World J. Surg. 29, 554–556 (2005)
DOI: 10.1007/s00268-005-7911-0

WORLD
Journal of
SURGERY

Why We Need Confidence Intervals

Douglas G. Altman, DSc

NHS/Cancer Research UK Centre for Statistics in Medicine, Old Road Campus, Headington, Oxford OX3 7LF, UK

Published Online: April 14, 2005

Abstract. The estimation approach to statistical analysis aims to quantify the effect of interest as an "estimate" of a clinically relevant quantity and to quantify the uncertainty in this estimate by means of a confidence interval (CI). As such, results expressed in this form are much more informative than results presented just as *p* values. This article focuses on the principles rather than the mathematics of CIs and discusses interpretation of CIs and some common misuses. CIs can be constructed for almost all analyses. They are especially useful for avoiding misinterpretation of nonsignificant results of small studies. CIs should be provided routinely for the main results of trials and observational studies.

A dramatic change in published reports of medical research over the last 20 years has been the wide adoption of confidence intervals (CIs) as a standard part of the presentation of the quantitative results of studies. The steady increase in the use of CIs, which has been encouraged by the fact that some leading journals require them, has in general not been instead of *p* values but as a supplement to them. Despite these encouraging changes, it may be useful to refresh memories about some key concepts: What do confidence intervals tell us, and why do we need them?

Principles of Statistical Inference

Two different, but complementary, approaches to statistical analysis are testing and estimation. Hypothesis testing has long been the mainstay of statistical analysis in medical research. However, it is well recognized that providing just a single, somewhat impenetrable "*p* value" represents a serious overreduction of the data. A *p* value for a comparison of two groups is the probability of observing a result at least as extreme as the observed result if in truth there is no difference between the groups (i.e., under the "null hypothesis" of no effect). For example, in a randomized controlled trial (RCT) the null hypothesis is that the proportions with the outcome of interest are the same in both treatment groups, so the risk difference is zero and the risk ratio is 1. A very small *p* value suggests that the null hypothesis is highly unlikely to be true, and thus there is evidence of an effect. By convention $p < 0.05$ is usually termed "statistically significant" and is widely considered adequate evidence of effect. In reality, a result with $p < 0.05$ provides marginal evidence of effectiveness, and rather smaller *p* values are needed for a convincing result.

By contrast, the estimation approach to statistical analysis aims to quantify the effect of interest as an "estimate" of a clinically relevant quantity and to quantify the uncertainty in this estimate by means of a CI. For example, we can obtain CIs for means or proportions in a group of individuals or for the difference between two such estimates. The CI is a range of values either side of the estimate between which we can be 95% sure that the true value lies. A series of identical studies carried out on different samples of patients from the same population would yield varying results spread around the true, but unknown, effect. The CI obtained from the results of a single study provides a range of uncertainty due to this "sampling variation." The main purpose of confidence intervals is thus to indicate the (im)precision of the sample study result as an estimate of the population value.

As the name implies, the range specified by the CI indicates how confident we can be in the observed results. A narrow CI indicates little imprecision (uncertainty) and hence a high degree of confidence. Such confidence in general comes only from large studies. The convention of using 95% "coverage" for CIs is arbitrary, as is that of taking $p < 0.05$ as being significant, and authors sometimes use 90% or 99% CIs. Note that the word "interval" means a range of values and is thus singular. The two values that define the interval are known as confidence limits.

Confidence intervals can be calculated for most statistical estimates, including summaries of single samples and the difference between two samples, as well as for regression coefficients. I focus here on RCTs comparing two health care interventions. In most circumstances the CI is calculated from the observed estimate of the treatment effect, such as the difference (d) between two proportions, and the standard error (SE) of that estimate. A 95% CI is obtained here as $d \pm 1.96SE$. (The formula varies according to the nature of the outcome measure and the coverage of the CI, but it is of this general type.) Full details of methods for calculating CIs for various types of data and various study designs are given elsewhere [1, 2].

Correspondence to: Douglas G. Altman, DSc, e-mail: doug.altman@cancer.org.uk

Example

A trial comparing open mesh and laparoscopic mesh repair of inguinal hernias found that the 2-year recurrence rate was 87/862 (10.1%) patients in the open group and 41/834 (4.9%) in the laparoscopy group [3]. The relative risk is 0.49 with the 95% CI from 0.34 to 0.70. We can interpret this finding as saying that our best estimate is that the risk of recurrence is about halved in the laparoscopy group (relative risk reduction 51%) but that the results are compatible with a reduction in risk of recurrence between 30% and 66%. (The authors cited the odds ratio, which is similar to the risk ratio but less easy to interpret.) Thus even in this large trial there is a lot of uncertainty about the magnitude of the benefit of the laparoscopic approach.

Patients in the open repair group also experienced greater levels of pain. At 2 weeks after surgery the difference in mean pain scores (by visual analogue scale) was 6.1 (95% CI 1.7–10.5). Again, although there is a significant difference, there is considerable uncertainty about the size of the effect.

Two Common Errors

In a comparative study such as an RCT, a common, serious mistake is to conclude from a nonsignificant result (i.e., with $p > 0.05$) that the groups are "the same." Yet this serious error is extremely common. CIs are especially useful here, as they show whether the data are compatible with clinically useful true effects.

Leung et al. [4] carried out a randomized trial comparing laparoscopy assisted resection versus open resection for patients with rectosigmoid carcinoma. They sought an increase in 5-year survival probability from 60% to 75%. The results from a trial of 337 patients showed 5-year survival probabilities of 76.1% in the laparoscopy group and 72.9% in the open resection group. The authors reported $p = 0.61$ for the comparison and concluded that laparoscopic surgery "does not jeopardize survival." They did not present a confidence interval for the difference in survival, yet this is quite informative. Using the general formula from above and the standard errors they provided of 3.7% and 4.0%, respectively, the 95% CI for the difference in 5-year survival was −7.5% to 13.9%. In other words the study result is compatible with a range of results between laparoscopic resection leading to a 7.5% worse survival or 13.9% better survival than open resection. The difference sought was 15%, only just outside the confidence interval. Thus even with more than 350 randomized patients there is still quite a lot of uncertainty about the relative survival associated with the two procedures.

Koivunen et al. [5] concluded from their trial that "adenoidectomy...is not effective...it cannot be recommended"; yet the 95% confidence interval (CI) for the primary outcome (further episodes of otitis media) was compatible with an 18% absolute risk reduction at 24 months. The clinically important difference they sought was a 25% reduction. The study is compatible with a smaller benefit of, say, 15%, which others may judge would be clinically useful. Even when a clinically useful effect has been ruled out, phrases such as "is not effective," "did not reduce," and "has no effect" are not justified [6].

Small trials are likely to have nonsignificant results, and thus there is great scope for drawing a misleading conclusion based only on a p value. For example, Widman et al. [7] concluded that drainage reduced the hematoma volume after total hip arthroplasty, but their trial included only 22 patients and they did not present confidence intervals. Only if the CI excludes clinically useful benefit would it be reasonable to conclude that a study has demonstrated no benefit, and even then aspects of how that particular trial was conducted may prevent safe generalization. This error is not confined to reports of RCTs but, rather, is a general misinterpretation of not significant as "not present." Absence of evidence is not evidence of absence [8, 9].

Another common misuse of CIs in a comparative study is the presentation and comparison of separate CIs for each group rather than consideration of a CI for the contrast. This practice leads to inferences based on whether the two CIs, such as for the means in each group, overlap; or whether one group has a CI including the value for no effect whereas the other does not. This is not the appropriate comparison and may mislead. In such cases the correct approach is to construct a CI to compare the two groups, such as for the ratio of two relative risks or the difference between the change from baseline in each group [10].

Discussion

The CI gives a measure of the precision (or uncertainty) of study results for making inferences about the population of all such patients. A strictly correct definition of a 95% CI is that 95% of such intervals contain the true population value. Little is lost by the common but less pure interpretation of the CI as the range of values within which we can be 95% sure the population value lies. The uncertainty (imprecision) expressed by a CI is to a large extent affected by the square root of the sample size. Small samples provide less information than large ones, and CI is correspondingly wider in a smaller sample.

Presentation of a CI places a clear emphasis on quantification, in direct contrast to p values. The p value is not an estimate of any quantity but, rather, a measure of the strength of evidence against the null hypothesis of "no effect." The p value by itself tells us nothing about the size of a difference nor even the direction of that difference. Thus p values on their own are not informative in articles or abstracts. By contrasts, CIs indicate the strength of evidence about quantities of direct interest, such as treatment benefit. They should be given in the main text and in the abstract of published articles reporting RCTs [11] and other studies.

Despite the considerably different philosophical approaches, CIs and significance tests are closely related. Thus a "significant" p value of $p < 0.05$ corresponds to a 95% CI that excludes the value indicating equality; this value is 0 for the difference between two means or proportions and 1 for a relative risk, odds ratio, or hazard ratio. The prevailing view is that estimation, including CIs, is the preferable approach to summarizing the results of a study, but CIs and p values are complementary and many articles use both.

Confidence intervals reflect only uncertainty arising from sampling variation, not additional uncertainty due to failure to follow the protocol, nonrandom loss to follow up, and so on. True uncertainty is greater, therefore, than indicated by the CI [6]. These considerations are especially relevant to nonrandomized studies. Such studies generally require adjustment for important baseline variables to try to make the groups more similar. For example, in a nonrandomized study evaluating valve surgery for adults with valve endocarditis, the unadjusted analysis showed a large reduction in risk of mortality for those undergoing surgery

(16% vs. 33%; hazard ratio = 0.43, 95% CI 0.29–0.63; $p < 0.001$) [12]. Adjusted analyses using two different strategies gave rather similar results, although with larger p values, but in one case with a much wider confidence interval (hazard ratio = 0.45, 95% CI 0.23–0.86). Such adjustment may not be fully convincing. A systematic review and meta-analysis of published studies including 16,000 patients comparing unilateral and bilateral mammary artery bypass grafting showed a significantly reduced risk of death in the bilateral group (hazard ratio = 0.81; 95% CI 0.70–0.94) [13]. Nonetheless, because of uncertainties in the result and the considerable public health importance of the question a 10-year randomized trial has recently been funded.

Confidence intervals can be constructed for most common statistical estimates or comparisons [1]. For randomized trials and other comparative studies, these include differences between means or proportions, relative risks, odds ratios, hazard ratios, and the number needed to treat (NNT). Likewise, CIs can be obtained for all the main estimates arising in studies of diagnosis—sensitivity, specificity, positive predictive value (all of which are simple proportions)—and estimates derived from meta-analyses and case-control studies. A computer program for personal computers that covers these and other methods is available [1].

Although CIs are desirable for the primary results of a study, they are not needed for all results. Furthermore, it is important that when given they relate to the contrast of interest. In particular, when two groups are compared, the appropriate CI is that for the difference between the groups, as illustrated in the above examples. Not only is it not helpful to give separate CIs for the estimates in each group, this presentation can be quite misleading. When the authors have not provided CIs, they can often be constructed using the results provided in their article.

Conclusions

The most appropriate methods of statistical analysis and presentation must be largely a matter of personal judgment, although increasingly journals are requesting or requiring authors to use CIs when presenting their key findings. The wide adoption of CIs in medical research papers has been of great benefit to a more correct understanding of the information provided by the results of medical research. CIs should be provided routinely for the main results of trials and observational studies.

References

1. Altman, DG, Machin, D, Bryant, TN, et al. (2000), *Statistics with Confidence*, 2nd ed., BMJ Books, London, (including CIA software)
2. Altman, DG (2000) "Confidence intervals" In: Sackett, DL, Straus, S, Richardson, WS (editors), *Evidence-based Medicine: How to Practice and Teach EBM*, 2nd ed., Churchill-Livingstone, Edinburgh, pp 233-243
3. Neumayer L, Giobbie-Hurder A, Jonasson O, et al. Open mesh versus laparoscopic mesh repair of inguinal hernia. N. Engl. J. Med 2004;350:1819–1827
4. Leung KL, Kwok SP, Lam SC, et al. Laparoscopic resection of rectosigmoid carcinoma: prospective randomized trial. Lancet 2004;363:1187–1192
5. Koivunen P, Uhari M, Luotonen J, et al. Adenoidectomy versus chemoprophylaxis and placebo for recurrent acute otitis media in children aged under 2 years: randomized controlled trial. B. M. J 2004;328:487–490
6. Altman DG, Bland JM. Confidence intervals illuminate absence of evidence. B. M. J 2004;328:1016–1017
7. Widman J, Jacobsson H, Larsson SA, et al. No effect of drains on the postoperative hematoma volume in hip replacement surgery: a randomized study using scintigraphy. Acta. Orthop. Scand 2002;73:625–629
8. Altman, DG, Bland, JM (1995) "Absence of evidence is not evidence of absence" B.M.J. 311: 485
9. Alderson P. Absence of evidence is not evidence of absence. B. M. J 2004;328:476–477
10. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. B. M. J 2003;326:219
11. Moher D, Schulz KF, Altman DG, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191–1194
12. Vikram HR, Buenconsejo J, Hasbun R, et al. Impact of valve surgery on 6-month mortality in adults with complicated, left-sided native valve endocarditis: a propensity analysis. J. A. M. A 2003;290:3207–3214
13. Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. Lancet 2001;358:870–875