Results should not be reported
as statistically significant or statistically non-significant

"Statistical significance" is so commonly misreported and misinterpreted, that we recommend that terms such as 'not significant', 'not statistically significant', 'significant', 'statistically significant', 'trend towards [an effect]', 'borderline significant') should not be used in EPOC reviews. In general point estimates and confidence intervals, when possible, or p-values should be reported. Plain language should be used to describe effects based on the size of the effect and the quality of the evidence. (See Worksheets for preparing summary of findings tables using GRADE).

Authors frequently make yes/no judgments about whether an intervention does or does not work based tests of statistical significance (e.g. a p value < 0.05 or a confidence interval that excludes no effect). This is likely to be misleading, as explained below. Imprecision of an estimate of effect is only one of several factors that may decrease our confidence in an estimate of effect.[1] It should be interpreted, together with other factors that can increase or decrease our confidence. (See Worksheets for preparing summary of findings tables using GRADE).

A common mistake made in instances when evidence is inconclusive is the confusion of a lack of evidence of an effect (due to imprecision) with 'evidence of no effect'.[2] It is wrong to claim that inconclusive evidence (when there is a wide confidence interval that includes potentially important benefits or harms) shows that an intervention has had 'no effect'. 'Statistical significance' should *not* be confused with the size or importance of an effect.

When results are not 'statistically significant' it cannot be assumed that there was no impact. Typically a cut-off of 5% is used to indicate statistical significance. This means that the results are considered to be 'statistically non-significant' if the analysis shows that differences as large as (or larger than) the observed difference would be expected to occur by chance *more* than one out of twenty times ($p \geq 0.05$). There are, however, two problems with this assumption. Firstly, the cut-off point of 5% is arbitrary. Secondly, 'statistically non-significant' results (sometimes mislabelled as 'negative'), might or might not be inconclusive. The figure below illustrates how the use of the terms 'statistically non-significant' or 'negative' can be misleading. Similarly, 'statistically significant' results might or might not be important.

Trends that are 'positive' (i.e. in favour of an option), but 'statistically non-significant', are often described as 'promising' and this can also be misleading. 'Negative' trends of the same magnitude, in contrast, are not typically described as 'warning signs'.
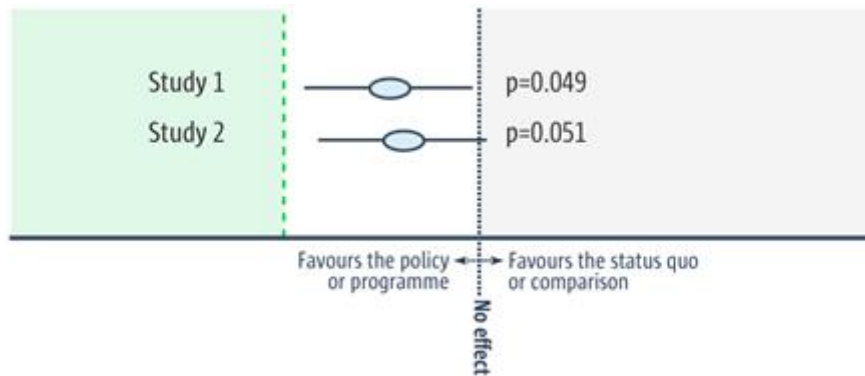
References

[1] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Devereaux PJ, et al. GRADE guidelines 6. Rating the quality of evidence – imprecision. J Clin Epidemiol 2011; 64:1283-93.
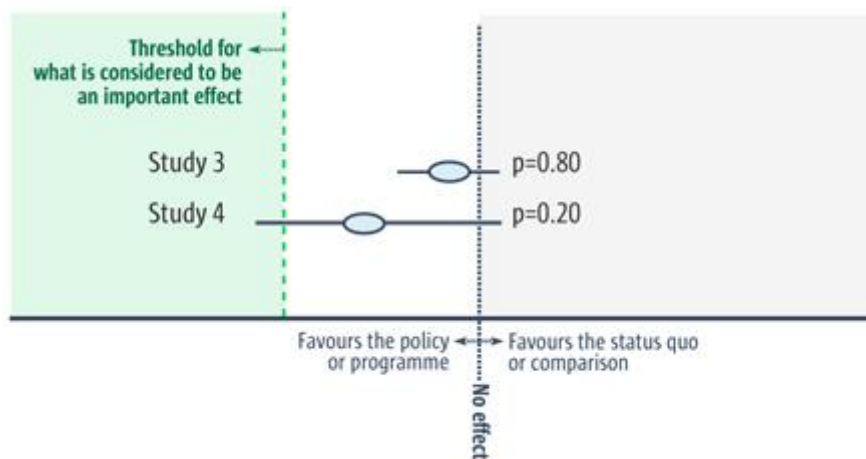[2] Alderson P, Chalmers I: Survey of claims of no effect in abstracts of Cochrane reviews. BMJ 2003, 326:475.

Figure. Two problems with classifying results as 'statistically non-significant' or 'negative'

1. 'Statistical signficance' is based on an arbitrary cut-off

Study 1   p=0.049
Study 2   p=0.051

Favours the policy ◄─► Favours the status quo
or programme   or comparison

No effect

2. 'Statistically non-significant' results may or may not be inconclusive

Threshold for ◄───
what is considered to be
an important effect

Study 3   p=0.80
Study 4   p=0.20

Favours the policy ◄─► Favours the status quo
or programme   or comparison

No effect

The blue dots in this figure indicate the estimated effect for each study and the horizontal lines indicate the 95% confidence intervals. A 95% confidence interval means that we can be 95 % confident that the true size of the effect is between the lower and upper confidence limit (the two ends of the horizontal lines). Conversely, there is a 5% chance that the true effect is outside this range.

The figure illustrates two problems that arise when results are classified as 'statistically non-significant' or 'negative':

1. *The classification is based on an arbitrary cut-off*. The results of Study 1, for example, are marginally different from the results of Study 2. But by using the conventional cut-off of P < 0.05, the results of Study 1 are considered 'statistically significant' and the results of Study 2 'statistically non-significant'.

In addition, in the example shown in the illustration the confidence intervals for both Study 1 and Study 2 rule out an important effect (i.e. an effect that is as large as or larger than the indicated threshold for what is considered to be an important effect). Although the results for Study 1 would be interpreted as 'statistically significant', the size of the effect was not important.

2. *'Statistically non-significant' results may or may not be inconclusive*. If the short green vertical line in the lower part of the figure indicates the smallest effect considered important, the results for Study 3 would be conclusive, since an important impact is highly unlikely. The results for Study 4 would be 'inconclusive' since it is not unlikely that there would be an important impact (the 95% confidence interval crosses the threshold for what is considered to be an important effect). Both results, however, might be regarded as 'statistically non-significant' or 'negative'