IMPROVING THE RELIABILITY OF ORTHOPAEDIC MEASUREMENTS

JAMES G. WRIGHT, ALVAN R. FEINSTEIN

From Yale University School of Medicine, Connecticut

The three purposes of this article are: 1) to describe the three sources of measurement variability, 2) to discuss methods of evaluating measurement variability and, 3) to suggest techniques for improving the reliability of clinical measurements in orthopaedics.

In making clinical decisions, orthopaedic surgeons rely heavily on measurements obtained from the patient’s history, physical signs and radiological examination. A measurement is regarded as reliable, or reproducible, if the same result is obtained when the same entity is measured again.

The importance of measurement reliability can be illustrated with an example of a child with scoliosis. An incremental difference of 5° between two measurements of the Cobb angle is generally believed to represent a deterioration that may require spinal fusion, but two recent studies have shown that different observers may vary up to 10° in the measurement of the Cobb angle (Carman, Browne and Birch 1990; Morrissy et al 1990). Therefore, true progression of a scoliotic curve is proven only when the difference between two measurements exceeds this limit. Any change less than 10° may simply arise from measurement variability and may possibly lead to an unnecessary operation.

The problem creates a scientific challenge for orthopaedic surgeons in documenting, and if necessary, reducing the variability of orthopaedic measurements.

Measurement variability is present when different results are obtained on repeated measurement of the same entity. Although laboratory measurements are generally regarded as trustworthy because of well-established methods and standards for quality control, much less attention has been directed towards measurements performed by clinicians (Feinstein 1987).

In addition to reliability, validity (or accuracy) is an important quality of scientific measurement. A measurement is regarded as valid or accurate (Healy 1989) if the attribute of interest is correctly measured. Reliability is a necessary but not sufficient condition for validity (Feinstein 1977; Nunnally 1978). For example, a scale that consistently measures 5 kg higher than its true value would be reliable but not valid. A measurement that is completely unreliable, however, cannot be valid. For instance, if we measured the degree of scoliosis in a patient from a metre length radiograph, and obtained a dramatically different result when the radiograph was repeated a day later when no clinical change had occurred, we would conclude that the measurement technique was neither reliable nor valid. Validity is discussed extensively elsewhere (Nunnally 1978; Feinstein 1987); the rest of this article is concerned exclusively with reliability.

The term ‘observer variability’, often used to describe any study of measurement variability, may be misleading because variation in measurement can arise from sources not involving an observer, such as a change of position of a patient during a radiological examination. To avoid ambiguity in this essay, measurement variability will encompass all sources of variability; variability arising from an observer will be called clinician variability and the term observer variability will not be used.

SOURCE OF MEASUREMENT INCONSISTENCY

An unreliable measurement can arise from three sources of variability: the patient, the procedure, and the clinician (Sackett et al 1980; Feinstein 1987). The measurement process is divided into these component parts because each component may have to be considered separately to reduce measurement variability. Although an investiga-
tor may want to focus on only one source of variability, a complete appraisal requires consideration of all three sources.

**Patient variability.** The first source of measurement inconsistency arises biologically when the phenomenon being measured varies with the patient's particular state or condition; an acceptable 'normal' resting heart rate has wide limits. Patients may also vary when examined in different conditions such as fasting, drunk, cooperative, or even in differing physical positions; for example, in patients with spondylolisthesis, the actual slip of the fifth lumbar vertebrae on the sacrum may increase by up to 20% between the standing and supine position (Lowe et al 1976).

*Evaluating patient variability.* To evaluate the biological variability of patients, suitable definitions are required for the patient study group, for the spectrum of disease within that group, and for the testing conditions (Feinstein and Kramer 1980; Haynes 1980; Feinstein 1985, 1987).

1) **Patient group.** The group of investigated patients should consist of people for whom the measurement is intended. That is, the measurement variability found for knee laxity in a group of normal healthy college students would not apply to a group of patients with chronic knee ligament tears.

2) **Spectrum of disease.** The spectrum of disease in the study group should be sufficiently broad, ranging from mild to severely afflicted patients. For instance, when the measurement of an anatomical axis is evaluated with metre-length standing radiographs, the investigated group should include patients with large angular deformities. If the study group is confined to patients with straight limbs, the variability of measurement will be underestimated.

3) **Testing conditions.** Patients should be examined in test conditions that match the clinical setting.

*Reduction of patient variability.* Patient variability in clinical practice may be reduced by examining patients under comfortable circumstances, and by using consistent, standardised test conditions. For example, lateral radiographs of patients with spondylolisthesis should always be taken in the standing position (Lowe et al 1976).

**Procedural variability.** The second source of measurement inconsistency is the procedure itself. The term procedure is here used exclusively for a technical examination of the patient, such as a biopsy or a scanogram, rather than the clinician's unaided observation. Measurement variability during the procedure may arise either from the technician's performance or from his equipment. For instance, procedural variability of metre-length radiographs could be introduced by inconsistencies in both the technician's positioning of the patient, or by erratic radiographic equipment. The guidelines used to perform the procedure, to be discussed subsequently, are called the procedural criteria.

As shown in Figure 1 (broken line from patient to clinician), a procedure and its variability are absent when the clinician examines the patient directly. For example, limb alignment may be determined by examining a radiograph of the patient's lower limbs or by examining the patient's legs directly. If a clinician measures limb alignment from a radiograph, both the clinician and the intervening radiographic procedure are sources of variability; if the clinician examines the patient directly, the procedure and its variability do not apply.

![Fig. 1](attachment:figure1.png)

**Sources of measurement variability.**

Variability is rarely documented for radiographic procedures because repeated radiographs cannot ethically be taken on the same patient at the same visit. Therefore measurement variability for radiographic measurements, such as the Cobb angle, is usually ascribed only to clinicians observing radiographs. Consequently, measurement variability is underestimated because both patient and the procedural source have been excluded. Several techniques have been used to circumvent this difficulty. The variability of measurement has been documented for limb alignment from Styrofoam models of lower limbs (Floyd 1988), and from amputated specimens (Wright, Treble and Feinstein 1991). Measurement variability can also be studied on occasions when multiple films of the same patient, taken for another purpose, are available. Repeated films from intravenous pyelogram studies on the same day were used to examine the variability of measurement for radiological features of dislocated hips (Broughton et al 1989).

*Evaluating procedural variability.* Procedural variability should be evaluated in a setting that reproduces, as closely as possible, the measurement procedure to which the patient will be subjected. For example, a single highly-motivated radiographer in a research study using high quality equipment is likely to have much less variability than multiple radiographers in a service setting.

*Reducing procedural variability.* To reduce variability, the operational methods should be described in enough detail to allow replication of the technique for repeated measurements. Whenever possible, similar or identical
equipment should be used, that is, positional jigs designed to standardise limb positions (Wevers, Siu and Cooke 1982).

**Clinician variability.** The third source of inconsistency arises from clinicians observing, extracting, and interpreting information. As shown in Figure 1, a clinician’s observational activities can be directed towards the product of a previous procedure, such as a radiograph or biopsy specimen, or toward examining the patient directly. Observational activities have two components: the first step is the observational process with which the clinician generates raw data; the second step is the clinician’s conversion of the raw data into categories.

Measurement variability may arise in the observational process from the clinician or from measurement devices used in this process. For example, unreliable measurements of the Cobb angle may come from the clinician’s inconsistent choice of the end-vertebrae, or from variable markings on protractors used to measure the angle (Morrissy et al 1990). The guidelines for the performance of an observational process are called process criteria. For the measurement of limb alignment, process criteria would include instructions on how to (a) choose the landmarks, (b) draw the lines, (c) use the protractor, and (d) report the angle (Moreland, Bassett and Hanker 1987).

The second phase of observational activity occurs when the raw data are converted into categories. The criteria used to make this conversion have been called the conversion criteria (Feinstein 1985). Conversion criteria are used when the raw data of anatomical limb alignment may be categorised as: normal alignment for 5° to 7° of valgus, genu varum for less than 5° of valgus, and genu valgum for more than 7°. In other measurements, clinicians are sometimes unaware that they use conversion criteria. Clinicians often distinguish ‘sick’ from ‘not-so-sick’ patients but the criteria for this categorisation may not appear in any publication. The categories are based on a combination of the patient’s physical appearance, the way the patient responds to questions and the results of investigations. Each clinician may have distinctively individual conversion criteria.

Special terms are in common use to describe the situations when clinicians do not agree. In measurements of the same phenomenon, intra-clinician variability occurs when the same person obtains different repeated results, and inter-clinician variability occurs when two people disagree.

**Evaluating clinician variability.** When tested, the variability of clinicians should be checked by determining whether they are ‘blind’, and whether they are ‘practised’ or ‘unpractised’ (Haynes 1980; Feinstein 1985).

‘Blind’ clinicians have no knowledge of prior measurements, by themselves or other clinicians, nor preferably, of the clinical status of the patient. Knowledge of any of these features can create bias in the observations.

If the clinicians are very experienced or practised, their measurement variability will not accurately reflect the usual variability of inexperienced or unpractised clinicians.

**Reduction of clinician variability.** Clinician variability may be reduced by clarifying both the observational process and conversion criteria (Feinstein and Kramer 1980; Haynes 1980; Feinstein 1985; Morrissy et al 1990).

1) Process criteria. When unreliability is encountered, variability may be reduced by a meeting between the clinicians, involving a frank, open discussion of their process and conversion criteria. Inconsistency during the observational process may be reduced both by consistent choice of landmarks, such as end-vertebrae for the Cobb angle, or by using the same equipment. The latter procedure in one study would have avoided inconsistently manufactured protractors (Morrissy et al 1990).

Measurements also vary less when performed under ideal circumstances, such as a well-rested clinician (Haynes 1980). Measurement skills improve with practice, and more precise measurements are obtained with repeated assessments (Fleiss 1986). Finally, Morrissy et al 1990 noted that intra-clinician variability is generally smaller than inter-clinician variability.

2) Conversion criteria. Important issues of variability arise from the number of categories in the scale used for expressing ratings, or from the specifications of conversion criteria. An increased number of categories in a rating scale allow a greater number of clinical distinctions, but will usually increase the measurement inconsistency. The ideal number of categories for a scale cannot be defined, but increased clinical discrimination may be offset by greater measurement variability (Feinstein 1987). For example, in classifying Perthes’ disease, the Salter–Thompson rating scale, with only two categories, has less observer variability than the four-category rating scale of Catterall (Simmons, Graham and Szalai 1990). The Catterall scale, however, may allow greater prognostic stratification than is possible with only two categories.

Explicit conversion criteria may reduce measurement variability. For instance, ‘limp’ can be graded slight, moderate or severe, but the absence of criteria for these ratings may lead to substantial variability. A different problem occurs when the conversion criteria are cited but are ambiguous, incomplete, or contradictory (Feinstein 1979). Walking ability can be divided into bed-to-chair, indoors only, two to three blocks, six blocks, or unlimited; with no clear guide for how to categorise a patient who can walk between three and six blocks (Harris 1969) or even how long a ‘block’ is.

**THE IMPORTANCE OF CONSISTENCY**

There are three reasons why clinical measurements should be reliable: an interval change in a reliable clinical measurement can be confidently attributed to a true change in clinical status of the patient; reliable measurements can improve the efficiency of clinical trials; and
reliability is a minimal requirement for valid and cogent clinical outcomes.

**True clinical change or measurement variability?** For a difference between two measurements of the same patient at different times to be attributable to a true change in the patient's condition, it must be greater than the simple variability of measurement. Reliable measurements, therefore, allow clinicians to be increasingly confident in attributing a difference between two measurements to a change in the clinical status of the patient (Cowell 1990).

**Efficiency of clinical trials.** Comparing a new with an old treatment, measurement inconsistency will obscure a true difference in effect. The solution to this dilemma is usually to enlarge the number of people in each treatment group of a clinical trial. Unreliable measurements therefore decrease the efficiency of clinical trials by requiring more patients to demonstrate a statistically significant difference between two treatments (Fleiss 1986).

**Reliable clinical outcomes: subjective or objective data?** The common belief that a patient's subjective complaints are 'soft' but that objective data are 'hard', implies that subjective data are less reliable and are always inferior. Attention to objective measurements only, however, may create two problems.

First, the patients' perceived pain and disability are the usual orthopaedic indications for most elective operations. If a patient's subjective complaints were replaced by objective data, often the most cogent and relevant end-points would be ignored. Secondly, objective measurements are not always 'hard'. For example, despite the objectively measured output, the KT1000 arthrometer, used to quantify knee ligament laxity, gives great variation in the measurement of subluxation (Forster, Warren-Smith and Tew 1989). The crucial attribute of 'hardness' in any kind of measurement is reliability (Feinstein 1977). With suitable attention to procedural and conversion criteria (Feinstein 1987), many 'soft' measurements can be 'hardened'.

In conclusion, measurement of patient outcomes after orthopaedic treatments should be cogent, relevant, and reliable. If a subjective measurement, such as pain or patient disability, is the outcome of interest for a clinical trial but is shown to be unreliable, two options are available. The first is to reject the measurement in favour of a more reliable but perhaps less cogent end-point: the second is to improve the 'hardness' of the subjective measurement by investigating the source of variability and correcting it, using the principles outlined in the previous section.

**STATISTICAL ANALYSIS**

In this last section we consider a group of statistical indices, called tests of concordance, that are used to quantify measurement reliability. Measurements can be expressed in several types of scales, with categories that are dichotomous, nominal, ordinal, or dimensional, also called continuous (Nunnally 1978; Feinstein 1985, 1987). Binary scales have two mutually exclusive or dichotomous categories, such as male or female, yes or no. Nominal scales, such as diagnostic labels or religious preference, contain categories with no inherent rank or order. Ordinal scales, such as grading knee ligament instability from 1 to 3, have directionally ordered ranks, but the differences between ranked categories are not necessarily equal. Dimensional data, such as deformity on a radiograph or range of motion, have ordered ranks and the increments between categories have measurably equal magnitude.

The most appropriate index of concordance depends on the type of rating scale. Concordance for binary data can be expressed with percentage agreement, the kappa statistic (Landis and Koch 1977; Kramer and Feinstein 1981; Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990), or the McNemar chi-square (Fleiss 1981). The percentage agreement, kappa statistic, and the Stuart–Maxwell statistic (Fleiss 1981) are indices of concordance for nominal data. Ordinal data can be compared with a simple percentage agreement and kappa statistics, or more appropriately, with a weighted kappa statistic (Kramer and Feinstein 1981). The conventional correlation coefficient, which measures the interdependency of two variables, is an index of trend or relatedness, but is unsatisfactory for quantifying agreement. If two observers measured flexion contracture of the hip and one observer consistently obtained a measurement 20° higher than the other, the results would be perfectly correlated but the two observers never agree. The appropriate concordance statistic for dimensional data is the intra-class correlation coefficient (Bartko 1966; Cicchetti and Sparrow 1981; Kramer and Feinstein 1981; Bland and Altman 1986; Fleiss 1986).

**Conclusions.** When investigators study reliability of measurement, all three possible sources of measurement variability should be considered. The patients should be chosen as a population for which the measurement is intended and the conditions of measurement should mimic the clinical situations as closely as possible. If measurement reliability is found to be unsatisfactory, the investigator should be prepared to discover the source or sources of unreliability and correct them.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

REFERENCES


THE JOURNAL OF BONE AND JOINT SURGERY


Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.


