On the precision and reliability of IOP measurements

Many studies have evaluated the quality of intraocular pressure (IOP) measurements. This commentary has been prompted by reading one of the most recent, which raises questions concerning criteria for the acceptability and standardisation of clinical instruments. Although this discussion relates specifically to tonometry, the general issue is relevant in every area of clinical measurement.

Studies to evaluate clinical methods fall into two categories—reliability studies which ask 'Is this a good method?' and validity studies which ask 'Is this the right method?' Reliability studies determine the consistency of repeated measurements taken by the same method on the same patients on a number of occasions, whereas validity studies compare measurements taken by one method with those taken by another 'gold standard' method. Since a method may be expected to compare with itself better than with another, it is generally considered that validity cannot exceed reliability. Thus reliability studies are fundamental in determining method quality.

Reliability studies permit evaluation of two aspects of method quality—precision and reliability. These terms have well established meanings in psychometric/statistical literature. Precision (given by the standard error of measurement) relates to the consistency of the measurements, and is a general term covering both repeatability and reproducibility, whereas reliability (given by the test-retest correlation coefficient) relates to the quality of inferences that may be drawn from the measurements. Precision tells us how much an individual patient's measurements may be expected to vary from one occasion to another as a result of random error, and thereby sets limits on how much change is significant. Reliability tells us how well the measurements can discriminate between patients. Good precision supports good reliability, but reliability depends also on how much variation there is between patients. Thus we cannot, even with the most precise measurements, discriminate reliably between patients whose true states are identical, while at the other extreme we do not need a very precise measure to discriminate between patients who are markedly different. It is notable that, in spite of many studies, the quality of IOP measurements has never been clearly established in terms of precision and reliability.

Many researchers studying IOP measurement have chosen to evaluate how one tonometer compares with another rather than assessing the precision and reliability of a particular tonometer. Of those studies concerned with precision and/or reliability, some have used different examiners and others the same examiner to take repeated measurements, some have evaluated contact and others non-contact tonometry, and none has analysed data in the terms outlined above. In fact, there are no two studies which provide independent evaluation of tonometry in a directly comparable form. What can we conclude about the quality of IOP measurements from such studies?

A basic requirement of a reliability study is that repeated measurements should be taken under identical conditions and that the measuring procedure itself should not cause any change in the condition being measured. With contact tonometry, however, it is well documented that IOP decreases on repeated measurement. It is therefore difficult, if not impossible, to make meaningful estimates of contact tonometry precision, unless a state is eventually reached at which IOP no longer changes with repeated measurement. On the other hand, repeated measurement with non-contact (air puff) tonometers appears to cause minimal IOP decrease. So estimates of precision for these instruments will have greater validity and, indeed, may set a limit for the best precision that can be achieved in tonometry.

The most recent study of non-contact IOP precision, and the only one whose data may be interpreted in the terms discussed, is by Vernon. Vernon assessed the repeatability of IOP measurements using the Pulsair 2000 tonometer. One hundred non-glaucomatous patients provided three measurements from each eye within a 15 minute period. All measurements were taken by the same experienced examiner. Results showed: (i) the sample mean of the first measurement in both right and left eyes was significantly higher than that of the second or third measurement, and (ii) comparison of second and third measurements gave 'coefficients of repeatability' (2 standard deviation of test-retest differences) of right eye 4.2 mm Hg and left eye 3.6 mm Hg. It was concluded that the instrument 'passed the British Standard for reproducibility of a standard test method'.

Although he did not analyse them in this manner, Vernon's data may be expressed in terms of precision and reliability. Firstly, how close is any single measurement to the patient's true IOP? A 95% confidence interval for true IOP may be approximated as: observed IOP plus or minus 2 standard errors of measurement (SEM=standard deviation of test-retest differences/sqrt(2)). In Vernon's study SEM obtained by averaging the two eyes is 1.4 mm Hg. This means that any individual measure of IOP will lie in a range of plus or minus 3 mm Hg (rounding to the nearest whole millimetre) from the true IOP, so a patient who measures 21 mm Hg could have a true IOP of anything from 18 to 24 mm Hg.

Secondly, how much difference do we need between measurements taken on two occasions to conclude that IOP has truly changed? A 95% confidence interval for stable IOP values may be estimated as plus or minus 2 standard deviations of test-retest differences. In Vernon's study this is represented by the reported 'coefficient of repeatability' which averages and rounds to 4 mm Hg. Thus any difference in individual IOP of 4 mm Hg or less is indistinguishable from measurement error; significant change must exceed 4 mm Hg. Corresponding values from a study of Goldmann tonometry cited by Vernon are: true IOP = measured IOP plus or minus 4 mm Hg, and criterion for true change of 5 mm Hg; in each case 1 mm higher (that is, less precise) than corresponding Pulsair values. This natural variation of IOP measurements in stable patients, over ranges of 6 mm Hg (Pulsair) and 8 mm Hg (Goldmann), will surely contribute to the occurrence of some 'erroneously high IOPs ... which are never substantiated in hospital'. There may also, of course, be erroneously low IOPs but these will generally not reach...
hospital. In short, IOP measurements appear not to be very precise. Vernon does not actually report a value for Pulsair reliability but this may be inferred from his data, using estimates of the standard error of measurement and between patient standard deviation, to be approximately 0.83.

Are these levels of precision and reliability acceptable, and is the method standardised? Vernon's study of the Pulsair tonometer concludes 'yes' to both questions. Vernon's conclusion that Pulsair repeatability is acceptable is based on the observation that >95% of the differences between second and third IOPs fell within 2 SD from the mean difference (p 557) and that it therefore 'passes the British Standard for reproducibility of a standard test method' (abstract). The suggestion is that this distribution of differences represents a criterion for judging acceptable precision and, by reference to BS 5497, standardisation. This, however, is not the case. Rather, the data are simply conforming to a statistical principle that 95% of values in any normal distribution will fall plus or minus 2 SD from the mean and, since test-retest differences will tend always to be normally distributed, the principle applies here. One must therefore consider the questions of acceptability and standardisation from a different perspective.

With regard to acceptability (is it a good method?), this is partly a matter of whether the measurements are more precise than those given by other methods, and partly of whether this precision is good enough in the clinic. In this case, Vernon's data suggest that the precision of the Pulsair is slightly better than that of the accepted gold standard Goldmann tonometer as reported by Phelps and Phelps. However, the two studies span almost a 20 year interval, data are not presented in the same form and, as already noted, precision estimates for contact tonometry may be problematic. Unfortunately, data from other published studies do not permit any detailed comparison of precision estimates for non-contact tonometers, so it is difficult to conclude whether the precision of any tonometer is acceptable compared with any another. Clinical acceptability on the other hand may be judged by reliability—that is, a reliable method should discriminate effectively between the patients on whom it is used. In this case, the estimated reliability of 0.83 from Vernon's data may represent acceptable though not very good discrimination, bearing in mind the fact that visual acuity and contrast sensitivity letter charts generally achieve reliabilities in excess of 0.9 even within 'normal' populations. Bearing in mind also that the estimated precision and reliability of Pulsair tonometry would very likely be even poorer if different and/or inexperienced operators were included in the study, it might be argued that the reason the method is acceptable is not because it is high precision and reliability, but because clinicians screening for glaucoma do not rely on tonometry alone and therefore poor IOP measurement does not significantly increase the risk of false negative results, even though it may generate some (safe) false positives.

With regard to standardisation, an essential principle in BS 5497 is that 'preparing a standard requires careful evaluation of the method by a number of laboratories' and 'for a standardised method, repeatability will be approximately the same for all laboratories applying the method'. There is no evidence in the tonometry literature of any systematic effort to meet these conditions. Indeed, if one is strict, they cannot be met since BS 5497 applies to studies in which all the samples being measured are assumed to be identical. Clearly this is not the case with clinical measurement. In terms of BS 5497, a tonometer might be standardised for measurement of IOP in normal corneas but not on actual patients. To achieve standardisation in clinical application would require controlled studies of precision and reliability carried out in different clinical centres, with well defined patient populations and protocols and a unified approach to data analysis. This approach may represent evaluation 'overkill' for IOP measurement, which already has clinical acceptability as discussed above, but for other forms of clinical measurement it may be desirable. Probably the greatest impediment to achieving unequivocal standardisation of clinical methods has been the lack of any standardised approach to the design, conduct, and analysis of evaluation studies. BS 5497 provides a framework for such an approach, but it cannot be literally applied to the clinical assessment of patients. Surely, in the present era of quality consciousness, now is the time for the clinical community to tackle this issue.

JAMES M GILCHRIST
Department of Optometry, University of Bradford, West Yorkshire