EVIDENCE-BASED CHILD HEALTH: A COCHRANE REVIEW JOURNAL
Evid.-Based Child Health **3**: 904–906 (2008)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/ebch.260

Tips and Tricks

Tips and tricks for understanding and using SR results – no 11: *P*-values and Confidence Intervals

L. C. M. Kremer[1,2]* and E. C. van Dalen[1]

[1] *Department of Paediatric Oncology, Emma Children's Hospital/Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands*
[2] *Department of Paediatrics, Emma Children's Hospital/Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands*

This eleventh article for 'Tips and tricks for understanding and using SR results' in *Evidence Based Child Health* is, like the previous ones, aimed at helping readers to understand the results of systematic reviews and to use the results in clinical practice. This time, we focus on P-values and Confidence Intervals. The information in this article is based on earlier papers, the *Cochrane Handbook*, and the collective experience of the authors in teaching evidence-based medicine (1–4).

Understanding SR Results

Limited value of a p-value

The p-value estimates the probability that the observed effect would have occurred by chance if the null hypothesis – that there is no difference between the effects of a treatment and a control intervention – was true. The results of a study lead us either to reject (if the p-value >0.05) or to accept (if the p-value <0.05) a null hypothesis. So, the p-value represents the probability that an observed or greater difference occurred by chance. The 0.05 threshold is an arbitrary one that is commonly used in medical research. A p-value that is very small indicates that the observed effect is very unlikely to have arisen purely by chance, and therefore provides evidence against the null hypothesis. P-values <0.05 are often reported as statistically significant.

The simple statements '$p < 0.05$', '$p > 0.05$' or 'p ns (not significant)', which are commonly used, have led to the mistaken belief that studies should aim at obtaining 'statistical significance'. However, small differences of no real clinical interest can be statistically significant if the number of study participants is large, whereas clinically important effects may be statistically non-significant for no other reason than that the number of participants was small. The p-value fails to provide clinical relevant information about the range of values within which the true effect lies.

Confidence Intervals

The results of individual studies and meta-analyses in systematic reviews should be presented as a point estimate together with a Confidence Interval. A 95% Confidence Interval is often interpreted as indicating a range within which we can be 95% certain that the true effect lies. However, the strictly correct interpretation of a Confidence Interval is based on the hypothetical notion that if a study was repeated infinitely often, and on each occasion a 95% Confidence Interval was calculated, then 95% of these intervals would contain the true effect.

The most commonly used level of confidence is 95%, but a Confidence Interval may be reported for any level of confidence (like 90% or 99%). The higher the confidence level, the wider the Confidence Interval. However, the width of the Confidence Interval also depends on the sample size of the study. Reducing the sample size leads to less precision and an increase in the width of the Confidence Interval. Furthermore, the width of the Confidence Interval depends on the Standard Deviation for continuous variables, the risk of events for dichotomous outcomes, and the number of events observed for time-to-event outcomes.

For a meta-analysis, the Confidence Interval depends on the precision of the individual study estimates and on the number of studies included in the meta-analysis. The width of the Confidence Interval usually decreases if the individual study estimates are more precise and if more studies are included in the meta-analysis. For random effects models, the width of the Confidence Interval also depends on the degree of heterogeneity within the studies. A high level of heterogeneity possibly increases the width of the Confidence Interval.

Using SR Results

P-value and Confidence Interval

In Figure 1 we present six hypothetical trials with equal interventions, but with different sample sizes. The outcome of interest is the absolute risk reduction

*Correspondence to: L. C. M. Kremer, Department of Paediatric Oncology, Emma Children's Hospital/Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands. E-mail: l.c.kremer@amc.uva.nl in infection rate. How can we interpret the results of these individual studies? The horizontal line at 0% represents a risk reduction of 0: infection rate in the intervention group is exactly the same as in the control group (so no statistically significant difference). From a statistically significant point of view, Confidence Intervals lying above this line favour the intervention treatment, whereas Confidence Intervals lying below this line favour the control treatment. The horizontal line at 5% represents the threshold for a clinically relevant difference. For the intervention being investigated in these hypothetical trials, we consider a 5% absolute risk reduction the smallest benefit that would outweigh the negative effects of therapy.

Trial 1 represents the results of a study with 50 patients. In this trial the point estimate is 2.5% with a wide 95% Confidence Interval, from -10–10%. Would you recommend the intervention to a patient based on the results of trial 1? Most likely you would not. If the 95% Confidence Interval includes 0 there is no statistically significant difference between the intervention and control group. However, the Confidence Interval includes the 5% we used as a threshold for a clinical relevant difference. Perhaps a trial with more patients would show a statistically significant difference. The inference of this study is no evidence of a statistically significant effect or clinically relevant effect.

Trial 2 represents the results of a trial which included 500 patients. The 95% Confidence Interval does not include the 0%. As a result there is a statistically significant reduction in infection rate compared to the control group. Would you recommend the intervention to a patient? Here, the concept of a minimal clinically relevant treatment effect proves to be useful. You have to consider the smallest amount of benefit that would justify therapy. The upper limit of the Confidence Interval of this trial lies below the threshold for a clinically relevant effect. Therefore, the intervention should not be recommended to a patient. The inference of this study is evidence of a statistically significant effect and also evidence of no clinically relevant effect.

In trial 3 the point estimate is above the threshold for clinical relevance; however, the Confidence Interval includes 5% (the threshold for a clinically relevant difference). So the real absolute reduction could be below 5%. What advice would you give to a patient? The sample size in this trial was inadequate to provide a definitive conclusion. The inference of this study is evidence of a statistically significant effect but no evidence of a clinically relevant effect.

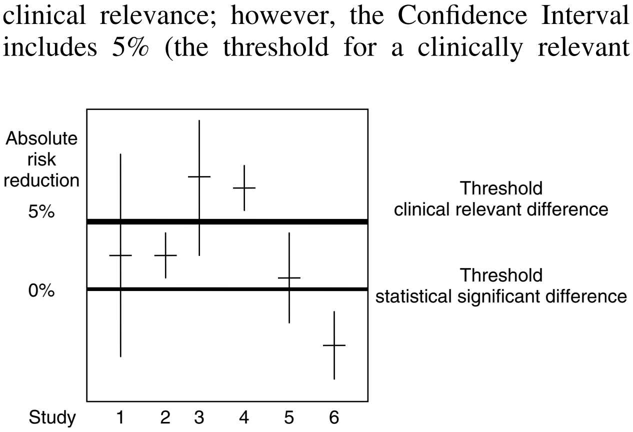

Figure 1. Examples of the effect (point estimate and 95% Confidence Intervals) of six hypothetical trials

Trial 4 included a higher number of patients than trial 3, and, as a result, the width of the Confidence Interval is smaller. The lower boundary of the Confidence Interval does not include 5% (the threshold for a clinically relevant difference). We would advise to use the intervention, assuming the patient's values and circumstances are congruent with its use. The inference of this study is evidence of both a statistically significant effect and a clinically relevant effect.

In trial 5 the point estimate is higher than 0%. However, the Confidence Interval is wide and no statistically significant results have been identified. The upper boundary of the Confidence Interval does not include 5%. We can be very confident that if there is a positive effect, it is trivial and not clinically relevant. We will therefore decide not to use the intervention treatment. The inference of this study is no evidence of statistically significant effect along with evidence of no clinically relevant effect.

In trial 6 the point estimate is lower than 0%, and the upper boundary of the Confidence Interval does not cross the 0% line. A statistically significant higher risk of infections is observed in the treatment arm over the control arm, so the intervention treatment should not be recommended to patients. The inference of this study is evidence of statistically significant harm of the intervention treatment.

Bottom Line

- The *p*-value fails to provide clinically relevant information about the range of values within which the true effect lies.
- The results of individual studies and meta-analyses in systematic reviews should be presented as a point estimate together with a Confidence Interval.
- For individual studies, the width of the Confidence Interval depends on the sample size of the study and on the Standard Deviation for continuous variables, the risk of events for dichotomous outcomes, and the number of events observed for time-to-event outcomes.
- For a meta-analysis, the width of the Confidence Interval depends on the precision of the individual study estimates, the number of studies included in the meta-analysis and on the degree of heterogeneity within the studies.
- For interpretation of results of individual studies and systematic reviews a threshold for a clinically relevant difference should be defined. If the upper boundary of the Confidence Interval is below the threshold, the inference of the study is 'evidence of

DOI: 10.1002/ebch.260

no clinically relevant effect'; if the lower boundary of the Confidence Interval is above the threshold, the inference of the study is 'evidence of a clinically relevant effect'; if the Confidence Interval includes the threshold the inference of the study is 'no evidence of a clinically relevant effect'.

References

1. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed).* 1986; **292**: 746–750.
2. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians: 4. Correlation and regression. *CMAJ.* 1995; **152**: 497–504.
3. Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Moyer V, Guyatt G; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *CMAJ.* 2004; **171**: 611–615.
4. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions Version* 5.0.0 [updated February 2008]. *The Cochrane Collaboration,* 2008; Available from www.cochrane-handbook.org.

DOI: 10.1002/ebch.260