Accurate accuracy assessment

Review of basic principles

Jonas Ranstam[1], Leif Ryd[2] and Ingemar Önsten[3]

[1]Health and Society, Malmö University, SE-205 06 Malmö, Sweden. Tel +46 40-343633. Fax -921520. E-mail: jonas.ranstam@hs.mah.se; Departments of Orthopedics, [2]University Hospital, Lund, Sweden and [3]Malmö University Hospital, Malmö, Sweden
Submitted 99-02-18. Accepted 99-03-18

It is vital for scientific communication and for the development of new methods that measurements are correctly and uniformly assessed and interpreted. There is an international standard which should be used, see Appendix 1. However, in orthopedic research, there seem to be many confusing definitions. For instance, in one report (Manadan et al. 1997) accuracy is measured as the mean of absolute errors between "the measured and the calculated values" in 3 repeated measurements by 2 different observers. In another report (Buckland-Wright et al. 1995), accuracy is measured as the mean error in 4 repeated measurements, where "the error in an individual measurement was the absolute difference between that measurement and ... the mean value". In a third paper, the mean of absolute values is used instead of the true mean (Malchau et al. 1995). In a fourth paper (Feng et al. 1996), "accuracy" is used frequently, but never defined.

Accuracy is related to measurement error; both terms relate to the deviation from a true value. The deviation can be of two kinds, systematic and random. The systematic component is related to the trueness of the instrument (the opposite of bias), the random component to its precision (the opposite of variance).

Trueness can be assessed by using the difference between true and measured values in a large series of such results. When assessing trueness, a major problem is often finding the true values to compare measurements with (a "gold standard"). If no gold standard exists, and we wish to compare two imperfect instruments, other techniques must be employed, see, e.g., Bland 1995.

An instrument with only random errors

Assuming that an instrument has no bias, i.e., only random errors, its accuracy can be assessed using repeated measurements on the same subject. The standard deviation of the differences of repeated recordings will give a measure of both its precision and accuracy; since no systematic errors exist, the two concepts are congruent. For example, assume that in a series of subjects we study a new technique to assess the distance between two well-defined points on the skeleton, Table 1.

The standard deviation is the square root of the variance, which usually is defined as

$$\frac{1}{n-1}\sum(x_i - \bar{x})^2$$

In this formula, n is the number of observations, $\bar{x}$, the mean in our study population, is an estimate of an implicit target population mean, and x_i is the value of the i^{th} observation. In our example, the mean difference is 0 (the mean absolute difference is 1.6). Hence, the variance of the difference can be calculated as

$$\frac{1}{n-1}\sum d^2 = \frac{14}{4}$$

Pairs of measurements

	Reading (mm)		
Subject	First	Second	Difference (d)
1	49	48	1
2	39	38	1
3	51	53	–2
4	43	41	2
5	48	50	–2

The standard deviation, s, is the square root of this, i.e., 1.87. However, the reason for using $(n - 1)$ instead of n when calculating the standard deviation is the loss of one degree of freedom in estimating the mean. This is not the case when assessing accuracy; we assume that we have only random errors. Therefore, we should not estimate the mean; it is zero by assumption. We should divide by n instead of by $(n - 1)$. Our standard deviation is therefore $\sqrt{\frac{14}{5}} = 1.67$. However, when n is large, this is not important.

Hitherto we have considered differences in recordings from two measured values only. This is known as the instrument's repeatability or reproducibility, see Appendix 1. The concept of accuracy and measurement error is, however, related to deviation from a true value. These deviations have only half the variance of differences between two measurements. To obtain an "accuracy-variance", we should thus divide the "repeatability-variance" by 2. The correct standard deviation of the measurement error in our example is, therefore,

$$\sqrt{\frac{14}{10}} = 1.18$$

This result is identical with what we would obtain if we used the square root of the "residual mean square" from an algorithm for repeated measurement analysis of variance, which is an often recommended approach, see, e.g., Bland, 1995.

To obtain a limit, below which at least 95% of all errors can be expected to fall (assuming normal distribution), we multiply the standard deviation by 1.96. In our example, this is

$$1.96 \times \sqrt{\frac{14}{10}} = 2.31$$

An instrument with both systematic and random errors

If an instrument is biased (has systematic errors), one must assess the magnitude of this. A series of measurements of the same object will be necessary; the mean difference between these measurements and the true value estimates the bias. A 95% confidence interval of the mean difference will provide a margin of error for the estimated bias:

$$d_0 \pm 1.96 \times \frac{s_{d_0}}{\sqrt{n}}$$

Note that, when we calculate the standard deviation of individual measurements for the construction of this confidence interval, we should use $(n - 1)$ in the denominator; we do, indeed, estimate the mean:

$$s_{d_0} = \sqrt{\frac{\sum (d_0 - d_0)^2}{n - 1}}$$

An assessment of the statistical uncertainty in the precision (the random errors) of the instrument can be based on the same measurements, but by constructing a 95% confidence interval for the reproducibility or the repeatability (see Appendix 2). Note, however, that when reliability is assessed using deviations from a true value the variance should be doubled by analogy with the previously described difference between "reliability-variance" and "accuracy-variance".

Other techniques

We have stressed the importance of using coherent definitions of accuracy and measurement errors and on the existence of internationally accepted standard definitions; we suggest that they should be used. However, we have not attempted to present a complete operational description of how accuracy should be assessed in practice. Various methods and techniques exist. For instance, in some circumstances it would be appropriate to use a coefficient of variation, a kappa coefficient or an intraclass correlation coefficient (see Bland 1995).

Buckland-Wright J C, Macfarlane D G, Williams S A, Ward R J. Accuracy and precision of joint space width measurements in standard and macroradiographs of osteoarthritic knees. Ann Rheum Dis 1995; 54: 872-80.

Bland M. An introduction to medical statistics. Oxford Medical Publications, Oxford 1995.

Feng Z, Ziv I, Rho J. The accuracy of computed tomography-based linear measurments of human femora and titanium stem. Invest Radiol 1996; 31: 333-7.

ISO. Accuracy (trueness and precision) of measurement. International Standard ISO 5725-1:1994. Switzerland 1998.

Malchau H, Kärrholm J, Wang Y X, Herberts P. Accuracy of migration analysis in hip arthroplasty. Acta Orthop Scand 1995; 66: 418-24.

Manadan P, Rankin R, Askew M J, Gradisar I A. Accuracy assessment of a technique for contact point determination from planar radiographs. Biomed Sci Instrum 1997; 33: 321-5.

Appendix 1

International Standard and definitions (ISO 1998)

Accuracy – The closeness of agreement between a test result and the accepted reference (the 'true') value.

Trueness – The closeness of agreement between the average value obtained from a large series of test results and an accepted reference ('the true') value.

Precision – The closeness of agreement between independent test results obtained under stipulated conditions.

Repeatability – Precision under repeatability conditions.

Repeatability conditions – Conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment, within a short interval of time.

Repeatability limit – The value less than or equal to which the absolute difference between two test results obtained under repeatability conditions may be expected to have a probability of 95%.

Reproducibility – Precision under reproducibility conditions.

Reproducibility conditions – Conditions where test results are obtained by the same method on identical test items in different laboratories with different operators, using different equipment.

Reproducibility limit – The value less than or equal to which the absolute difference between two test results obtained under reproducibility conditions may be expected to have a probability of 95%.

Appendix 2

Formulas for assessing accuracy and repeatability

Instrument with only random errors

Repeatability, assuming no bias: $1.96 \times \frac{\sum d^2}{n}$

Accuracy, assuming no bias: $1.96 \times \frac{\sum d^2}{2n}$

Instrument with both systematic and random errors

Interval estimate of bias: $d_0 \pm 1.96 \times \frac{s_{d_0}}{\sqrt{n}}$

where $s_{d_0} = \sqrt{\frac{\sum (d_0 - d_0)^2}{n-1}}$

Repeatability: $1.96 \times \sqrt{2} \times \sqrt{\frac{\sum (d_0 - d_0)^2}{n-1}}$

d = paired difference between repeated measurements of the same object.

d_0 = measured deviation from a true value.

n = number of (pairs of) measurements.