Thorax 1991;46:454–456

Statistics in respiratory medicine · 2

Repeatability and method comparison

S Chinn

Repeatability and reference ranges for change

A measurement that is totally unrepeatable clearly has no validity. Repeatability, however, or reproducibility, is an ambiguous concept without precise definition. To a clinical chemist it may mean reproducibility of results from an autoanalyser for a single blood sample. A technician may spend hours perfecting a non-automatic technique. But if two blood samples taken from the same subject within hours give very different results laboratory repeatability may be relatively unimportant.

The repeatability of most respiratory measurements is necessarily of the "time to time" type with the interval between measurements measured at least in minutes but possibly in hours or days. Differences in the results obtained depend on the time gap, the variation increasing—that is, repeatability decreasing—with the length of the gap. Neild *et al*[1] measured forced expiratory volume in one second (FEV_1), peak expiratory flow (PEF), respiratory resistance, and specific airway conductance in 25 non-asthmatic subjects three times on each of three consecutive days, and reported the "within subject within day" variance of each, and also the estimate of *additional* variance due to variation within subjects but between days.

Variances were used in the paper by Neild *et al*[1] because *components* of variance may be added. The usual measure of repeatability when only within day or only between day variation is studied is the within subject standard deviation. If more than two repeat measurements are carried out for some or all subjects then repeatability is most easily calculated from the results of a one way analysis of variance, with "subjects" as the "group" variable. The within subject standard deviation is the square root of the pooled within subject sum of squares divided by its degrees of freedom (that is, for those used to analysis of variance terminology, of the residual "mean square").

Frequently repeatability studies are carried out with just two repeat measurements per subject; indeed, for most purposes this is the most efficient design. Then it is natural to take the *differences* between the first and second measurements for each subject. For example, Chinn *et al*,[2] from a study designed to assess repeatability of histamine challenge tests, reported the mean difference between the post-saline FEV_1 values measured in 107 subjects on two occasions as 0·01 litres and the standard deviation of the differences as 0·42 l. To convert a standard deviation of *differences*, which has double the variance of a single FEV_1, to a within subject standard deviation of a single FEV_1 we divide 0·42 by $\sqrt{2}$ to get 0·30. Strictly speaking, the fact that the mean difference is not exactly zero but 0·01 should be taken into account. The within subject standard deviation is actually

$$\sqrt{\frac{106(0{\cdot}42)^2 + 107(0{\cdot}01)^2}{107 \times 2}}$$

but the result is the same to 2 decimal places.

The fact that either the standard deviation of the differences or the within subject standard deviation may be reported as a measure of repeatability may lead to confusion. The within subject standard deviation is referred to as the *single determination* standard deviation, and the 95% range (see article—May 1991, p391) derived from it as the single determination 95% range. This will indicate the limits around a single measurement that must be regarded as possible values for the true measurement—that is, how much reliance, say for diagnosis, can be placed on that reading.

If a patient is being monitored then we are interested in the *change* in values. The inherent variability, as measured by the single determination standard deviation or range, enters into both the initial and the subsequent measurement, and so the standard deviation and 95% range for change, are greater, by a factor of the square root of 2, than the single determination values. The 95% range for change can be calculated directly from the standard deviation of differences between the repeat measurements, or from the within subject standard deviation provided that the extra factor of the square root of 2 is remembered.

It is recommended that the single determination standard deviation or 95% range is used for measuring repeatability as such and the 95% range for change for use in assessment of patients. It is, however, of prime importance that whichever is used is clearly stated, so that results will not be misinterpreted and can be converted to the alternative form when this is needed.

Not all calculations are carried out on the original scale of measurement. The third article in this series will explain how to choose the scale. Many measurements—for example, PD_{20} in bronchial challenge testing—are analysed on a log scale. *All* calculations should be carried out on the log values, but an

Department of Public Health Medicine, United Medical and Dental Schools of Guy's and St Thomas's Hospitals, St Thomas's Campus, London SE1 7EH
S Chinn

Reprint requests to:
Miss Chinn

arithmetic mean on the log scale can be antilogged to give the geometric mean value and the 95% range, expressed as $\pm$ k, can be antilogged to $\times/\div$ antilog(k). Thus Chinn *et al*[2] reported a within subject standard deviation of $\log_{10}$ PD_{20} histamine as 0·27. The single determination 95% range was thus log $PD_{20} \pm 2 \times 0{\cdot}27$—that is, $\log(PD_{20}) \pm 0{\cdot}54$, or PD_{20} $\times/\div$ 3·47 μmol. Such a range is usually expressed, however, in units of doubling doses, obtained by dividing 0·54 by $\log_{10}(2)$ (= 0·301), to give $\pm$ 1·79 doubling doses.

The possibility of an overall shift in mean value between the first and the second test can be investigated by calculating the standard error and 95% confidence interval for the mean difference. For the post-saline FEV_1 mentioned above the standard error of the mean difference was $0{\cdot}42/\sqrt{107} = 0{\cdot}04$. The 95% confidence interval for the mean difference was thus from $-0{\cdot}07$ to $+$ 0·09 l. This tells us only the likely size of the *bias* between the first and the second occasion post-saline FEV_1, and little about repeatability. A true confidence interval is *not* a measure of repeatability.

Method comparison

One aspect in which methods can be compared is in their repeatability. For example, Oldham and Cole[3] reported the repeatability of nine indices of FEV_1, each calculated from five repeat blows. The most repeatable on the basis of the within subject standard deviation was the mean of all five, and the least repeatable the maximum of the first three.

With most alternative methods there is a possibility that they do not, on average, give the same answer. Bland and Altman[4] described in detail how to compare two methods on the same scale of measurement, illustrating their recommendations with PEF data. They give several reasons why the correlation coefficient should not be used, the most important of which are, firstly, that to agree perfectly the results from two measurements must lie on the line of identity, not just any straight line, and, secondly, that the correlation coefficient is influenced by the range of variation between the subjects, blood samples, or other units chosen to test the two methods. For a given level of agreement the correlation coefficient increases as the variance between the units increases.

Bland and Altman[4] recommended plotting the difference in the two results against the mean value from the two methods. Provided that there is no relation between differences and mean values, "limits of agreement" can be calculated as $\bar{d} \pm 2s$, where $\bar{d}$ is the mean difference and s the standard deviation of the differences. If the sample size, *n*, is less than 100, 2 should be replaced[5] by

$$t_{n-1,0{\cdot}05}\sqrt{(n+1)/n}.$$

Agreement, or more accurately lack of it, thus has two components, the *relative bias* as estimated by $\bar{d}$ and the *random variation* as estimated by s, which is at least as great as that predicted by the repeatability of each method. If the within subject (or between replicate if more appropriate) standard deviations are s_1 and s_2, then

$$s \geqslant \sqrt{s_1^2 + s_2^2}.$$

A 95% confidence interval for the relative bias can be calculated as for a paired *t* test—that is,

$$\bar{d} \pm t_{n-1,0{\cdot}05}\frac{s}{\sqrt{n}}.$$

As $\bar{d}$ estimates only the systematic component of lack of agreement, this is *not* a measure of agreement. As with repeatability, care must be taken to distinguish between confidence interval and the 95% limits of agreement.

Measurements on different scales

Bland and Altman[4] dealt only with measurements on the same scale. We cannot compare repeatability as measured by standard deviations in different units, as may be required in comparing repeatability of different indices of histamine challenge tests[6]—for example, PC_{20} in log (mg/ml) and the slope of the FEV_1 dose-response curve in $l(mg/ml)^{-1}$. The solution is to calculate a dimensionless statistic. Although dividing the standard deviation by the mean to give the coefficient of variation is still used, it is valid only in certain circumstances[7] (to be described in the third article in this series). It is better to calculate the ratio of between subject to total variation, known as the intraclass correlation coefficient, as used by Dehaut *et al*.[6] The maximum value of the intraclass correlation coefficient is 1, achieved only when repeatability is perfect. A value of zero (or less) denotes repeatability that is no better (or worse) than would be expected by chance. To be useful a measurement should have an intraclass correlation coefficient of at least 0·6. Baseline FEV_1 measured on two occasions 1–14 days apart[2] in 111 subjects had an intraclass correlation coefficient of 0·88. Repeated measurements of FEV_1 on the same day may give a value as high as 0·99.[3] In the simplest case of two components of variation, between subject (or other unit) and within subject, estimating the two components as described by Armitage and Berry[8] is straightforward. A statistician should be consulted before data collection, however, if there are more than two components. The components of variance are of direct use—for example, the effect of averaging three or five measurements can be compared.[1]

The use of the intraclass correlation coefficient implies that each component of variance has been estimated appropriately, from sufficient data (at least 25 degrees of freedom) and from a sample representing the population to which the results will be applied. When intraclass correlation coefficients are compared they should be obtained from data on the same sample of subjects, or from samples from the same population.

With two methods on different scales there is no "line of identity" on which the data should lie for perfect agreement. Indeed, there is no

reason why the relation between them should be a straight line. All that we require is the possibility of perfect calibration—that is, a smooth curve, increasing or decreasing, that describes one and only one value on each scale corresponding to a point on the other. Thus a straight line or exponential relation allows calibration whereas a sinusoidal relation does not: for any y value there would be several corresponding x values. Of course, it may be possible to transform one of the measurements into the same unit as the other, and calibration implies this. How we initially choose the transformation or scale of measurement will be described in the third article.

1 Neild JE, Twort CMC, Chinn S, *et al*. The repeatability and validity of respiratory resistance measured by the forced oscillation technique. *Respir Med* 1989;**83**:111–8.
2 Chinn S, Britton JR, Burney PGJ, Tattersfield AE, Papacosta AO. Estimation and repeatability of the response to inhaled histamine in a community survey. *Thorax* 1987;**42**:45–52.
3 Oldham PD, Cole TJ. Estimation of the FEV_1. *Thorax* 1983;**38**:662–7.
4 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
5 Healy MJR. Notes on the statistics of growth standards. *Ann Hum Biol* 1974;**1**:41–6.
6 Dehaut P, Rachiele A, Martin RR, Malo JL. Histamine dose-response curves in asthma: reproducibility and sensitivity of different indices to assess response. *Thorax* 1983;**38**:516–22.
7 Chinn S. The assessment of methods of measurement. *Statistics in Medicine* 1990;**9**:351–62.
8 Armitage P, Berry G. *Statistical methods in medical research*. 2nd ed. Oxford: Blackwell, 1987:196–200.