of no difference; (ii) an over-emphasis on statistical significance; iii) showing preference for one treatment over the other; and iv) acknowledging some uncertainty in the overall conclusions. Discrepancies were resolved through discussion. The first category (stating no evidence of a difference) likely represents conflating evidence of absence with absence of evidence (misinterpretation). The second category represents those which place an over-emphasis on statistical non-significance in abstract conclusions (technically correct but arguably uninformative). The category "showing preference for one treatment over the other" is included as despite non-statistical significance, some trial reports do make definitive conclusions about one treatment being preferable; and the category "acknowledging uncertainty" included to identify those which conclude uncertainty around the effectiveness of the active treatment. Further information on these classifications is provided in Table 3. In addition, we determined whether studies provided an appropriate justification for their conclusion, either by reference to values supported by the confidence interval or by reference to other contextual information (as this might override overall inferences focusing on the primary outcome). Finally, we also identified from each conclusion key phrases which we assessed as

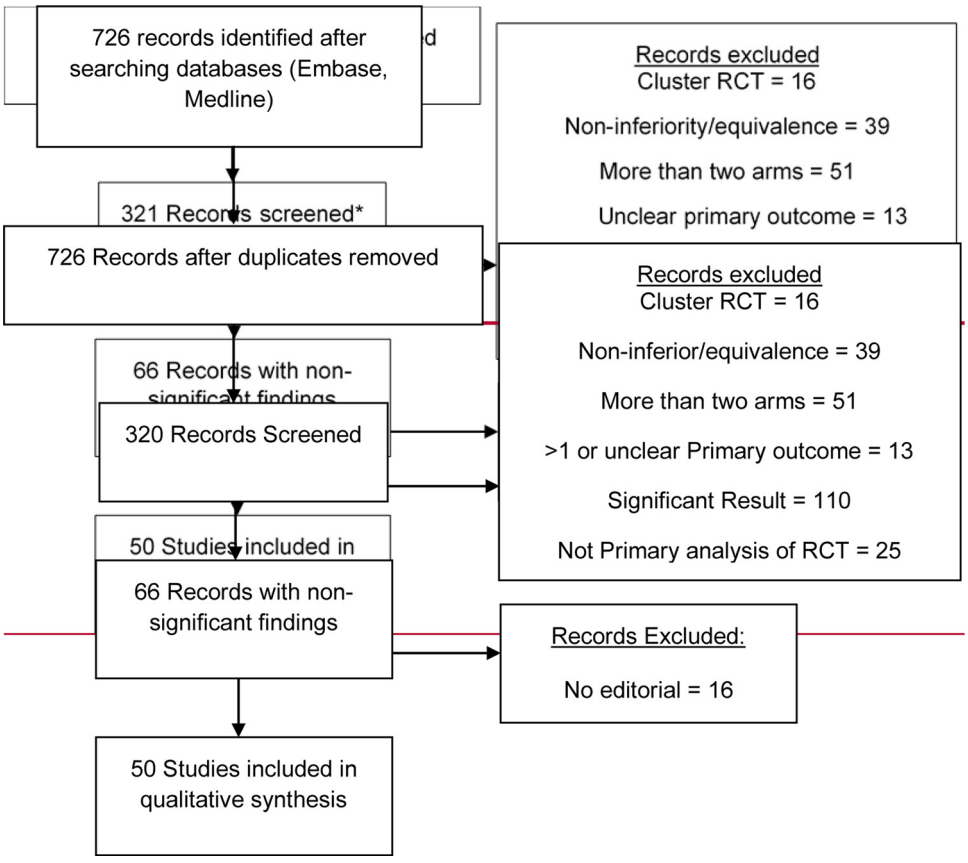

Fig. 2. PRISMA flow chart for included randomised control trials. Record screening was halted once 50 papers had been identified as meeting the inclusion criteria.

constituting the overall meaning of the conclusion. These key statements were then summarised in a narrative form.

5. Results

The database search yielded 726 papers that potentially met the inclusion criteria after duplicates were removed (Fig. 2). These were screened in date order, screening 320 papers, until our pre-specified sample size of 50 was achieved. The characteristics of the final 50 papers are summarized in Table 1. Most were published in either NEJM (18, 36%) or JAMA (25, 50%); just over half had a binary outcome (28, 56%) while 17 (34%) had a continuous outcome. The average (median) number of participants randomized (total across both arms) was 694 [IQR: 357 to 2,275]. The papers were published from 2017 to 2020.

Adherence to CONSORT guidance on the reporting of the primary outcome was high: 94% (47/50) of papers reported summaries of the outcome for each arm separately, 98% (49/50) reported the effect size, 96% (48/50) reported a confidence interval, 94% (47/50) reported the *P*-value, while 86% (43/50) reported all of these (Table 2). However, less than half tudies reported the absolute difference for binary outcomes: for the subset of 28 studies with binary primary outcomes only 21% (6/28) reported both relative and absolute measures of effect; 32% (9/28) reported only an absolute measure of effect while the majority reported only a relative measure (13/28, 46%).

In 56% (28/50) of the studies the overall conclusion of the abstract was classified as suggesting no difference in effectiveness between the two treatments being compared (Table 2); whilst 38% (19/50) were classified as placing an over-emphasis on statistical significance (see Supplementary Table 2 for a breakdown by each study). For the three remaining studies, only 1 (2%) clearly reported in the abstract conclusion that the finding was inconclusive and that more research was required; 2 (4%) concluded one treatment was preferable to the other. When considering appropriate justification for conclusions, in only one study (2%) was there a consideration of values supported by the confidence interval in the overall conclusion of the abstract and only three (6%) considered other contextual evidence (Table 2).

Most studies summarized the interpretation in the conclusion of the abstract using a phrase which was suggestive of no difference between the two treatments under comparison (Table 3). Examples of such phrases included “treatment A did not improve (or reduce, or increase) out-

Table 1. Characteristics of included studies

Characteristic	All
Journal	N = 50
New England Journal of Medicine	18 (36%)
JAMA	24 (48%)
Annals of Internal Medicine	3 (6%)
The BMJ	5 (10%)
Year of Publication	
2019	15 (30%)
2018	12 (24%)
2017	23 (46%)
Outcome Type	
Binary	28 (56%)
Continuous	17 (34%)
Survival	5 (10%)
Study size	
Number randomised (median, IQR)	694 [355–2,423]
Number randomised (range)	94 – 12,092

IQR, inter-quartile range.

comes compared to treatment B" or "the treatment has no benefit (or no impact, no effect)". Some studies used phrases such as "treatment A is no more effective than treatment B" or "treatment A was not superior to treatment B". Others made more reference to outcomes, for example "outcomes did not differ (or were similar) between the two treatments." Other studies were more ambiguous about the conclusion, averting to some notion of statistical significance, for example "outcomes did not differ significantly (or did not show any significant difference)". Of note, many reports used the term significance but without clarity about whether this related to statistical or clinical significance.

6. Discussion

6.1. Summary of findings

Almost all abstracts of RCTs published in high impact journals with non-statistically significant primary outcomes appropriately report treatment effects and confidence intervals, yet most make definitive conclusions about active treatments being no different to the comparator treatment, despite this being prima facia inconsistent with a non-statistically significant primary outcome result. Few made any reference to other contextual evidence to support this finding; and few justified this statement of no difference by giving due consideration to values supported by the confidence interval. In addition, a large number of studies unhelpfully provide no informative interpretation: in the overall conclusion they simply state that the result is non-statistically significant, despite having reported confidence

Table 2. Summary reporting of results of primary outcome in abstract

	Number adhering (%)
Adherence to CONSORT guidance on reporting of primary outcome result	
Numerical reporting of results	N = 50
Results for each arm	47 (94%)
Effect size	49 (98%)
Confidence interval	48 (96%)
P-value	47 (94%)
All of the above	43 (86%)
Results for binary outcomes	N = 28
Binary Outcomes	28 (56%)
Relative and absolute effect	6 (21%)
Relative effect only	13 (46%)
Absolute effect only	9 (32%)
Interpretation of primary outcome result in abstract conclusion	
Classification of interpretation^	N = 50
No difference	28 (56%)
Over-emphasis on statistical significance	19 (38%)
One treatment preferable	2 (4%)
Inconclusive	1 (2%)
Use of appropriate justification for interpretation	N = 50
Consideration of values supported by CI	1 (2%)
Consideration of other contextual evidence	3 (6%)

CI, Confidence Interval
^ see Table 3 for definitions and examples

Table 3. Classification and example terms used in abstract conclusions

Classification	Definition	Phrases used in reporting conclusion of abstract
No difference	A statement that suggests there is evidence of no difference between the treatments.	"similar rates of safety and efficacy" "did not reduce" "no effect" "did not improve" "does not prevent" "was no more effective than placebo" "did not concur survival advantage" "was not found to reduce" "was not associated with" "no evidence of clinical benefit"
Over-emphasis of statistical significance	A statement that suggests the findings are not statistically significant without any further interpretation.	"did not significantly improve" "did not differ significantly" "did not show any significant difference" "no significant difference" "did not result in significantly lower" "did not result in a rate that was significantly lower" "did not have significantly better outcomes" "did not result in a statistically significant difference" "no significant difference" "did not significantly improve" "did not result in significantly lower risk" "did not significantly improve symptoms" "did not provide significant benefit"
Showing preference for one treatment justification	A statement that is directive in its conclusion suggesting a meaningful difference between the two treatments being compared.	"results in less severe PPH"
Inconclusive	A statement that acknowledges some uncertainty over the finding.	"and confidence intervals for the treatment effect that included the minimally important difference"

intervals in the results section. This finding suggests that authors are abiding by reporting guidelines when it comes to the types of statistical measures that should be reported, but fall back on statistical significance when it comes to interpretation. Clear statements that the study finding is inconclusive (i.e., when the confidence interval provides support for both benefit and harm) in reports of RCTs in high impact journals are rare. Despite high profile campaigns in 2016 to put a stop to this poor practice [38], our review demonstrates that the practice of misinterpretation is still highly prevalent.

6.2. Research in context

Scale of the problem: Correct interpretation of non-statistically significant primary outcome results from randomised controlled trials can be challenging but is very important [35]. Previous reviews of trials conducted between 2000 and 2017 have demonstrated that misinterpretation is common [3,10,11]. Our review of more recent trials conducted between 2017-2020, after publication of the American Statistical Association statement, shows that misinterpretation persists in the study abstract, arguably the most influential part of the trial report [20]. Repeated demonstration of this problem ultimately signifies that the contribution of randomised trials to evidence-based medicine is being undermined because they are not interpreted properly.

Why is it happening: Misinterpretation might stem from a desire to make a definitive conclusion about effect. Investigators, authors, editors of journals, clinicians and patients all have a desire for definitive answers, and this might encourage investigators to translate their findings into definitive statements (e.g., no effect) even when this is not supported by the statistical findings. Yet, equally problematic, editors and reviewers might insist on specific language (e.g., non-significant), sometimes against the better judgement of the authors [33]. Often this insistence on specific language originates from the desire not to be seen to be creating any spin whereby borderline statistical significance is over-interpreted [6,29]. Indeed, this concern is likely behind the NEJMs continued focus on statistical significance, particularly when multiplicity adjustments are to be made [16]. Yet, confidence intervals can be adjusted for multiplicity – essentially becoming wider as the number of tests increases [9]. Furthermore, simply concluding a finding is non-statistically significant — whilst technically correct, is arguably unhelpful and not aligned with

the move away from statistical significance [40]. Moreover, if editors and reviewers are actively insisting on specific language, much of this language is misleading as in many study reports the language used is highly suggestive of no effect, when often at least a small (and indeed often moderate or even large) effect cannot be ruled out.

Solutions: The statement on P-values by the American Statistical Association, has been cited more than 4000 times, signifying the scale of the problem [38]. Yet, rather than being told what not to do, investigators need to be given clear direction on what is appropriate [26,27]. First and foremost the focus should not be placed on the point estimate as this can be misleading, especially in small samples [3,23]. Others have attempted to suggest appropriate language for interpretation and have tried to provide guidance for good practice [10,18,19]. Reporting on absolute scales also needs to be improved, because without an absolute measure of effect it is unlikely that a real impact of clinical importance can be made [22]. The perpetuation and frequency of misinterpretation, coupled with arguably no satisfactory solution have motivated some to suggest an entire paradigm shift is needed with Bayesian methods holding the solution [10,39]. For example, a Bayesian reanalysis of the RECOVERY convalescent plasma trial, with a non-significant primary outcome (survival to 28 days, rate ratio 1.00, 95% CI 0.93–1.07; $P = 0.95$), when analysed with a vague prior, estimated the likelihood of any benefit to be 64% [15]. When framed like this, readers can clearly see how this finding differs from one that conclusively indicates no benefit.

7. Limitations

Focus on the abstract: Our assessment of reporting focused on study abstracts which are limited by word counts [20]. More nuanced interpretations of study findings are likely contained in full text discussion sections. For example, other important contextual information includes other evidence in the field, harms or side effects, costs, invasiveness and patient preferences [30]. Based on such contextual information, authors might have been justified in making a conclusion that appears to be inconsistent with the primary outcome result. However, this does not undermine the need to ensure the conclusion in the abstract is aligned with results reported in the abstract, and if there is other mitigating evidence to override the primary outcome result this should be transparent. Concise phrases such as "whilst not statistically significant, the findings do not rule out a positive effect," or "the results are consistent with both no effect, a small effect, or an effect of clinical importance", or "the results are mostly consistent with a negative effect but do not rule out a small positive impact" convey the uncertainty and implicitly acknowledge values supported by the confidence interval and do so using few words.

True null effects: We did not attempt to determine if the authors' conclusions of no effect were consistent with the effects supported by the confidence intervals. Full interpretation of confidence intervals requires consideration and knowledge of what are clinically important effects [5,8,28]. Yet, minimum clinically important effects are seldom reported [7]. The target effect size used at the design stage might be inferred to be at least as small as the minimally clinically important effect [7]. Under this assumption, we could have considered whether confidence intervals for primary outcome results included target effect sizes, thus indicating that clinically important effects could not be ruled out. However, target effect sizes are rarely based on true minimally important differences, and are mostly determined by values that are believed to be feasible or amenable to detection at affordable sample sizes [7]. Others have shown a marked mismatch between target effect sizes and average treatment effects – with target effect sizes being much greater than observed average effects [Rothwell 2018]. Due to these difficulties we therefore did not attempt to make our own assessment of the study finding, but rather simply described practices of stating no effect without clearly stating the confidence interval ruled out clinically important effects. Thus, it might be possible that some studies which reported an overall interpretation of no difference between the two treatment arms were correct in this interpretation: some of these associated confidence intervals might well have excluded clinically important differences, although this was not transparent in the abstract [21].

Subjectivity: Whilst our assessments of abstract conclusions were undertaken independently and in duplicate, the assessment requires some subjective judgement. We identified common terms and phrases used to describe and interpret key results from trials, and unsurprisingly found little variation across trials, many using what might be referred to as "stock phrases". When classifying language used to describe key findings, others have used a finer classification than the one used here. For example, a previous review [10] separated statements like "the intervention was not beneficial" and "outcomes were similar" from "no difference" – but all such statements are likely to be interpreted similarly in practice. A reviewer pointed out that statements such as "mean outcomes were similar in both groups" might be technically correct if authors were referring to point estimates; however, such statements do not provide any reflection on the uncertainty associated with the confidence interval [23]. Related to this, many of the conclusions used what might be considered "directional claims" indicating that the active intervention "did not reduce" the outcome. In our case study evaluating the impact of GTN which found that the intervention was associated with mostly harmful outcomes, such an interpretation would have been justified. However, many of the studies which used directional claims did so despite their confidence intervals including moderate positive and negative effects, making no mention of confidence intervals mostly ruling out effects in one direction. Thus, many of the com-

mon phrases used to describe key findings are technically defensible, and so statistically might be considered as correct. However, the wording used to describe key findings needs to be both technically correct and convey the full meaning of the findings so it is properly understood by those who need to implement the findings.

Others also have differentiated "no clinical difference" from "no difference" – and whilst we agree there is a distinction here, in our review such statements (perhaps generously) were assumed to mean the confidence interval did not include clinically important effects. Of some note, it was almost never clear whether "significance" referred to clinical or statistical significance. We classified any mention of "significance" as referring to statistical significance unless clear otherwise. In a handful of studies this might be questionable, for example when the phrase "no significant improvement" or "no significant benefit" is used. However, these statements lack clarity and any reclassification would only shift between poor practice categories in our assessment.

Representativeness: Finally, we limited our assessment to four top journals. Top journals usually lead the way in good practice of reporting and methods [12,17]. It is possible that in other lower impact journals the problem of misinterpretation of primary outcome results is even more problematic. It might equally be the case that lower impact journals can improve the consistency of interpretation with the primary outcome result with less pressure to report only studies with definitive conclusions [37]. Furthermore, our choice of the four journals was arbitrary and might not be representative of other top journals. Moreover, unexpectedly, but perhaps predictable in hindsight, a large majority of our study reports were published in either JAMA or NEJM and so our review might be regarded as a review of practices in these two journals.

8. Conclusion

The phenomenon of conflating absence of evidence with evidence of absence seems to stubbornly persist, even against better judgement. Clear non-technical guidance is needed to help researchers interpret their findings. Statements that are technically correct can still mislead. Almost certainly the paradigm of frequentist statistics perpetuates the problem. A paradigm shift to Bayesian methods might help yield more interpretable answers.

Copyright

Transparency

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2022.01.014.

References

[1] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311(7003):485.
[2] Al-Lamee R, Thompson D, Dehbi HM, Sen S, Tang K, Davies J, et al. ORBITA investigators. percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial. Lancet 2018;391(10115):31–40.
[3] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567(7748):305–7.
[4] Bath PM, Woodhouse LJ, Krishnan K, Appleton JP, Anderson CS, Berge E, et al. Prehospital Transdermal Glyceryl Trinitrate for Ultra-Acute Intracerebral Hemorrhage: Data From the RIGHT-2 Trial. Stroke 2019;50(11):3064–71.
[5] Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 2002;14(2):109–14.
[6] Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA 2010;303(20):2058–64.
[7] Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA, et al. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. Trials 2018;19(1):606.
[8] Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7(5):541–6.
[9] Efird JT, Nielsen SS. A method to compute multiplicity corrected confidence intervals for odds ratios and other relative effect estimates. Int J Environ Res Public Health 2008;5(5):394–8 PMID: 19151434; PMCID: PMC3699999. doi:10.3390/ijerph5050394.
[10] Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open 2019;9(9):e024785.
[11] Gewandter JS, McDermott MP, Kitt RA, Chaudari J, Koch JG, Evans SR, et al. Interpretation of CIs in clinical trials with non-statistically significant results: systematic review and recommendations. BMJ Open 2017;7(7):e017288.
[12] Ghimire S, Kyung E, Kang W, Kim E. Assessment of adherence to the CONSORT statement for quality of reports on randomized controlled trial abstracts from four high-impact general medical journals. Trials 2012;13:77.
[13] Gonon F, Konsman J, Cohen D, Boraud T. Why most biomedical findings echoed by newspapers turn out to be false: the case of attention deficit hyperactivity disorder. PLoS ONE 2012;7(9):e44275.

[14] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al., GRADE Working Group GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924–6 Mar;567(7748):305-307.
[15] Hamilton FW, Lee T, Arnold DT, Lilford R, Hemming K. Is convalescent plasma futile in COVID-19? A Bayesian re-analysis of the RECOVERY randomized controlled trial. Int J Infect Dis 2021;109:114–17.
[16] Harrington D, D'Agostino RB Sr, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381(3):285–6.
[17] Hays M, Andrews M, Wilson R, Callender D, O'Malley PG, Douglas K. Reporting quality of randomised controlled trial abstracts among high-impact general medical journals: a review and analysis. BMJ Open 2016;6(7):e011082.
[18] Harrell 2021 https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/934 accessed 15 July 2021
[19] Hemming K, Taljaard M. Why proper understanding of confidence intervals and statistical significance is important. Med J Aust 2021;214(3):116–18 e1.
[20] Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al., the CONSORT Group CONSORT for reporting randomised trials in journal and conference abstracts. Lancet: 2008;371:281–3.
[21] Glasziou P, Doll H. Was the study big enough? Two café rules. Evid Based Med 2006;11(3):69–70.
[22] Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003;327(7417):741–4.
[23] Finch S, Cumming G. Putting research in context: understanding confidence intervals from one or more studies. J Pediatr Psychol 2008;34(9):903–16 2009 Oct.
[24] Jones MP, Beath A, Oldmeadow C, Attia JR. Understanding statistical hypothesis tests and power. Med J Aust 2017;207(4).
[25] Kahan BC, Jairath V. Outcome pre-specification requires sufficient detail to guard against outcome switching in clinical trials: a case study. Trials 2018;19(1):265.
[26] Matthews R, Wasserstein R, Spiegelhalter D. The ASA's p-value statement, one year on. Significance 2017;14(2):38–41.
[27] Matthews R. The p-value statement, five years on. Significance 2021;18:16–19. doi:10.1111/1740-9713.01505.
[28] McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA 2014;312(13):1342–3.
[29] Nascimento DP, Almeida MO, Scola LFC, Vanin AA, Oliveira LA, Costa L, et al. Letter to the Editor - Not even the top general medical journals are free of spin: A wake-up call based on an overview of reviews. J Clin Epidemiol Nov 2021;139:232–4 Epub 2021 Jun 29. PMID: 34214625. doi:10.1016/j.jclinepi.2021.06.016.
[30] Pocock SJ, Stone GW. The Primary Outcome Fails - What Next? N Engl J Med 2016;375(9):861–70.
[31] Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362–3 10.1056.
[32] Schulz KF, Altman DG, Moher Dfor the CONSORT Group. CONSORT 2010. Statement: updated guidelines for reporting parallel group randomised trials. Ann Int Med 2010;152(11):726–32.
[33] Shaqman M, Al-Abedalla K, Wagner J, Swede H, Gunsolley JC, Ioannidou E. Reporting quality and spin in abstracts of randomized clinical trials of periodontal therapy and cardiovascular disease outcomes. PLoS One 2020;15(4):e0230843.
[34] Turner L, Shamseer L, Altman DG, Schulz KF. Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev 2012;1:60.
[35] Resnik D. Scientific research and the public trust. Sci Engineer Eth 2010;17(3):399–409.
[36] Saha S, Saint S, Christakis D. Impact factor: a valid measure of journal quality. J Med Library Associat 2003;91(1):42–6.
[37] Tressoldi P, Giofré D, Sella F, Cumming G. High impact=high statistical standards? not necessarily so. PLoS ONE, 2013;8(2):e56180.
[38] Wasserstein Ronald L, Lazar Nicole A. The ASA statement on p–values: context, process, and purpose. The American Statistician 2016;70(2):129–33.
[39] Yarnell CJ, Abrams D, Baldwin MR, Brodie D, Fan E, Ferguson ND, et al. Clinical trials in critical care: can a Bayesian approach enhance clinical and scientific decision making? Lancet Respir Med 2021;9(2):207–16.
[40] Young PJ, Nickson CP, Perner A. When should clinicians act on non-statistically significant results from clinical trials? JAMA 2020;323(22):2256–7.