On the art of measuring

In this issue of *Acta Orthopedica Scandinavica* there is an article which focuses on the measuring process itself. Ryd and co-workers ("Knee scoring systems in gonarthrosis. Evaluation of interobserver variability and the envelope of bias", pp 41–45) conclude that many kinds of measuring instruments in use in clinical practice, despite a reasonable reproducibility, are not satisfactory but lend themselves to systematic errors. The article directs attention to the fact that the clinical measuring instruments used in orthopedics, and in most other disciplines of medicine, rest on interpretations by the assessors. It is a truism to state that the registration made by any measuring instrument is formed by a combination of the instrument itself and the investigator, who handles the instrument. This fact needs, however, always to be acknowledged.

Comparisons of different study results necessitate the use of instruments that really measure what they are intended to measure and that diseases are clearly defined. A striking example is studies of posttraumatic arthrosis of the knee, where different studies report quite different risks after meniscectomy. These differences could be caused by an unclear definition of the disease itself, heterogeneous groups of patients, or that the information regarding associated injuries, time from injury, etc., in many cases is omitted or incomplete. Moreover, many different radiographic classifications may have been used, or a low reproducibility of radiographic readings could be the cause. Unsatisfactory results were presented when the reproducibility of two radiographic shoulder fracture classifications were tested (Sidor et al. 1993, Siebenbrock and Gerber 1993), and it was even questioned whether such classification systems were useful at all (Burstein 1993). The results of the treatment of various conditions in orthopedics, especially concerning fractures, may be misleading because of the lack of consistent classifications (Cowell 1994).

Different scoring scales have been widely used in the research on knee ligament injuries. The Lysholm score was published in 1982 (Lysholm and Gillquist 1982), and it has been very frequently used in its revised form the last 10 years (Tegner and Lysholm 1985), especially in research of anterior cruciate ligament injuries and their treatment. Strikingly, a recent study showed that, among four diagnostic categories, the Lysholm score was least sensitive for the cruciate ligament patients (Bengtsson et al. 1996). It has lately become obvious to many orthopedic researchers that much effort must be made to do research on instruments that are meant to measure the results of our interventions–the research on outcome-instruments (Amadio 1993).

Outcome is something that follows as a result (Webster 1993), and it can be classified and measured differently. Outcomes are traditionally divided into objective and subjective, but can also be classified according the level at which the instrument is measuring: impairment level, disability level or handicap level (WHO 1980). An orthopedic example could concern a patient with an anterior cruciate ligament injury. If the result of an operation or other treatment were to be evaluated at *impairment* level, laxity could be measured (an indicator of the loss of an anatomical structure). If such patients were evaluated with an instrument measuring at *disability* level, the subject's inability to perform activities normal to that person would be considered and questions regarding ability to walk, run, etc., would be asked or else performance tests would be conducted. If, finally, outcome were measured at *handicap* level, we would evaluate whether the given individual was limited or was prevented from fulfilling his normal role, i.e., will the anterior cruciate ligament injury, resulting laxity (impairment) and inability to walk on uneven terrain (disability) prevent or limit the individual from returning to his work, sport or other activity normal to that individual (handicap)?

Systematic error

Apart from conventional characterization of measurement instruments, which includes evaluating the random error by statistical treatment (see below), some specific appraisal of the systematic error is desired. From the report by Ryd and co-workers it appears that clinical instruments are inaccurate and have a low measurement reproducibility (reliability) and therefore are distinctly susceptible to the action of systematic errors. This would be analogous to the following problem: Suppose somebody has a 1 meter long rod, but stated that "this rod is 1.1 meter long". For the observer it would be much easier to uncover the error, if he had a tape measure (with high reproducibility) than if he had to rely on ocular assessment (with low reproducibility). It is sometimes inferred that statisti-

cal treatment protects an investigation against faulty conclusions due to measurement errors. This applies only to random errors and, indeed, the entire statistical theory deals only with random (stochastic) variation. Systematic variation is statistically treated as a part of the signal and must be addressed beforehand by the design of the study. Thus an instrument with low reproducibility will not only give misclassifications and hence low sensitivity and specificity, which can be quantified. It will also allow an unknown and not quantifiable amount of bias to bear on the investigation.

Clinical medicine usually involves the assessment of subjective data, such as pain, well-being, patient satisfaction and the like. These data are often defined as "soft outcomes". Traditionally they have been collected by an observer, at worst by the operating surgeon, or better by a more or less (in)dependent observer. Certainly patients (usually) want to feel better after any particular treatment and doctors (usually), from their own inner needs, want to register an improvement from that treatment (Morris 1993). Thus, assessments are done in the framework of positive bias, which is inherently linked to the clinical situation. Positive results of any treatment in an open and uncontrolled clinical trial without at least independent observers are thus liable to reflect partly, sometimes entirely, this systematic error. The trend in outcome research today is towards measuring subjective outcomes, i.e., at disability and handicap level, since health-care providers, patients, insurers, government agencies, and employers seek to determine whether specific interventions satisfy the needs of patients. To avoid observer-introduced bias, this has increased the use of patient-administered questionnaires. These "soft" outcomes have in numerous studies proved to be at least as reliable, valid and responsive as many measurements considered objective (Bellamy et al. 1988, Levine et al. 1993, Ware et al. 1993).

Basic criteria for outcome instruments

All outcome instruments used for scientific purposes, objective or subjective, have to fulfil basic criteria. They must be standardized, permit quantification, be reliable, valid for the disease and subjects involved and responsive to clinical change (Liang and Jette 1981).

Firstly, it must be determined for what subjects and under what circumstances the instrument is to be used. *Content validity* should be determined by a consensus process, involving health-care providers interested in the outcome and supposed to use the instrument, and, of course, the patients involved. An excellent example of such a process was recently given by Martin et al. (1996).

Secondly, the instrument must be standardized and used in a way that introduces minimal bias. Outcome measures administered by an observer will introduce observer bias, however, an independent observer will minimize the bias. This is true both if the instrument is considered objective (measuring laxity, interpreting radiograms) or subjective (asking questions from a questionnaire).

Thirdly, the instrument must be tested for reliability, construct and criterion validity, and responsiveness to clinically important changes. This process involves clinical studies and statistical analyses.

Traditionally, *reliability* is assessed by test-retest procedures and assessment of intra- and interobserver reliability.

Construct validity focuses on the extent to which a measure performs in accordance with expectations (Carmines and Zeller 1979). Does the laxity tester measure what it is meant to measure, loss of anterior cruciate ligament, or does it measure something else? Do the questions asked in a questionnaire regarding physical function reflect the desired construct and physical function, or do they reflect some other construct, e.g., socioemotional function? *Criterion validity* has two components, concurrent and predictive. Both rely on the ultimate presence of some irrefutable standard of truth, i.e., a "gold standard". The greatest restriction on this form of validity-testing is the general absence of "gold standards".

Responsiveness or sensitivity to change is a critical attribute of an outcome measure. A measuring instrument on which the scores change only a little when tracking the change in a patient who has changed drastically, is termed "poorly responsive". Conversely, a measuring instrument on which the true scores show some change even when the change actually occurring is very small is termed "highly responsive". A clinical study using a highly responsive outcome measure needs fewer patients to yield statistically significant changes, making clinical studies more manageable. See also "effect size" (Ryd et al. p 42 in this issue).

Only a few instruments used in orthopedic research have been tested to meet these criteria.

Suggestions

The way to improvement lies in meticulous study designs when validated instruments are to be used. Ideally, one would select a single measure integrating

every relevant aspect of the disease (i.e., having content validity) and being extremely responsive to clinical change (i.e., requiring few study patients). In reality, one is often forced to compromise and trade off one consideration against another in order to stay within the limits of patient availability, logistic support, and budgetary constraints. In order to allow cross-study comparisons and improve orthopedic research, outcome measures should be used that have proven reliability, validity and responsiveness. Already existing tools should be tested for these important clinimetric properties and if no instrument exists which is suitable for the intended purpose, efforts should be made to develop and test such instruments.

Finally, prospectivity is a baseline requirement. Validation against known constructs and criteria, randomization, blinding of independent observers and the use of patient-administered assessments for subjective outcomes should constitute the normal design.

Harald Roos, Ewa Roos, Leif Ryd
Department of Orthopedics, Lund University Hospital, S-221 85 Lund, Sweden

References

Amadio P C. Editorial: Outcomes measurements. J Bone Joint Surg (Am) 1993; 75: 1583-4.

Bellamy N, Watson Buchanan W, Goldsmith C H, Campbell J, Stitt L W. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or the knee. J Rheumatol 1988; 15: 1833-40.

Bengtsson J, Möllborg J, Werner S. A study for testing the sensitivity and reliability of the Lysholm knee scoring scale. Knee Surg, Sports Traumatol, Arthroscopy 1996; 4: 27-31.

Burstein A H. Editorial: Fracture classification systems: do they work and are they useful? J Bone Joint Surg (Am) 1993; 75: 1743-4.

Carmines E G, Zeller R A. Reliability and validity assessment. SAGE Publications, Beverly Hills, London 1979.

Cowell H R. Editorial: Patient care and scientific freedom. J Bone Joint Surg (Am) 1994; 76: 640-1.

Levine D W, Simmons B P, Koris M J. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Joint Surg 1993; 100 (100): 1585-92.

Liang M, Jette A M. Measuring functional ability in chronic arthritis. Arthritis Rheuma 1981; 24 (1): 80-6.

Lysholm J, Gillquist J. Evaluation of knee surgery results with special emphasis on use of a scoring scale. Am J Sports Med 1982; 10: 150-4.

Martin D P, Engelberg R, Agel J, Snapp D, Swiontkowski M F. Development of a musculoskeletal extremity health status instrument: the musculoskeletal function assessment instrument. J Orthop Res 1996; 14: 173-81.

Morris R W. Comparative evaluation of outcome of knee replacement operations using alternative prostheses. Thesis, University of London, UK 1993.

Sidor M L, Zuckerman J D, Lyon T, Koval K, Coumo F, Schoenberg N. The Neer classification system for proximal humerus fractures. J Bone Joint Surg (Am) 1993; 75: 1745-50.

Siebenrock K A, Gerber C. The reproducibility of the classification of the fractures of the proximal end of the humerus. J Bone Joint Surg (Am) 1993; 75: 1751-5.

Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. Clin Orthop 1985; 198: 43-9.

Ware J, Snow K, Kosinski M, Gandek B. SF-36 health survey manual and interpretation guide. New England Medical Center, The Health Institute, Boston, MA 1993.

Webster's New encyclopedic dictionary. Black Dog and Leventhal 1993.

WHO. International classification of impairments, disabilities, and handicaps. World Health Organization, Geneva 1980.