*The Medical Journal of Australia* – Pre-print – 16 July 2020

# Why proper understanding of confidence intervals and statistical significance is important: Covid-19 randomised trials as a case in point

**Word count:** 1571 (excluding abstract and boxes) 16 references

Monica Taljaard
Associate Professor
Ottawa Hospital Research Institute
Clinical Epidemiology Program
Ontario, Ottawa, Canada

Karla Hemming
Senior Lecturer
University of Birmingham
Institute of Applied Health Research
University of Birmingham
Edgbaston, West Midlands, United Kingdom of Great Britain and Northern Ireland

## Abstract

Statistically non-significant findings are often misinterpreted as evidence that there is no difference in effectiveness between two interventions. Despite this fallacy being widely known and accepted, it unfortunately persists as a widespread phenomenon.

To illustrate why correct interpretation of statistical significance and confidence intervals is crucial to good evidence-based medicine we use two topical case studies. The first case study concerns the question of whether N95 respirators are more effective than medical masks in preventing respiratory infections when used by health care providers. This question was assessed by a pragmatic well conducted cluster randomised trial that reported in 2019, designed to determine if the efficacy of the N95 respirator in controlled settings could be maintained in real life, where compliance may be sub-optimal. The trial reported a non-significant finding which it interpreted as “no significant difference”. However, despite being non-significant, the confidence interval for the primary outcome rules out anything but a small possibility of any clinically important benefit of the N95 respirator, and in fact supports more that the N95 respirator might be associated with a small amount of harm. Our second case study concerns the question of whether Lopinavir-Ritonavir provides any treatment benefit in patients with severe Covid-19. This question was assessed by a small single centre open label randomised trial. The trial reported non-significant findings which it interpreted as “no difference” or “no benefit”. In fact confidence interval for the primary and other important secondary outcomes are so wide they include the possibility of real benefit.

Our examples are topical and reiterate that correct interpretation of results from statistical significance testing is crucial. Firstly, a large pragmatic trial does not only not support the general use of the N95 respirator in clinical practice, but appeared to show it was associated with some increase in risk. Secondly, a small highly prominent randomised trial of a potential treatment (Lopinavir-Ritonavir) for patients with severe Covid-19 is compatible with the possibility of benefit, and the treatment deserves more research. These examples represent what is typical in the wider literature – inconsistent and inaccurate interpretation of statistical significance and confidence intervals. They serve to illustrate the point of why correct interpretation is crucial to evidence-based medicine.

## Background

Guidelines for reporting results from randomised trials have long underscored the importance of confidence intervals [Rothman 1978]. 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 [Jones 2017]. Reviews of contemporary trials reveal that researchers mostly adhere to this advice [Hays 2016]. Despite this, misinterpretation stubbornly persists [Gewandter 2017]. Firstly, 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. Secondly, 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 whilst researchers are abiding by reporting guidelines and including confidence intervals, these are rarely fully interpreted in the conclusions (i.e. 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 [Ronald 2016, Greenland 2019]. 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 [see Box 1], bridging the gap between statistical philosophy and the desire to draw conclusive findings from most trials. 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. Whilst we advocate for a more holistic interpretation considering contextual factors, we urge for transparency in these arguments. We illustrate these recommendations using two topical case studies [see Box 2 and 3].

## Practical guide

### *Interpreting a confidence interval*

The imperfect nature of any approach to hypothesis testing is now widely recognised [Jones 2017]. One approach, advocated by Neyman-Person, uses an objective but arbitrary cut-point (usually a p-value of 0.05) for statistical significance. Fisher on the other-hand, argued for an approach based on a continuum with no set threshold, also arguing for the consideration of other contextual factors. Yet neither approach acknowledges the importance of the size of any treatment effect. Focus thus shifted to the reporting of confidence intervals [Schulz 2010]. 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) (Figure 1, row 2). Directive, yet not statistically significant results, can also arise when the confidence interval mostly overlaps with values indicative of benefit (or harm), i.e., when the interval covers treatment effects mostly in one direction (Figure 1, row 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 (Figure1, row 4). The primary outcome in the ResPECT trial (Box 2) is an example of a statistically non-significant primary outcome, but which probably rules out any meaningful benefit, whereas the outcome mortality in the Lopinavir–Ritonavir trial (Box 3) 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 (i.e. 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 [Cook 2018]. 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 [Hislop 2014]. Secondly, the minimally clincially important difference in a superiority trial is related to the non-inferiority or equivalence margin considered in non-inferiority or equivalence trials [Dun 2018]. 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 [McGlothlin 2014]. 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 [Laupacis 1998]. No minimally important clinical differences were specified in either of the two case studies considered here, but in both examples we use logical reasoning to consider what plausible smallest important differences might be (see Box 2 and 3).

### *Importance of informative conclusions*

In almost all situations a technically correct conclusion of “statistically not-significant” is unhelpful. Those in need of information, wish to know whether the findings are inconclusive (more research is required) or whether the trial can be directive in its conclusions. Both case studies are statistically not-significant for their primary outcomes, however concluding “not-significant” in the study conclusions is unhelpful. Despite being statistically not-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. Whilst 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 [Schulz 2010].

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 thus 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, multiplicity of analyses and provides guidelines on good practice, such as pre-trial registration. Increasingly investigators are adhering to this advice. Yet, investigators, reviewers and editors are still falling foul of correct interpretation of statistically non-significant results. Clinical interpretation of trial results needs to shift to being centred on whether the results (i.e. 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 over treatment 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 [Rothwell 2018].

**References**

[Bender 2001] Bender R. Calculating Confidence Intervals for the Number Needed to Treat. Controlled Clinical Trials 22:102–110; 2001.

[Cao 2020] Cao B, Wang Y, Wen D, et al. A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2001282.

[Cook 2018] Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA, Ashby D, Emsley R, Fergusson DA, Walters SJ, Wilson ECF, Maclennan G, Stallard N, Rothwell JC, Bland M, Brown L, Ramsay CR, Cook A, Armstrong D, Altman D, Vale LD. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. Trials. 2018 Nov 5;19(1):606.

[Dunn 2018] Dunn DT, Copas AJ, Brocklehurst P. Superiority and non-inferiority: two sides of the same coin? Trials. 2018 Sep 17;19(1):499.

[Gewandter 2017] Gewandter JS, McDermott MP, Kitt RA, Chaudari J, Koch JG, Evans SR, Gross RA, Markman JD, Turk DC, Dworkin RH. Interpretation of CIs in clinical trials with non-significant results: systematic review and recommendations. BMJ Open. 2017 Jul 18;7(7):e017288. doi: 10.1136/bmjopen-2017-017288. Review. PubMed PMID: 28720618; PubMed Central PMCID: PMC5726092.

[Greenland 2019] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. doi: 10.1038/d41586-019-00857-9.

[Hays 2016] 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 Jul 28;6(7):e011082. doi: 10.1136/bmjopen-2016-011082. PubMed PMID: 27470506; PubMed Central PMCID: PMC4985789.

[Hislop 2014] Hislop J, Adewuyi TE, Vale LD, Harrild K, Fraser C, Gurung T, Altman DG, Briggs AH, Fayers P, Ramsay CR, Norrie JD, Harvey IM, Buckley B, Cook JA; DELTA group. Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. PLoS Med. 2014 May 13;11(5):e1001645. doi: 10.1371/journal.pmed.1001645. eCollection 2014 May. Review. PubMed PMID: 24824338; PubMed Central PMCID: PMC4019477.

[Jones 2017] Jones MP, Beath A, Oldmeadow C and Attia JR. Understanding statistical hypothesis tests and power, Med J Aust 2017; 207 (4): doi: 10.5694/mja16.01022

[Laupacis 1998] Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988 Jun 30;318(26):1728-33.

[McGlothlin 2014] McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014 Oct 1;312(13):1342-3. doi:10.1001/jama.2014.13128.

[Radonovich 2019] Radonovich LJ Jr, Simberkoff MS, Bessesen MT, Brown AC, Cummings DAT, Gaydos CA, Los JG, Krosche AE, Gibert CL, Gorse GJ, Nyquist AC, Reich NG, Rodriguez-Barradas MC, Price CS, Perl TM; ResPECT investigators. N95 Respirators vs Medical Masks for Preventing Influenza Among Health Care Personnel: A Randomized Clinical Trial. JAMA. 2019 Sep 3;322(9):824-833.

[Ronald 2016] Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

[Rothman 1978] Rothman KJ. A show of confidence. *N Engl J Med* 1978;299:1362–3.

[Rothwell 2018] 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 Oct 10;19(1):544.

[Schulz 2010] Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010 Mar 23;340:c332. doi: 10.1136/bmj.c332.

**Box 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 (i.e., 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.

**Box 2: The N95 respirators for health care professional's trial (ResPECT trial) [Radonovich 2019]**

*Background*

The ResPECT trial is a pragmatic, unblinded multicentre two arm cluster randomized trial of N95 respirators versus medical masks for the prevention of respiratory infections in health care workers. The trial enrolled 4,051 health care professionals working across multiple in-patient and out-patient settings across more than 190 sites across seven US hospitals between 2011 and 2015.

*Design*

The primary outcome was laboratory-confirmed influenza. The trial was designed to detect a target effect size of a 25% relative risk reduction (i.e. from 7.5% to 5.6%, or approximately a 2 percentage point reduction in risk) at 80% power and 5% significance. However, like most studies, this was a target effect size (i.e. the effect the trial was powered to detect) rather than being the minimally important difference.

*Primary outcome result*

The authors state that: "There were 207 laboratory-confirmed influenza infection events (8.2%) in the N95 respirator group and 193 (7.2%) in the medical mask group (difference, 1.0%, [95%CI, −0.5% to 2.5%]; P = .18) (adjusted odds ratio [OR], 1.18 [95%CI, 0.95-1.45])."

*Results on absolute scale*

Percentage point difference: 1% (95% CI: -0.5% to 2.5%); NNH 100 (95% CI: NNH 200, NNT 40)

*Investigators conclusions*

"Among outpatient health care personnel, N95 respirators vs medical masks as worn by participants in this trial resulted in no significant difference in the incidence of laboratory-confirmed influenza."

*Suggested clinically important differences*

In the ResPECT trial the event rate (laboratory confirmed infection) is about 7% under the medical mask group. Subject expertise will help in considering that might constitute a meaningful change in this event rate. We might say that differences of up to 1.5% might be clinically non-important. Whilst setting this value might be considered arbitrary, note that in our interpretation we interpret this value with considerable caution.

*Interpretation of the confidence interval*

The results for the primary outcome tell us that the true effect of N95 respirators might be to increase risk of laboratory confirmed influenza by up to 2.5% (i.e. a result consistent with clinically important harm) or to decrease the risk of influenza by up to 0.5% (i.e. small but clinically unimportant benefit). Whilst this trial is therefore compatible with an effect which might be beneficial or harmful the majority of the confidence interval supports small and unimportant effects. Because values at the tails of the confidence intervals are less likely, this trial is finds little support for any benefit or harm of the N95 respirator.

*Suggested Interpretation*

The ResPECT trial supports a conclusion that wearing N95 respirators has minimal impact or might even slightly increase the risk of influenza. This counter intuitive finding might be a consequence of risk compensation that can occur in the real world. However, because secondary outcomes do not support any increased risk it might well be be due to random chance. Moreover, this trial suggests there is no beneficial effect of N95 respirators over medical masks.

*Notes*

Values taken are those reported in trial reports, except the NNH, NNH confidence intervals were estimated using 1/reported absolute risk difference. An alternative would be the Wilson Score method [Bender 2001].

NNT: Number Needed to Treat for benefit; NNH: Number Needed to treat for Harm; CI: Confidence Interval.

**Box 3: A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19 [Cao 2020]**

*Background*

The Lopinavir-Ritonavir trial is an unblinded single centre two arm individually randomized trial of Lopinavir-Ritonavir treatment for 14 days versus standard care for the treatment severe illness in patients with laboratory confirmed SARS-CoV-2. The trial enrolled 199 patients between 18th January and 3rd February 2020.

*Design*

The primary outcome was time to clinical improvement on a 7-point ordinal scale, but this is a difficult outcome to interpret. Moreover, in hospitalised patients with confirmed SARS-CoV-2 infection mortality is clearly an important outcome. We therefore focus on mortality here. Under the control arm around 20% of patients died. A trial with 200 patients has 90% power to detect a 15% absolute risk reduction (i.e. invention mortality rate of 5%).

*Results*

The trial observed a median time to clinical improvement of 16 days (IQR: 13.0 to 17.0) in the intervention arm and 16 days (IQR 15 to 18) in the control arm. For the primary outcome the hazard ratio for clinical improvement was 1.24 (95% CI: 0.90 to 1.72).

Mortality at 28 days was 19.2% in Lopinavir–Ritonavir group and 25% in the standard-care group (risk difference, −5.8 percentage points 95% CI: −17.3 to 5.7).

*Result on absolute scale*

Percentage point difference (mortality): -5% (95% CI: -17.3% to 5.7%); NNT 20 (95% CI: NNT 6, NNH 18)

*Investigators conclusions*

“In hospitalized adult patients with severe Covid-19, no benefit was observed with Lopinavir–Ritonavir treatment beyond standard care.”

*Suggested minimally important differences*

The trial had power to detect a 15% absolute risk reduction in mortality. This is a very large effect size that, whilst clearly of clinical importance, is arguably not likely to be achieved by many drug interventions. Smaller effect sizes would also be clinically important. Reducing mortality even by 1 percentage point is likely to be considered important, provided there are no major adverse side effects.

*Interpretation of the confidence interval*

The results for the outcome mortality tell us that the true effect of Lopinavir–Ritonavir treatment might be to increase risk of mortality by up to 5.7% (i.e. a result consistent with clinically important harm) or to decrease the risk of mortality by up to 17.3% (i.e. large clinically unimportant benefit). This trial is therefore compatible with an effect which might be beneficial or harmful, and the range of these effects so large that the result is uncertain.

*Suggested Interpretation*

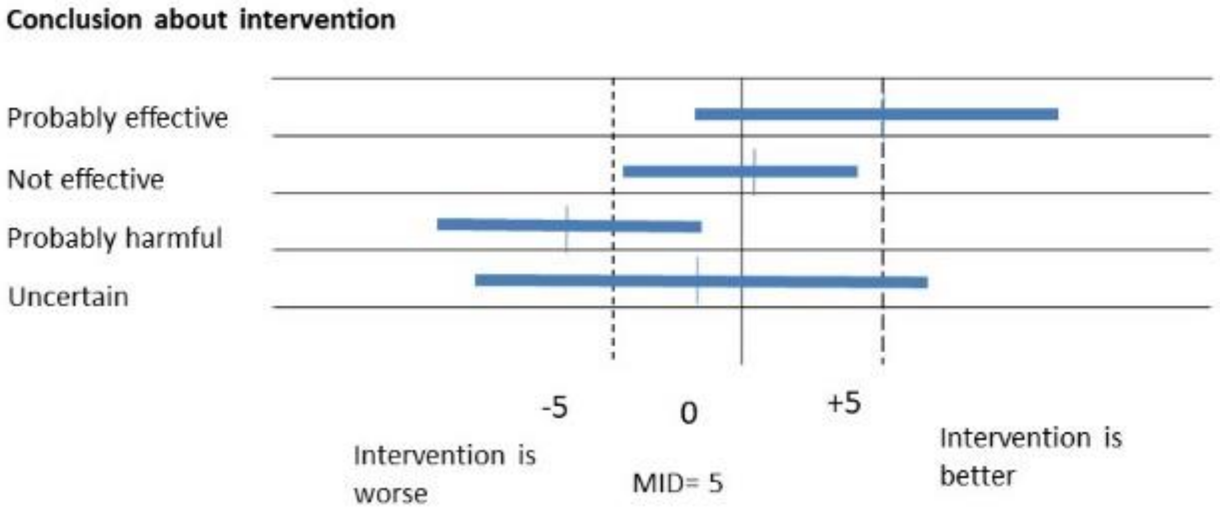
Mortality, despite not being pre-specified as the primary outcome is arguably the more important outcome in this trial. The trial findings are mostly supportive of positive effects but the trial does not rule out small to moderate harm. Secondary outcomes are mostly also inconclusive. There is insufficient evidence to support whether treatment with Lopinavir–Ritonavir can improve mortality in patients with severe Covid-19. More research is needed.

*Notes*

Values taken are those reported in trial reports, except the NNT, NNT confidence intervals were estimated using 1/reported absolute risk difference. An alternative would be the Wilson Score method [Bender 2001].

NNT: Number Needed to Treat for benefit; NNH: Number Needed to treat for Harm; CI: Confidence Interval.

Figure 1: Illustrative fictitious example of how to interpret a statistically non-significant confidence interval (MID: minimal important difference, or minimally important effect size).



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