Assessing new methods of clinical measurement

' … the effect of measurement error on clinical practice continues to be underestimated.'[1]

Measurement is such a routine part of clinical and research practice that it is taken for granted and its importance forgotten. Poor quality measurements will impair reliable diagnosis and prognosis, and inaccurate measurements will hamper the conduct of randomised trials and epidemiological studies.

Progress in clinical measurement comes from new technology or techniques. There should be evidence that a new method of measurement performs well before it can enter clinical practice, yet there is no such requirement. Indeed, there is no clear definition of 'satisfactory'.

In a few areas measurement issues are taken seriously; for example, guidelines were developed to evaluate and grade the performance of semi-automated blood pressure machines in carefully designed experiments.[2] In some fields researchers are trying to establish standard outcome measures for randomised trials; more should do so.[3] Overall, though, clinical measurement does not receive the same attention as laboratory measurement.

Most medical specialties have no organised evaluation of methods of measurement. Nor is it easy to find guidance in textbooks on how to assess new or existing methods of measurement, and these issues do not feature in the set of research methods usually taught. Perhaps as a consequence, measurement studies have been carried out in an ad hoc manner. However, many such studies are performed and published across all medical specialties, indicating wide recognition of the need to assess measurements.

Unfortunately, the standard approach to method comparison for some decades was to calculate the correlation coefficient between the values obtained by two methods on the same individuals. The deficiencies of this approach have been described frequently[4] but that approach is still in use. In brief, correlation assesses overall association across many measurements not individual agreement, and thus does not address a clinically relevant question. It is clear that peer review does not prevent the publication of papers with inappropriate statistical methods. And once faulty methods are in the literature, they will be copied.

When assessing a new measurement method, we may wish to answer questions such as:

1. How variable are the measurements made using two different methods on the same individual?
2. How variable are repeated measurements made using a specific method by the same observer on the same individual?
3. How variable are measurements made using a specific method by different observers on the same individual?

For simplicity, I will refer to all of these as method comparison problems. Studies that collect a single measurement by each method on each individual can only address question 1.

In laboratory medicine it is usually possible to compare a new assay with a highly accurate reference standard method, but in most clinical areas assessment of measurement methods must be done in the absence of a (near) truth; familiar examples include blood pressure, lung function, and skin-fold thickness.

A new clinical method will be compared with a current standard method (there may be more than one in use). We wish to know how well the new method agrees with the standard method. To me that question must be addressed in the context of the individual patient, a notion that underpinned the development in around 1980 of the limits of agreement method for method comparison studies: the so-called 'Bland-Altman method'.[4,5] In brief, the idea is simply to examine the distribution of the differences between measurements by two methods for each individual in the study, in particular calculating the mean and standard deviation (SD) of the distribution. Assuming that the differences between the methods have a normal distribution, we would predict that 95% of such differences in future would lie between mean–2SD and mean+2SD, which we called 95% limits of agreement.[4,5] (We also implicitly assume that all observers take equally good measurements.) Various extensions to the basic method have been developed.[6,7]

We suggested a histogram of differences to check that the assumption of (approximate) normality was reasonable, and a plot of the difference against the mean to check that the overall results were relevant across the relevant range of the measurements. The latter plot has become synonymous with the method; indeed, many seem to believe that the plot is the method.

A specific example is a comparison of tympanic infrared and axillary mercury thermometry for 94 children presenting with acute cough.[8] The mean difference between the axillary and tympanic measures was 1.18°C (SD = 0.96°C), and the 95% limits of agreement were –0.73 to +3.09°C. Twenty (21%) of the children were febrile as judged by the standard mercury method, and just four (4%) by infrared thermometry. The plot suggested that agreement tended to deteriorate with falling temperatures. The authors concluded that ' … [t]he mean difference and limits of agreement are too large for this tympanic thermometer to replace the mercury thermometer in normal clinical practice.'[8]

In contrast, Cuschieri *et al* used the same methods to explore differences between central and (the standard) mixed venous pCO_2 measurements in 83 critically ill patients. They reported that ' … the data points in the Bland/Altman plot appear to be well scattered, and the limits-of-agreement band (mean±SD) of –0.64 to

+0.92 appears to be relatively narrow, suggesting good overall agreement between the mixed and central pCO_2 differences'.[9]

In developing the limits of agreement method, our main goal was to address question 1. The same approach can also be used for questions 2 and 3, when there are replicate measurements or multiple observers. In general, a full examination of the properties of a new method requires such investigations.

Some studies have a more specific target which can be adequately addressed by single measurements; the study by Nicolaï *et al*[10] in this issue of the *BJGP* is one such. They showed that although measurements of ankle brachial index (ABI) were on average the same in primary care practices and a vascular laboratory, the individual differences had a very wide scatter, indicating that the measurements are not comparable and hence not interchangeable. The authors identified as a factor the lack of a standardised measurement method in primary care practices, but they did not compare the performance of those various approaches and the study was not really large enough to do so. Further studies might be done to ascertain which method should be recommended in that setting. Also, it is important to assess the repeatability of laboratory measurements for comparison.

Two methods will never agree exactly. And whether a method provides measurements that are acceptable cannot be answered directly by statistical analysis alone (and certainly not by a *P* value). Rather, statistical analysis can provide a meaningful summary of the evidence from a study to inform clinical judgement. For example, it might be felt that the ABI measurements in the clinic should generally (say 95% of the time) be within 0.2 units of the value obtained in the laboratory. If that is not achieved then one way to improve measurements is to take the average of 2 or 3 readings. If agreement fails to meet the target the method should probably not be used.

Clinical measurements are taken for a reason. The consequences of classifying patients based on clinic measurements, whether ABI or blood pressure, should be considered, as was done by Nicolaï *et al*.[10]

Douglas G Altman,
Director of Centre for Statistics in Medicine, University of Oxford.

Provenance
Commissioned; not peer reviewed.

Funding body
Douglas G Altman is supported by Cancer Research UK.

REFERENCE

1. McDonough PG. Measurement error — 'How much of a difference does it take to make a difference?'. *Fertil Steril* 1997; **67(4):** 790–791.
2. O'Brien E, Petrie J, Littler W, *et al*. The British Hypertension Society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems. *J Hypertens* 1990; **8(7):** 607–619.
3. Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials* 2007; **8:** 39.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1(8476):** 307–310.
5. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32:** 307–317.
6. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8:** 135–160.
7. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat* 2007; **17:** 571–582.
8. Hay AD, Peters TJ, Wilson A, Fahey T. The use of infrared thermometry for the detection of fever. *Br J Gen Pract* 2004; **54(503):** 448–450.
9. Cuschieri J, Rivers EP, Donnino MW, *et al*. Central venous-arterial carbon dioxide difference as an indicator of cardiac index. *Intensive Care Med* 2005; **31(6):** 818–822.
10. Nicolaï SPA, Kruidenier LM, Rouwet EV, *et al*. Ankle brachial index measurement in primary care: are we doing it right? *Br J Gen Pract* 2009; **59(563):** 422–427.

DOI: 10.3399/bjgp09X420905

ADDRESS FOR CORRESPONDENCE

Douglas G Altman
Centre for Statistics in Medicine,
University of Oxford,
Wolfson College Annexe,
Linton Road, Oxford OX2 6UD.
E-mail: doug.altman@csm.ox.ac.uk