Key research skills

Why proper understanding of confidence intervals and statistical significance is important

Karla Hemming[1]

Monica Taljaard[2]

1 Institute of Applied Health Research, University of Birmingham, Edgbaston, UK.

2 Ottawa Hospital Research Institute, Ottawa, Canada.

k.hemming@bham.ac.uk

Series editors

John R Attia

Michael P Jones

doi: 10.5694/mja2.50926

Guidelines for reporting results from randomised trials have long underscored the importance of confidence intervals.[1] Confidence intervals are informative as they show the likely range of effect sizes supported by the findings, whereas *P* values dichotomise the findings based on statistical significance at an arbitrary cut-off.[2] Reviews of contemporary trials show that researchers mostly adhere to this advice.[3] Despite this, misinterpretation stubbornly persists.[4] First, many trials are interpreting absence of evidence as evidence of no effect, concluding an intervention is ineffective when in fact the results suggest its effectiveness is uncertain. Second, in some trials it might be correct to conclude that a treatment is effective (or harmful), despite the non-statistically significant result; yet researchers persist in unhelpful language such as "not statistically significant".

Thus, while researchers are abiding by reporting guidelines and including confidence intervals, these are rarely fully interpreted in the conclusions (ie, researchers are not abiding by the philosophy underpinning the reason for the guidelines). It is this paradox that led to recent campaigns demanding appropriate interpretation of confidence intervals.[5,6] Despite availability of publications addressing the statistical philosophy underpinning hypothesis testing, there is a dearth of practical guidelines for investigators, reviewers and editors in correct interpretation of findings from randomised controlled trials.

Here we provide a practical guide, bridging the gap between statistical philosophy and the desire to draw conclusive findings from most trials (Box 1). To this end, we provide recommendations for the interpretation of the primary outcome result, where we urge interpretation of the full range of the confidence interval and its overlap with effect sizes considered to be clinically important. While we advocate for a more holistic interpretation considering contextual factors, we urge transparency in these arguments. We illustrate these recommendations using two topical case studies (online Supporting Information).[7,8] The first study aimed to determine whether the efficacy of the N95 respirator in controlled settings could be maintained in real life, where compliance may be suboptimal. The trial reported a non-significant finding which it interpreted as "no significant difference".[7] The second study investigated whether lopinavir–ritonavir provides any treatment benefit in patients with severe coronavirus disease 2019, reporting non-significant findings which it interpreted as "no difference" or "no benefit".[8]

Practical guide

Interpreting a confidence interval

The imperfect nature of any approach to hypothesis testing is now widely recognised.[2] One approach, advocated by Neyman–Pearson, uses an objective but arbitrary cut-point (usually a *P* value of 0.05) for statistical significance. On the other hand, Fisher argued for an approach based on a continuum with no set threshold, also arguing for the consideration of other contextual factors. However, neither approach acknowledges the importance of the size of any treatment effect. Focus therefore shifted to the reporting of confidence intervals.[9] The confidence interval can be interpreted as providing a range of treatment effects supported by the study. Not all values within the interval are equally supported: those closer to the point estimate have more support, and support tapers the closer to the bounds of the interval.

When interpreting confidence intervals in relation to clinically important effect sizes (see below), primary outcome results can be directive despite not being statistically significant. This can arise when the confidence interval excludes a clinically meaningful benefit (or harm) (Box 2, row 2). Directive, yet not statistically significant results, can also arise when the confidence interval mostly overlaps with values indicative of benefit (or harm), that is, when the interval covers treatment effects mostly in one direction (Box 2, rows 1 and 3). In reality, statistically non-significant results can also arise in situations where the confidence interval is wide and includes treatment effects that are both beneficial and harmful. Such results should be interpreted as inconclusive (Box 2, row 4). The primary outcome in the ResPECT trial[7] (Supporting Information) is an example of a

1 Recommendations for interpreting results from randomised controlled trials

The interpretation of the findings should be informative:

- Directive conclusions from randomised trials are desirable. Investigators should refrain from using unhelpful statements like "statistically not significant" in the overall conclusion of the trial findings.

Confidence intervals should be properly interpreted:

- Interpretation of the trial findings should consider the range of effects supported by the confidence intervals.
- Values at the tails of the confidence interval are less supported by the data from the trial.

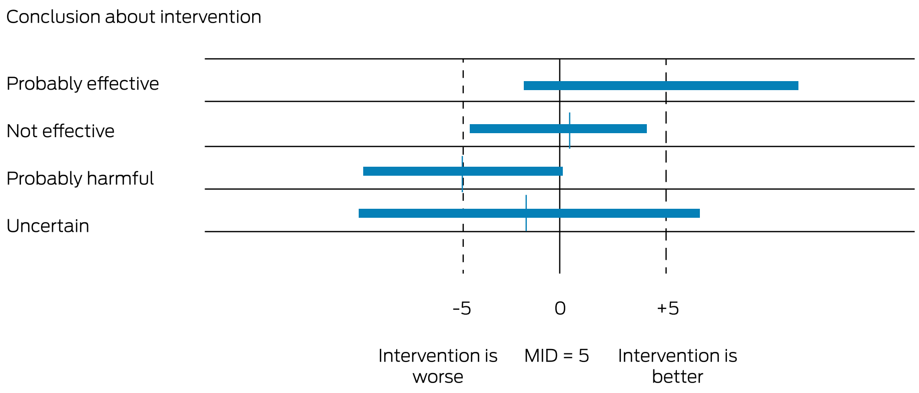
Clinically important treatment effects must be considered:

- Only when the confidence interval conclusively rules out (ie, does not overlap with) any treatment effect considered to be clinically important can a directive conclusion of no effect be made.
- A confidence interval result that unequivocally includes both benefit and harm should be interpreted as inconclusive.

The overall conclusion should be justified:

- Overall conclusions should be contextualised. This contextualisation includes not only the primary outcome but also associated harms, costs, secondary outcomes or other supplementary considerations. These arguments should be transparent.

2 Illustrative example of how to interpret a statistically non-significant confidence interval



MID = minimally (clinically) important difference. This illustration considers a continuous outcome for which changes in the region of 5 points are considered to be probably unimportant. The MID is therefore around 5. The exact position of the point estimate has no relevance to the interpretation. All confidence intervals are symmetric.

statistically non-significant primary outcome, but which probably rules out any meaningful benefit, whereas the outcome mortality in the lopinavir–ritonavir trial[8] (Supporting Information) is an example of a statistically non-significant result which is probably inconclusive.

Clinically important treatment effects

Confidence intervals need to be interpreted with an understanding of what are clinically important changes in outcomes — referred to as clinically important treatment effects. The notion of the minimum clinically important effect size (ie, the smallest effect size that is thought to be of any clinical importance) will be familiar to many researchers. Ideally, the sample size should be based on being adequately powered to detect this effect size.[10] However, given the nature of research which is often constrained by limited budgets and resources, sample size calculations are often based on effect sizes that are thought to be achievable or that yield a feasible sample size.[11] Second, the minimally clinically important difference in a superiority trial is related to the non-inferiority or equivalence margin considered in non-inferiority or equivalence trials.[12] Minimum clinically important effect sizes thus inform what a clinically important effect is. Ideally, what constitutes a clinically important effect should be pre-specified, well justified and include opinions of both clinicians and patients.[13] Moreover, it should not be taken as absolute and should be interpreted on a continuum. For trials that evaluate effects on outcomes such as mortality, any positive effect of the intervention (however small) might be clinically important. Reporting results on the absolute scale, perhaps as a number needed to treat, can aid in interpretation of clinical importance.[14] No minimally important clinical differences were specified in either of the two studies considered here,[7,8] but in both examples we use logical reasoning to consider what plausible smallest important differences might be (Supporting Information).

Importance of informative conclusions

In almost all situations, a technically correct conclusion of “statistically non-significant” is unhelpful. Those seeking information wish to know whether the findings are inconclusive (more research is required) or whether the trial can be directive in its conclusions. Both studies are statistically not significant for their primary outcomes; however, concluding that the study was “non-significant” in the study conclusions is unhelpful. Despite being statistically non-significant, the primary outcome result from the ResPECT trial suggests there is probably no benefit from the N95 respirator; moreover, the confidence interval also covers regions which might be considered as clinically important increases in risk. However, when considering the full range of treatment effects supported by the confidence intervals for both primary and secondary outcomes for the lopinavir–ritonavir trial, we see that the results are compatible with both benefit and harm, and this result is therefore inconclusive.

Holistic interpretation

There are of course many considerations other than the primary outcome result. First and foremost, it is necessary to consider the robustness of the trial design, risks of bias, and generalisability. Both the ResPECT trial and the lopinavir–ritonavir trial appear to be free from any obvious bias. While the interpretation of the trial findings should focus on the result of the primary outcome, other contextual factors are important. These might include secondary outcomes, harms, costs, and evidence from other trials. This more holistic interpretation is endorsed by the CONSORT statement.[9]

In the case of N95 respirators, the community wishes to know whether N95 respirators, which are more expensive and uncomfortable to wear, provide any extra benefit over medical masks. Concluding any risk of harm from the N95 respirator mask appears counter-intuitive, but might be a reflection of risk compensation. Moreover, given this suggestion of increased risk appears only at the extreme tail of the confidence interval, it might reflect random chance. If the N95 respirator is truly compatible with harm, the secondary outcomes would likely have shown that signal too. In fact, all of the secondary outcomes seemed to indicate either no effect or a likely protective effect. Thus, a reasonable interpretation based on the primary outcomes is that N95 respirators probably provide no added protection. The primary outcome result for the lopinavir–ritonavir trial is uncertain; other outcomes were also mostly uncertain. Evidence from other trials was rapidly evolving, but none pointed convincingly to any suggestion that lopinavir–ritonavir could be abandoned as an ineffective treatment, at least not just yet.

Summary

Evidence-based medicine requires careful execution of randomised trials of high internal validity and high generalisability. A growing body of literature on the conduct and reporting of randomised trials warns against such things as manipulation of outcome selection and multiplicity of analyses, and provides guidelines on good practice, such as pre-trial registration. Increasingly, investigators are adhering to this advice. However, investigators, reviewers and editors are still failing to correctly interpret statistically non-significant results. Clinical interpretation of trial results needs to shift to being centred on whether the results (ie, values supported by the confidence intervals) are consistent with a clinically important effect.

Pre-specification and justification of clinically important effect sizes should become the norm. Minimally important effect sizes have been a feature of sample size calculations, especially in non-inferiority trials, but they are fundamental for the interpretation of all randomised trials. Although there is as yet no consensus on how to determine these values, it does not mean this issue can be ignored. Reporting on absolute scales is almost certainly helpful here.

The interpretation of the primary outcome result is not the only consideration when determining the final conclusions. There may, for example, be side effects from treatments, cost considerations or issues of overtreatment or invasiveness and secondary supportive outcomes. These other considerations might lend support to an overall conclusion that the treatment is unlikely to be beneficial, despite a non-significant finding. However, it is crucial that there is transparency in how this conclusion is reached.

Clear and conclusive findings are more appealing to journal editors and to their readership. Sometimes, but not always, trials which are statistically not significant can still be directive. Unfortunately, many trials ultimately end up being uncertain simply because they are too small. Trials undoubtedly need larger sample sizes to reduce uncertainty and to ensure they are powered to detect clinically important effect sizes.[15]

Competing interests: No relevant disclosures.

Provenance: Commissioned; externally peer reviewed. ■

The unedited version of this article was published as a preprint on mja.com.au on 16 July 2020.

References are available online.

1 Rothman KJ. A show of confidence. *N Engl J Med* 1978; 299: 1362–1363.

2 Jones MP, Beath A, Oldmeadow C, Attia JR. Understanding statistical hypothesis tests and power. *Med J Aust* 2017; 207: 148–150. https://www.mja.com.au/journal/2017/207/4/understanding-statistical-hypothesis-tests-and-power

3 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: e011082.

4 Gewandter JS, McDermott MP, Kitt RA, et al. Interpretation of CIs in clinical trials with non-significant results: systematic review and recommendations. *BMJ Open* 2017; 7: e017288.

5 Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat* 2016; 70: 129–133.

6 Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019; 567: 305–307.

7 Radonovich LJ Jr, Simberkoff MS, Bessesen MT, et al. N95 respirators vs medical masks for preventing influenza among health care personnel: a randomized clinical trial. *JAMA* 2019; 322: 824–833.

8 Cao B, Wang Y, Wen D, et al. A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. *N Engl J Med* 2020; 382: 1787–1799.

9 Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 2010(340): c332.

10 Cook JA, Julious SA, Sones W, 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: 606.

11 Hislop J, Adewuyi TE, Vale LD, et al. Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. *PLoS Med* 2014; 11: e1001645.

12 Dunn DT, Copas AJ, Brocklehurst P. Superiority and non-inferiority: two sides of the same coin? *Trials* 2018; 19: 499.

13 McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA* 2014; 312: 1342–1343.

14 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988; 318: 1728–1733.

15 Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal. *Trials* 2018; 19: 544. ■