Use of confidence intervals to indicate uncertainty in research findings

One of the most dramatic changes in published reports of medical research over the past 10 years is the wide adoption of confidence intervals (CIs) as part of the presentation of the quantitative results of studies. Although the basic methods have been around for some time, between 1986 and 1987, for unclear reasons, a cluster of publications in medical journals espoused the cause of improved reporting by using CIs (1-3). Since then, the use of CIs has steadily increased, partly encouraged by the fact that some leading journals require them! It seems that the understanding and usefulness of CIs have also increased. Despite these encouraging changes, this may be a good time (and certainly a good place) to refresh memories about the key concepts.

The estimation approach to statistical analysis aims to quantify the effect of interest and also to quantify the uncertainty in this estimate by means of a CI, which is usually in a range of values that are on either side of the estimate in which we can be 95% sure that the true value lies. The convention of using the value of 95% is arbitrary, just as is that of taking $P < 0.05$ as being significant. Some authors use 90% or 99% CIs. Note that the word “interval” means a range of values and is thus singular. The 2 values that define the interval are known as confidence limits.
The CI is based on the idea that if the same study were done with a different sample of patients, the results would not be identical but would be spread around the true but unknown result. The CI estimates this "sampling variation." In most circumstances, the CI is calculated from the observed estimate of the quantity of interest, for example, the difference (d) between 2 proportions, and the standard error (SE) of the estimate. A 95% CI is obtained here as d ± 1.96 SE. (The formula varies according to the nature of the outcome measure and the coverage of the CI, but it is of this general type.) For example, in a randomised, placebo-controlled trial of acellular pertussis vaccine (4), 71 of 1670 (4.3%) infants developed pertussis amongst those who received the vaccine, and 240 of 1665 (14.4%) did so among the control group. The difference in percentages, known as the absolute risk reduction, is 10.1%. The SE of this difference is 0.99%, so that the 95% CI is 10.1 ± 1.96 x 0.99 (i.e., from 8.2% to 12.0%).

What does a CI tell us? 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, somewhatopaquely, that 95% of these intervals will contain the true population value. Little is lost by the less than pure interpretation of the CI as the range of quantities of direct interest such as treatment benefit, which makes them of particular relevance to Evidence-Based Medicine.

Despite the considerably different philosophical approaches, CIs and significance tests are closely related mathematically. Thus, a "significant" P value of P < 0.05 will correspond to a 95% CI that excludes the value that indicates equality. For example, this value is 0 for the difference between 2 means or proportions and 1 for a relative risk or odds ratio. (The equivalence of the 2 approaches may not be exact in some circumstances.) The prevailing view is that estimation, including CIs, is the preferable approach to summarising the results of a study, but CIs and P values are complementary and many studies use both. Of the 48 abstracts published in the first 2 issues of Evidence-Based Medicine, 21 used both, 9 presented only CIs, and 8 presented only P values. Perhaps it is surprising that only about two thirds of the abstracts in Evidence-Based Medicine included CIs, but some of the original articles do not provide sufficient data to permit their calculation. It is also the current policy of the journal not to provide CIs for results that are not statistically significant.

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 do large ones, and the CI is correspondingly wider in a smaller sample. For example, a recent review in Evidence-Based Medicine of a study that compared the characteristics of 3 tests to diagnose H. pylori (5) included the sensitivity of the 14C urea breath test as 95.8% (95% CI 75% to 100%). Although the figure of 95.8% is impressive, the small sample of 24 adults with H. pylori means that considerable uncertainty exists in the estimate, as shown by the wide CI. If the same sensitivity had been observed in a sample of 240, the 95% CI would have been from 92.5% to 98.0%.

Non-significant results in randomised trials (i.e., those with P > 0.05) are especially prone to misinterpretation. CIs are especially useful here because they show whether the data are compatible with clinically useful true effects. For example, a recent abstract in Evidence-Based Medicine presented results of a randomised trial to compare suturing with stapling for large-bowel anastomosis (6). One of the outcomes was wound infection, which occurred in 10.9% and 13.5% of patients, respectively (P = 0.30). The Evidence-Based Medicine review did not give the 95% CI for the difference of 2.6%, which is between 2% and 8%. Thus, even in this study of 652 patients, there remains the possibility that a modest difference in wound infection rates existed for the 2 procedures. In smaller studies there is greater uncertainty. Sung and colleagues (7) did a randomised trial with 100 patients to compare ocreotide infusion with emergency sclerotherapy for acute variceal haemorrhage. The observed rates of controlled bleeding were 84% in the ocreotide group and 90% in the sclerotherapy group (P = 0.55). Note that the figures for uncontrolled bleeding are similar to those for wound infection in the study just considered. In this case, however, the 95% CI for the treatment difference of 6% is 2.7% to +19%. This interval is very wide in relation to the 5% difference that was of interest. It is clear that the study cannot rule out a large difference in effectiveness; therefore, the authors' conclusion that "ocreotide infusion and sclerotherapy are equally effective in controlling variceal haemorrhage" (7) is certainly not valid.

CIs can be constructed for most common statistical estimates or com-
comparisons (8). For randomised controlled trials, these include differences between means or proportions, relative risks, odds ratios, and the number needed to treat (NNT). A computer program for personal computers that covers many of these methods is available from BMJ (9). 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.

Although CIs are desirable for the primary results of a study, they are not needed for all results. Further, when they are given, it is important that they relate to the contrast of interest. In particular, when 2 groups are compared, the appropriate CI is for the difference between the groups, as illustrated in the above examples. Not only is it unhelpful to give separate CIs for the estimate in each group, the appropriate CI is needed for all the main estimates available risks, odds ratios, and the number needed to treat (NNT). A computer program for personal computers that covers many of these methods is available from BMJ (9). 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.

Although CIs are desirable for the primary results of a study, they are not needed for all results. Further, when they are given, it is important that they relate to the contrast of interest. In particular, when 2 groups are compared, the appropriate CI is for the difference between the groups, as illustrated in the above examples. Not only is it unhelpful to give separate CIs for the estimate in each group, but this presentation can also be quite misleading. When the authors do not provide CIs, these often can be constructed by using the results provided in their study. Some variations of the test performance of less costly methods of detection. Clin Invest Med. 1995;18:177-85.


