



Presenting results with confidence

Angie M Wade

The *Journal of Clinical Pathology* publishes about nine papers a month presenting comparative (mostly primary) data. Their aim is to inform the clinical readership about the effects of disease or the comparative effectiveness of differing treatments.

During the last 10 years many internationally recognised reporting guidelines within health research have been developed. The most notable of these are CONSORT (CONsolidated Standards Of Reporting Trials¹), STROBE (STrengthening the Reporting of Observational studies in Epidemiology²) and STARD (STAndards for the Reporting of Diagnostic accuracy studies³). However, these guidelines have been used infrequently. Last year the umbrella network EQUATOR (Enhancing the QUALity and Transparency Of health Research⁴) was officially launched with the aim of enhancing the reliability of medical research literature by promoting transparent and accurate reporting of research results. One way the network aimed to do this was by increasing the usage of robust reporting guidelines such as CONSORT, STROBE and STARD. The *J Clin Pathol* supports this initiative, and the EQUATOR network is cited within the author instructions.

All the guidelines require that estimates should be given with some measure of precision that is dependent on the sample size. This precision usually takes the form of a confidence interval around the point estimate of effect. A review of the 3 months April to June 2009 revealed that of 28 *J Clin Pathol* articles presenting sample data, only six mentioned confidence intervals, and only four presented them correctly for all relevant estimates. To facilitate adherence to the current guidelines for contributors to *J Clin Pathol*, this paper gives an overview of confidence intervals and their interpretation in the most commonly encountered data scenarios. We do not give details of calculation as this can be done by any good statistics package, and the researcher

does not need to get involved with the intricacies of the process. The emphasis is on understanding the rationale and application.

AN ILLUSTRATIVE EXAMPLE

Suppose we want to know what percentage of malignant mesotheliomas (MMs) express GATA-6. We will look for GATA-6 expression in a random sample of MMs. The percentage that expresses GATA-6 in that sample will give an estimate of the population percentage.

THE SAMPLE ESTIMATE WILL NOT USUALLY BE THE SAME AS THE POPULATION VALUE.

How good an estimate the sample yields depends on the sample size. Larger samples give more precise estimates. It makes sense to take into account the sample size when interpreting the sample results.

WHAT IS A CONFIDENCE INTERVAL?

A confidence interval gives the range of population scenarios that the sample is compatible with. It is a measure of precision attached to, and built around, a study estimate.

Suppose in a random sample of 10 MMs we find that eight express GATA-6. Our best estimate of the percentage of all MMs that express GATA-6 is $8/10 = 80\%$.

BUT THIS DOES NOT MEAN THAT THE POPULATION PERCENTAGE IS 80%. We could quite reasonably expect to obtain 8/10 if the population percentage were 90% or 70% (when we would expect to obtain about seven or nine of our sample expressing GATA-6). However, if the population percentage were actually 20%, for instance, then we would expect to see about two expressing GATA-6 and would be surprised to find as many as eight.

The confidence interval is built around the sample estimate (80%) and gives the range of population values that can reasonably be expected to yield a sample estimate of 80% from a sample of that size. How "reasonable" it is for the interval to contain the population value is quantified by the percentage confidence interval that we choose to give:

- ▶ The 95% confidence interval gives the range within which we are 95% confident the population value lies (based on our sample)
- ▶ The 90% confidence interval is narrower and gives the range within which we are 90% confident the population value lies
- ▶ The 80% confidence interval is even narrower and gives the range within which we are 80% confident the population value lies. The interval is narrower as we are less confident it actually contains the population value.

It is not impossible that the population value lies outside the confidence interval; we just know that it is unlikely with a given level of confidence. A 95% confidence interval is estimated from our sample. We may be unlucky and obtain one of the 5% of random samples that yield a confidence interval that does not contain the population value.

If we calculate the 95% confidence interval around our sample estimate of 8/10, it is found to be (49% to 94%). This means that we are 95% confident that the population percentage of MMs that express GATA-6 is between 49% and 94%.

The 80% confidence interval for the sample estimate (8/10) is (60% to 91%), which, as expected, is narrower, since we are less confident (80% as opposed to 95%) that it contains the population value.

It might seem odd that on the basis of 10 MMs we obtain limits that are not values that could be obtained from a sample of 10 (ie, with 10 MMs we could not obtain a sample estimate of 49% as we could have either 4/10 (40%) or 5/10 (50%), but not 4.9 of the 10, expressing GATA-6). This does make sense though because the population value might actually be 49% even though we would never obtain this sample estimate from 10 MMs. As stated above, **THE SAMPLE ESTIMATE IS UNLIKELY TO BE EXACTLY THE SAME AS THE POPULATION VALUE**, particularly where small numbers are sampled.

LARGER SAMPLES

If we take a larger sample, the specified percentage confidence interval will be narrower as we will have a more precise estimate.

For example, if 80 of 100 randomly sampled MMs express GATA-6, our sample estimate is still 80% (80/100) but the 95% confidence interval is now (72% to

Correspondence to: Dr A M Wade, UCL Institute of Child Health and Great Ormond Street Hospital, MRC Centre of Epidemiology for Child Health, 30 Guilford Street, London WC1N 1EH, UK; awade@ich.ucl.ac.uk

88%), which is more precise than the (49% to 94%) obtained with the smaller sample.

The 80% confidence interval for the 80/100 is (75% to 85%).

Hence, as we would expect, the larger sample gives a more precise estimate.

TERMINOLOGY

The edges of the confidence interval are known as the confidence limits. For example, the 95% confidence limits for the example above are 49% and 94%. The 80% confidence limits are 60% and 91%. Sometimes "confidence interval" is abbreviated as "CI" and "confidence limits" as "CL".

The limits may be separated by the word "to" as we have done (60% to 91%) or by a comma (60%, 91%). Although a dash is sometimes used (60–91%), this is not recommended as it can cause confusion with minus signs.

The width of a confidence interval is the difference between the confidence limits (ie, how far the interval spans). For example, the 95% confidence interval in the example based on a sample of 10 is (49% to 94%), and hence of width 45% (94–49), while based on a sample of 100 the width is only 16% (88–72).

Note that it is most common to cite 95% confidence intervals and if no percentage is given then 95% confidence should be assumed.

CONFIDENCE INTERVALS FOR OTHER POPULATION PARAMETERS

The example just given considered a single population percentage (the percentage of MMs expressing GATA-6) and illustrates the simplest case. Confidence intervals can, and should, be built around any sample estimate of a population value. For example, the mean age of those expressing GATA-6 or the difference in average ages of those presenting with two different diagnoses.

How wide a particular 95% confidence interval is always depends on the sample estimate and the sample size. For numeric values, such as mean age or Ki-67 value, the width of the confidence interval will also depend on the variability of the sample measurements.

CALCULATION OF CONFIDENCE INTERVALS

For the majority of sample estimates a measure of precision known as the standard error can be calculated. There are two main exceptions to this. First is where no distribution can be assumed

and non-parametric estimates such as the median are used. Second, if sample numbers are small and the standard error cannot accurately be established, exact methods of confidence interval estimation need to be used.

If a standard error can be calculated then this is used to construct a confidence interval for the estimate. The standard error is positive and larger values mean that the estimate is less precise. Smaller samples yield larger standard errors and wider confidence intervals.

A 95% confidence interval is given by (sample estimate \pm 1.96 standard errors) = ((sample estimate–1.96 standard errors) to (sample estimate+1.96 standard errors)).

An 80% confidence interval is given by (sample estimate \pm 1.28 standard errors) = ((sample estimate–1.28 standard errors) to (sample estimate+1.28 standard errors)).

For example, the standard error of the percentage expressing GATA-6 based on a sample of 100 MMs can be calculated to be 4%.

The 95% confidence interval for the 80% sample estimate is calculated as: $(80 \pm 1.96(4)) = (80 \pm 7.84) = (72.16\% \text{ to } 87.84\%)$ which can be rounded to (72% to 88%).

The 80% confidence interval is calculated as: $(80 \pm 1.28(4)) = (80 \pm 5.12) = (74.88\% \text{ to } 85.12\%)$ which can be rounded to (75% to 85%).

CLINICAL INTERPRETATION OF CONFIDENCE INTERVALS

It is important to give confidence intervals around all sample estimates as they allow clinical interpretation of the study results that take into account the sample size. We should believe in the estimate based on 100 MMs more than that based on only 10 MMs.

The 95% confidence interval for a population estimate gives the range of population scenarios with which the sample values are compatible (with 95% confidence). We can reasonably exclude values outside this interval as being unlikely, whereas we should consider the possibility that anything within the interval could reasonably be true.

RECENT PUBLISHED EXAMPLES

Florena *et al*⁵

This study compared megakaryocytes (MKCs) between 30 patients with essential thrombocythaemia (ET) and 30 patients with primary myelofibrosis (PMF). The average difference of 11

(PMF 52, ET 41) was not significant ($p = 0.068$). A 95% confidence interval for the difference (–1.5 to 23.5) shows the range of average differences with which these two samples are compatible. We cannot exclude an increased average of 23.5 in the PMF group.

In order to fully interpret the results we need to consider not only the p value but also the confidence limits, in particular whether 23.5 is a clinically important difference that we would want to investigate further. The sample data are compatible with MKC values for patients with ET being on average 1.5 higher and 23.5 lower than for patients with PMF.

By contrast, Florena *et al*⁵ also recorded the percentages of MKCs positive for BCL-XI for each of the 60 patients and found a significant difference of 15.5 (PMF 35, ET 50.5; $p = 0.036$) with a 95% confidence interval of (2.2 to 28.8). This interval shows that the data are compatible with an average difference between PMF and ET as small as 2.2 or as large as 28.8. To interpret the data we need to consider the clinical relevance of those limits. Interpretation cannot be made merely using the p value.

Al-Mulla *et al*⁶

This study showed that the median age of onset of breast cancer was 55 years for individuals with mutation 185delAG in exon 2 (95% confidence interval 46.7 to 59.5). It is important to consider the confidence interval as this takes into account the sample size that the median age (55 years) is based on. If a much smaller sample were used which yielded the same average but a wider confidence interval of, for example (23 to 78 years), we would need to interpret the information differently. In the latter scenario we cannot draw any conclusions as the range given (23 to 78 years) is so wide as to be uninformative.

OTHER APPLICATIONS

A review of *J Clin Pathol* articles in the 3 months April to June 2009 shows that a variety of summary statistics are used regularly in *J Clin Pathol*. In all instances confidence intervals are appropriate and should be presented alongside the summary estimates. The main forms are as follows.

Single percentage

Prevalence studies yield sample estimates of population percentages. Other instances where single percentages are obtained are reliability studies (κ) and diagnostic studies (sensitivity, specificity,

positive and negative predictive values). Note that for diagnostic studies the sample size for the different estimates varies. For example, sensitivity is based on those with the disorder and specificity on those without, the size of these two groups possibly being quite different.

Single mean or median

We should be mindful of whether the measurements are normally distributed (hence the mean of the values will be a valid summary) or not (when the median is a better summary). Either way a confidence interval should be given alongside the mean or median.

Differences in percentages between two groups

The analysis here is typically χ^2 or Fisher's exact test. A confidence interval should be given for the percentage difference between the two groups.

Difference in means or medians between two groups

The analysis here is typically two sample t test (means of normally distributed data) or Mann-Whitney U test (medians of non-normally distributed data). Some skew data are actually log-normally

distributed and means can be calculated on the transformed scale. A confidence interval should be given for the difference in mean or median between the two groups.

For differences in percentages between two groups and difference in means or medians between two groups, note that a common mistake is to calculate confidence intervals for each group separately but this is not what is required. The confidence interval to present is for the DIFFERENCE between groups.

Correlations

The relationship between two variables using a correlation coefficient also requires a confidence interval since it is a sample estimate of the population value.

Hazard ratios

In *J Clin Pathol* it is not uncommon to see time to event (survival) data presented. The hazards of the event are compared between groups and it is important that their ratio is given with a measure of precision that is dependent on sample size.

OVERVIEW

This article has highlighted an area where vast improvements can be made

in the presentation and interpretation of study results in the *J Clin Pathol*. We expect authors to follow the guidelines of the EQUATOR network, and having a good understanding of confidence intervals is vital to this. Our hope is that this article will help researchers to make the best use of the data they have collected.

Competing interests: None.

Provenance and peer review: Commissioned; not externally peer reviewed.

Accepted 1 July 2009

J Clin Pathol 2009;**62**:1059–1061.
doi:10.1136/jcp.2009.069906

REFERENCES

1. **CONSORT**. <http://www.consort-statement.org> (accessed 29 September 2009).
2. **STROBE statement**. <http://www.strobe-statement.org> (accessed 29 September 2009).
3. **STARD statement**. www.stard-statement.org (accessed 29 September 2009).
4. **EQUATOR network**. www.equator-network.org (accessed 29 September 2009).
5. **Florena AM**, Tripodo C, Di Bernardo A, *et al*. Different immunophenotypical apoptotic profiles characterize megakaryocytes of essential thrombocythaemia and primary myelofibrosis. *J Clin Pathol* 2009;**62**:331–8.
6. **Al-Mulla F**, Bland JM, Serratt D, *et al*. Age-dependent penetrance of different germline mutations in the *BRCA1* gene. *J Clin Pathol* 2009;**62**:350–6.

BMJ Masterclasses

BMJ Masterclasses are educational meetings designed specifically to meet the learning needs of doctors. They help doctors keep up to date with the latest evidence and recent guidelines in major clinical areas, enabling them to use the latest evidence to make better decisions. The latest evidence, recent guidelines and best practice are delivered in an interactive and informative manner by leading experts. The speakers are specifically chosen as highly-skilled communicators who can authoritatively enthuse the audience and interpret the latest research and guidelines into practical tips for busy doctors. BMJ Masterclasses have proved a huge hit with clinicians, with many saying they have influenced their clinical practice.

<http://masterclasses.bmj.com/>

BMJ
masterclasses
Putting the latest evidence based medicine into practice