Supported by sanofi-aventis

What are confidence intervals and p-values?

Huw TO Davies PhD Professor of Health Care Policy and Management, University of St Andrews

Iain K Crombie PhD FFPHM Professor of Public Health, University of Dundee

- A confidence interval calculated for **a measure of treatment effect** shows the range within which the true treatment effect is likely to lie (subject to a number of assumptions).
- A p-value is calculated to assess whether trial results are likely to have occurred simply through chance (assuming that there is no real difference between new treatment and old, and assuming, of course, that the study was well conducted).
- Confidence intervals are preferable to p-values, as they tell us the **range of possible effect sizes** compatible with the data.
- p-values simply provide a cut-off beyond which we assert that the findings are 'statistically significant' (by convention, this is $p<0.05$).
- A confidence interval that **embraces the value of no difference between treatments** indicates that the treatment under investigation is not significantly different from the control.
- Confidence intervals **aid interpretation of clinical trial data** by putting upper and lower bounds on the likely size of any true effect.
- **Bias must be assessed** before confidence intervals can be interpreted. Even very large samples and very narrow confidence intervals can mislead if they come from biased studies.
- **Non-significance does not mean 'no effect'.** Small studies will often report non-significance even when there are important, real effects which a large study would have detected.
- Statistical significance does not necessarily mean that the effect is real: by chance alone about **one in 20 significant findings will be spurious.**
- Statistically significant does not necessarily mean clinically important. It is the **size of the effect** that determines the importance, not the presence of statistical significance.

For further titles in the series, visit:
www.whatisseries.co.uk

What are confidence intervals and p-values?

Measuring effect size

Clinical trials aim to generate new knowledge on the effectiveness (or otherwise) of healthcare interventions. Like all clinical research, this involves estimating a key parameter of interest, in this case the effect size. The effect size can be measured in a variety of ways, such as the relative risk reduction, the absolute risk reduction or the number needed to treat (NNT; Table 1). **Relative measures** tend to emphasise potential benefits, whereas **absolute measures** provide an across-the-board summary.[1] Either may be appropriate, subject to correct interpretation.

Whatever the measure used, some assessment must be made of the trustworthiness or **robustness** of the findings. The findings of the study provide a point estimate of effect, and this raises a dilemma: are the findings from this sample also likely to be true about other similar groups of patients? Before we can answer such a question, two issues need to be addressed. Does any apparent treatment benefit arise because of the way the study has been

Box 1. Hypothesis testing and the generation of p-values

The logic of hypothesis testing and p-values is convoluted. Suppose a new treatment appears to outperform the standard therapy in a research study. We are interested in assessing whether this apparent effect is likely to be real or could just be a chance finding: p-values help us to do this.

In calculating the p-value, we first assume that there really is no true difference between the two treatments (this is called the **null hypothesis**). We then calculate how likely we are to see the difference that we have observed just by chance if our supposition is true (that is, if there is really no true difference). This is the p-value.

So the p-value is the probability that we would observe effects as big as those seen in the study if there was really no difference between the treatments. If p is small, the findings are unlikely to have arisen by chance and we reject the idea that there is no difference between the two treatments (we reject the null hypothesis). If p is large, the observed difference is plausibly a chance finding and we do not reject the idea that there is no difference between the treatments. Note that we do not reject the idea, but we do not accept it either: we are simply unable to say one way or another until other factors have been considered.

But what do we mean by a 'small' p-value (one small enough to cause us to reject the idea that there was really no difference)? By convention, p-values of less than 0.05 are considered 'small'. That is, if p is less than 0.05 there is a less than one in 20 chance that a difference as big as that seen in the study could have arisen by chance if there was really no true difference. With p-values this small (or smaller) we say that the results from the trial are statistically significant (unlikely to have arisen by chance). Smaller p-values (say $p<0.01$) are sometimes called 'highly significant' because they indicate that the observed difference would happen less than once in a hundred times if there was really no true difference.

Table 1. Summary of effect measures

Measure of effect	Abbreviation	Description	No effect	Total success
Absolute risk reduction	ARR	Absolute change in risk: the risk of an event in the control group minus the risk of an event in the treated group; usually expressed as a percentage	ARR=0%	ARR=initial risk
Relative risk reduction	RRR	Proportion of the risk removed by treatment: the absolute risk reduction divided by the initial risk in the control group; usually expressed as a percentage	RRR=0%	RRR=100%
Relative risk	RR	The risk of an event in the treated group divided by the risk of an event in the control group; usually expressed as a decimal proportion, sometimes as a percentage	RR=1 or RR=100%	RR=0
Odds ratio	OR	Odds of an event in the treated group divided by the odds of an event in the control group; usually expressed as a decimal proportion	OR=1	OR=0
Number needed to treat	NNT	Number of patients who need to be treated to prevent one event; this is the reciprocal of the absolute risk reduction (when expressed as a decimal fraction); it is usually rounded to a whole number	NNT=∞	NNT=1/initial risk

conducted **(bias)**, or could it arise simply because of **chance?** The short note below briefly covers the importance of assessing bias but focuses more on assessing the role of chance.

Bias

Bias is a term that covers any **systematic errors** that result from the way the study was designed, executed or interpreted. Common flaws in treatment trials are:

- Lack of (or failure in) randomisation, leading to unbalanced groups
- Poor blinding, leading to unfair treatment and biased assessments
- Large numbers of patients lost to follow-up.

Assessment in these areas is crucial before the results from any trial can be assessed, and many useful guides exist to assist this process, such as an article by Guyatt *et al* and books by Sackett *et al* and by Crombie.[2–5] Interpretation of the effects of chance is only meaningful once bias has been excluded as an explanation for any observed differences.[6,7]

Chance variability

The results from any particular study will vary just by chance. Studies differ in terms of the people who are included, and the ways in which these specific individuals react to therapeutic interventions. Even when everything possible is held constant, there will still be some random variations. Hence we need some tools to help us to assess whether differences that we see between new treatment and old in any particular study are real and important, or just manifestations of chance variability. Confidence intervals and p-values help us to do this.

What are p-values?

Until comparatively recently, assessments of the role of chance were routinely made using **hypothesis testing,** which produces a 'p-value' (Box 1). The p-value allows assessment of whether or not the findings are 'significantly different' or 'not significantly different' from some reference value (in trials, this is usually the value reflecting 'no effect'; Table 1). A different and potentially more useful approach to assessing the role of chance has come to the fore: confidence intervals.[8] Although these might appear rather dissimilar to p-values, the theory and calculations underlying these two approaches are largely the same.

What are confidence intervals?

Confidence intervals provide different information from that arising from hypothesis tests. Hypothesis testing produces a decision about any observed difference: either that the difference is 'statistically significant' or that it is 'statistically non-significant'. In contrast, confidence intervals provide a **range** about the observed effect size. This range is constructed in such a way that we know how likely it is to capture the true – but unknown – effect size.

Thus, the formal definition of a confidence interval is: 'a range of values for a variable of interest [in our case, the measure of treatment effect] constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits'.[9]

It is conventional to create confidence intervals at the 95% level – so this means that 95% of the time properly constructed confidence intervals should contain the true value of the variable of interest. This corresponds to hypothesis testing with p-values, with a conventional cut-off for p of less than 0.05.

More colloquially, the confidence interval provides a range for our best guess of the size of the true treatment effect that is plausible given the size of the difference actually observed.

Assessing significance from a confidence interval

One useful feature of confidence intervals is that one can easily tell whether or not statistical significance has been reached, just as in a hypothesis test.

- If the confidence interval **captures** the value reflecting 'no effect', this represents a difference that is statistically non-significant (for a 95% confidence interval, this is non-significance at the 5% level).
- If the confidence interval **does not enclose** the value reflecting 'no effect', this represents a difference that is statistically significant (again, for a 95% confidence interval, this is significance at the 5% level).

Thus, 'statistical significance' (corresponding to $p<0.05$) can be inferred from confidence intervals – but, in addition, these intervals show the largest and smallest effects that are likely, given the observed data. This is useful extra information.

An example of the use of confidence intervals is shown in Box 2.[10]

Examining the width of a confidence interval

One of the advantages of confidence intervals over traditional hypothesis testing is the additional information that they convey. The upper and lower bounds of the interval give us information on how big or small the true effect might plausibly be, and the width of the confidence interval also conveys some useful information.

If the confidence interval is narrow, capturing only a small range of effect sizes, we can be quite confident that any effects far from this range have been ruled out by the study. This situation usually arises when the size of the study is quite large and, hence, the estimate of the true effect is quite precise. Another way of saying this is to note that the study has reasonable 'power' to detect an effect.

However, if the confidence interval is quite wide, capturing a diverse range of effect sizes, we can infer that the study was probably quite small. Thus, any estimates of effect size will be quite imprecise. Such a study is 'low-powered' and provides us with less information.

Errors in interpretation

Confidence intervals, like p-values, provide us with a guide to help with the interpretation of research findings in the light of the effects of chance. There are, however, three important pitfalls in interpretation.

Getting it wrong: seeing effects that are not real

First of all, we may examine the confidence interval and/or the p-value and observe that the difference is 'statistically significant'. From this we will usually conclude that there is a difference between the two treatments. However, just because we are unlikely to observe such a large difference simply by chance, this does not mean that it will not happen. By definition, about one in 20

Box 2. An example of the use of confidence intervals[10]

Ramipril is an angiotensin-converting enzyme (ACE) inhibitor which has been tested for use in patients at high risk of cardiovascular events. In one study published in the *New England Journal of Medicine*,[10] a total of 9,297 patients were recruited into a randomised, double-blind, controlled trial. The key findings presented on the primary outcome and deaths are shown below.

Incidence of primary outcome and deaths from any cause

Outcome	**Ramipril group**	**Placebo group**	**Relative risk**
	(n=4,645)	(n=4,652)	(95% CI)
	number (%)	number (%)	
Cardiovascular event (including death)	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)
Death from non-cardiovascular cause	200 (4.3)	192 (4.1)	1.03 (0.85–1.26)
Death from any cause	482 (10.4)	569 (12.2)	0.84 (0.75–0.95)

These data indicate that fewer people treated with ramipril suffered a cardiovascular event (14.0%) compared with those in the placebo group (17.8%). This gives a relative risk of 0.78, or a reduction in (relative) risk of 22%. The 95% confidence interval for this estimate of the relative risk runs from 0.70 to 0.86. Two observations can then be made from this confidence interval.

- First, the observed difference is statistically significant at the 5% level, because the interval does not embrace a relative risk of one.
- Second, the observed data are consistent with as much as a 30% reduction in relative risk or as little as a 14% reduction in risk.

Similarly, the last row of the table shows that statistically significant reductions in the overall death rate were recorded: a relative risk of 0.84 with a confidence interval running from 0.75 to 0.95. Thus, the true reduction in deaths may be as much as a quarter or it could be only as little as 5%; however, we are 95% certain that the overall death rate is reduced in the ramipril group.

Finally, exploring the data presented in the middle row shows an example of how a confidence interval can demonstrate non-significance. There were a few more deaths from non-cardiovascular causes in the ramipril group (200) compared with the placebo group (192). Because of this, the relative risk is calculated to be 1.03 – showing a slight increase in risk in the ramipril group. However, the confidence interval is seen to capture the value of no effect (relative risk = 1), running as it does from 0.85 to 1.26. The observed difference is thus non-significant; the true value could be anything from a 15% reduction in non-cardiovascular deaths for ramipril to a 26% increase in these deaths. Not only do we know that the result is not significant, but we can also see how large or small a true difference might plausibly be, given these data.

significant findings will be spurious – arising simply from chance. Thus, we may be misled by chance into believing in something that is not real – technically, this is called a **'type I error'.**

It is a frustrating but unavoidable feature of statistical significance (whether assessed using confidence intervals or p-values) that around one in 20 will mislead. Yet we cannot know which of any given set of comparisons is doing the misleading. This observation cautions against generating too many statistical comparisons: the more comparisons made in any given study, the greater the chance that at least some of them will be spurious findings. Thus, clinical trials which show significance in only one or two subgroups are unconvincing – such significance may be deceptive. Unless particular subgroup analyses have been specified in advance, differences other than for the primary endpoint for the whole group should be viewed with suspicion.

Statistical significance and clinical significance

Statistical significance is also sometimes misinterpreted as signifying an important result: this is a second important pitfall in interpretation. Significance testing simply asks whether the data produced in a study are compatible with the notion of no difference between the new and control interventions. Rejecting equivalence of the two interventions does not necessarily mean that we accept that there is an important difference between them. A large study may identify as statistically significant a fairly small difference. It is then quite a separate judgement to assess the clinical significance of this difference. In assessing the importance of significant results, it is the size of the effect – not just the size of the significance – that matters.

Getting it wrong again: failing to find real effects

A further error that we may make is to conclude from a non-significant finding that there is no effect, when in fact there is a real effect – this is called a **'type II error'.** Equating non-significance with 'no effect' is a common misconception. A non-significant confidence interval simply tells us that the observed difference is consistent with there being no true difference between the two groups. Thus, we are unable to reject this possibility. This is where confidence intervals are much more helpful than simple p-values: the observed difference will also be compatible with a range of other effect sizes as described by the confidence interval.[8] We are unable to reject these possibilities and must then assess whether some of them (usually the upper and lower limits of the confidence interval) might be important. Just because we have not found a significant treatment effect, it does not mean that there is no treatment effect to be found.[11] The crucial question is: how carefully have we interpreted the findings?

Extrapolating beyond the trial

For all the complexity of understanding bias and chance in the interpretation of the findings from clinical trials, another important consideration should not be forgotten. The findings from any given study relate to the patients included in that study. Even if an effect is assessed as probably real and large enough to be clinically important, a further question remains: how well are the findings applicable to other groups of patients, and do they particularise to a given individual?[12] Neither confidence intervals nor p-values are much help with this judgement. Assessment of this **external validity** is made based on the patients' characteristics and on the setting and the conduct of the trial.

Summary

Confidence intervals and p-values take as their starting point the results observed in a study. Crucially, we must check first that this is an unbiased study. The question that confidence intervals then answer is: what is the range of real effects that is compatible with these data? The confidence interval is just such a range, which 95% of the time will contain the true value of the main measure of effect (relative risk reduction, absolute risk reduction, NNT or whatever; Table 1).

This allows us to do two things. First, if the confidence interval embraces the value of no effect (for example, no difference between two treatments as shown by a relative risk equal to one or an absolute difference equal to zero), then the findings are non-significant. If the confidence interval does not embrace the value of no difference, then the findings are statistically significant. Thus, confidence intervals provide the same information as a p-value. But more than this: the upper and lower extremities of the confidence interval also tell us how large or small the real effect might be and yet still give us the observed findings by chance. This additional information is very helpful in allowing us to interpret both borderline significance and non-significance. Confidence intervals from large studies tend to be quite narrow in width, showing the precision with which the study is able to estimate the size of any real effect. In contrast, confidence intervals from smaller studies are usually wide, showing that the findings are compatible with a wide range of effect sizes.

References

1. Davies HT. Interpreting measures of treatment effect. *Hosp Med* 1998; **59:** 499–501.
2. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993; **270:** 2598–2601.
3. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: A basic science for clinical medicine,* 2nd edn. Boston, Massachusetts: Little, Brown and Company, 1991.
4. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence based medicine: how to practice and teach EBM.* London: Churchill Livingstone, 1997.
5. Crombie IK. *The pocket guide to critical appraisal.* London: BMJ Publishing, 1996.
6. Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ* 1994; **309:** 727–730.
7. Burls A. *What is critical appraisal?* London: Hayward Medical Communications, 2009.
8. Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. *BMJ* 1986; **292:** 746–750.
9. Last JM. *A dictionary of epidemiology.* Oxford: International Journal of Epidemiology, 1988.
10. The Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *N Engl J Med* 2000; **342:** 145–153.
11. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995; **311:** 485.
12. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* 1994; **271:** 59–63.

What are

confidence intervals and p-values?

First edition published 2003
Author: Huw TO Davies

This publication, along with the others in the series, is available on the internet at www.whatisseries.co.uk
The data, opinions and statements appearing in the article(s) herein are those of the contributor(s) concerned. Accordingly, the sponsor and publisher, and their respective employees, officers and agents, accept no liability for the consequences of any such inaccurate or misleading data, opinion or statement.

Published by Hayward Medical Communications, a division of Hayward Group Ltd.

Supported by *sanofi-aventis*