RESEARCH METHODS & REPORTING

STATISTICS NOTES

Uncertainty beyond sampling error

Douglas G Altman *professor of statistics in medicine*[1], J Martin Bland *professor of health statistics*[2]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK; [2]Department of Health Sciences, University of York, York YO10 5DD, UK

Statistical analysis of research results mainly uses confidence intervals and hypothesis tests to capture the uncertainty rising from our study being on a sample of participants drawn from a much larger population, in which our interest mainly lies.[1] But beyond the issue of sampling variation there are other sources of uncertainty that may be even more important to consider. In measurement, a distinction is made between precision, which is how variable are measurements on the same person by the same method made at the same time, and accuracy, which is how close the measurement is to what we actually want to know. For example, if we were to ask a group of patients on two occasions how much alcohol they typically consume, this would enable us to estimate precision, how repeatable answers are, but not how close these answers are to how much they actually drink, which we might suspect to be higher. In the same way, a confidence interval tells us about the precision of research results, what would happen if we were to repeat the same study, not their accuracy, which is how close the study is to the truth.

In general, beyond the imprecision or uncertainty of numerical results arising from sampling, the main concern is the possibility that the study results are biased. Recent developments in appraising published randomised trials have switched from considering "quality" (essentially undefinable) to assessing explicitly the risk of bias in relation to the way the study was done.[2] Here sources of possible bias include lack of blinding and losses to follow up (missing data). But bias is especially relevant in non-randomised studies, where there will be major concerns about possible confounding, where an apparent relationship between two things may be the result of the relationships of both to a third.

A further source of uncertainty concerns the extent to which the results of research conducted in a particular setting with selected participants can be taken as applying equally to a wider group of patients in a different location. Judgement of generalisability (also known as external validity) is challenging.[3] In a clinical trial, the part of the larger population we particularly care about does not yet exist. We want to know what would happen to future patients if we were to apply either of the trial treatments. Changes over time in the nature of disease, in the fitness of the patient, in nutrition, in ancillary treatments, and so on, may all make our sample unrepresentative as a guide for future action.

For example, the UK review of evidence relating to mammographic screening outlined three sources of uncertainty in relation to the pooled estimated effect from a meta-analysis of the results of all the randomised trials.[4] First, there was uncertainty due to sampling variation, as previously discussed.[1] Second, there was uncertainty from some methodological weaknesses of the trials. Third was uncertainty about whether the results from the trials were still relevant 25 years later, after major changes in cancer incidence, management of and treatments for breast cancer, and the technology of mammography. Unlike sampling variation, which is quantified in the confidence interval and is uncontroversial, the uncertainty from the other causes cannot easily be quantified and remains the source of fierce debate.

Large numbers, increasingly common in this era of big data, will produce narrow confidence intervals. These can create an illusion of accuracy, but they ignore all sources of possible bias that are not affected by sample size, and so these other sources become relatively more important.[5 6] A recent example of a very precise but seriously wrong answer purported to show that skin cancer was protective for heart attacks and all-cause mortality.[7] So, although confidence intervals are a valuable way of depicting uncertainty, they are always too narrow in the sense that they reflect only statistical uncertainty, precision rather than accuracy.

Many journals require authors to consider in their discussion the limitations of their study—some even require this in the article's abstract. Issues raised there help readers to judge what extra uncertainty might apply to the study, including whether observed effects may be affected by bias. It is common for authors to say that their results should be "interpreted with caution" (including >700 in the *BMJ*), but who knows what that means in practice? The GRADE group have developed a framework for a more structured approach to assessing the reliability of research findings that addresses the aspects outlined above.[8]

Correspondence to: D G Altman doug.altman@csm.ox.ac.uk

1   Altman DG, Bland JM. Uncertainty and sampling error. *BMJ* 2014;349:g7064.
2   Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
3   Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005;365:82-93.
4   Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013;108:2205-40.
5   Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol* 2004;33:1389-97.
6   Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014;7:342-6.
7   Lange T, Keiding N. Skin cancer as a marker of sun exposure: a case of serious immortality bias. *Int J Epidemiol* 2014;43:971.
8   Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-6.

Cite this as: *BMJ* 2014;349:g7065