Knee scoring systems in gonarthrosis

Evaluation of interobserver variability and the envelope of bias

Leif Ryd[1], Johan Kärrholm[2], Peter Ahlvin[3] and the Score Assessment Group

10 experienced orthopedic surgeons assessed 15 patients using 3 commonly used composite scoring systems and by some simple variables to evaluate knee replacements. Statistical evaluation showed that the scores were valid and reflected the disease process with a reasonable reproducibility. In the individual case, however, considerable changes of the total scores and the simple variables are needed to represent a true difference at the 95% confidence limit. The coefficient of repeatability varied from 45 to 72 for the scores.

Our study, which is suggested to represent any clinical investigation, showed that clinical measurements are not robust and meticulous efforts in terms of study design must be made to protect an investigation against the action of bias. Knee scores are exceedingly unreliable.

Departments of Orthopedics, [1]University Hospital, S-221 85 Lund, [2]Sahlgrens Hospital, Gothenburg and [3]Visby, Sweden.
Tel +46 46-17 15 10. Fax-13 07 32
Submitted 95-10-21. Accepted 96-10-10

Orthopedics, like all other natural sciences, use measurements to gain insights into the laws of nature. This places high demands on the measuring systems we use. They should be valid, reproducible, sensitive and specific enough to yield relevant results; the signal-to-noise ratio must be such that the phenomena studied do not drown in the bias inevitably introduced by the clinical situation (Ryd et al. 1994). We analyzed the quality and robustness of a number of scores and simple clinical measurements commonly used in orthopedic practice to evaluate the efficacy of total knee arthroplasty.

Materials and methods

At the 50th meeting of the Swedish Orthopedic Association 10 experienced orthopedic surgeons examined 15 patients suffering from gonarthrosis. 7 patients had been operated on with a total knee replacement about 1 year before the investigation, while 8 were on the waiting list. There were 8 right knees and 7 left index knees involved. The mean age of the patients was 74 (60–86) years. 9 patients had unilateral knee disease (Charnley stage A), 3 had bilateral knee disease (Stage B) while the remaining 3 had involvement of other parts of the locomotor system (Stage C). The assessment systems investigated were the Hospital for Special Surgery (HSS)-score (Ranawat et al. 1976), the Venn diagram scoring system (Jónsson 1981) and the Knee Society score (KSS) (Ewald 1989). The KSS involves one part which assesses the function of the knee and one part assessing the overall patient function. These were compiled by each investigator for each patient individually. The value of the HSS-score was also categorized as excellent (100–85), good (84–70), fair (69–60) and poor (<60) as suggested (Ranawat et al. 1976). Also the individual variables included in these scoring systems were investigated. Varus-valgus instability, sagittal instability, flexion and extension and alignment were measured using goniometers. Walking distances were recorded as subjectively estimated by the patients. Because pain is assessed differently in the three systems the investigators recorded pain on exertion and pain at rest (ache) on a VAS scale (0–100 mm). Also, the quantities of analgesics taken the week and the month before the investigation were recorded. Finally, an experienced nurse, not acquainted with the patients, assessed the "over all status" of the knee on a VAS scale after having had a free conversation with the patient.

The order by which each investigator examined each patient was recorded to assess possible learning phenomena. All investigators were experienced orthopedic surgeons well acquainted with the assessment of gonarthrotic patients. They had all used or were acquainted with the three different scoring systems.

The validity of the scores was assessed by a comparison to the only truly validated instrument to assess gonarthrotic knees; the Lequesne ISK (Index of Severity-Knee) (Lequesne et al. 1987) which was regarded as the gold standard.

The reliability was assessed by the Intraclass Correlation Coefficient (ICC) (Kramer et al. 1981) for the different variables tested as assessed by the 10 investigators. The validity of the variables was assessed by comparing them with the gold standard, ISK, using the Kendall Tau. This is a non-parametric correlation coefficient. For ordinal data (Venn and the HSS classification) the weighted kappa value was determined (Fleiss et al. 1973, Kramer et al. 1981). The mean and standard deviations (SD) from the 10 investigators obtained for all 15 patients regarding each variable was also compiled using 2-way ANOVA (Conboy et al. 1996). The coefficient of repeatability was computed by multiplying the SD by 2 times the square root of 2 (≈2.83) (Bland et al. 1986). Finally, mean values of these means and SDs were compiled and the coefficient of variation [(mean) SD/(mean) mean] for each parameter was computed to assess the signal-to-noise ratio. Another way to assess the signal-to-noise ratio is by the "effect size" concept (Kazis et al. 1989). Here the results of an intervention is put in relation to the accuracy by which the results can be assessed, by the formula:

(mean score of patients post op – mean score of patients preop) / SD of scores preop

Results

The different scores all correlated significantly with the Lequesne ISK (Table 1). The registrations by the nurse showed one of the best correlations with the ISK. The consumption of analgesics during the last month also reflected the status of the knee while consumption during the last week did not.

The reliability of the scores ranged from very low to very high as shown by the ICC. Many of the individual variables exhibited a good reliability while some, however, showed a ICC as low as 0.4. For the scores, the standard deviation of the individual assessments ranged from 16 (HSS) to 26 (KSS knee) and the coefficient of repeatability was as high as ≈ 72. The coefficients of variation for the scores ranged from 0.22 (HSS) to 0.41 (KSS knee) (Table 2).

For the individual variables, the largest standard deviation was found for the VAS assessments of pain. There was, however, a distinct difference between patients with little pain and with much pain; little pain seemed to be registered with a higher degree of consistency than did much pain (Table 3).

Computation of effect sizes showed that TKA is a worth-while procedure. The ranges, however, were large, for some of the investigators the effect size was as low as 0.3 (Table 4).

By comparing the ICC for the 5 examinations performed first on each patient with the five performed last learning effects were analyzed. No such effects were seen for any of the investigated variables except for walking distances.

Table 1. The relationship between the Lequesne ISK score and the means of the values for each patient regarding the over-all status of the knee as tested in this study

Comparison	Kendall Tau	P -value
ISK vs. HSS	0.65	0.0007
ISK vs. Venn	0.62	0.001
ISK vs.KSS patient	0.46	0.02
ISK vs.KSS knee	0.57	0.003
ISK vs. nurse	0.64	0.0008
ISK vs. analgesics month	0.47	0.02
ISK vs. analgesics week	0.30	0.12

Table 2. The consistency of some parameters reflecting the status of the knee

Parameter	ICC (r_i)	Weighted Kappa	SD	Coeff. of repeatability	Mean (mean)	Coeff. of variation (CV)
Varus-valgus instability	0.39		5.2	14.7	5.1	1.01 [a]
Sagittal instability	0.44		15.7	44.4	2.9	5.4 [a]
Maximum flexion	0.77		13.9	39.3	118.4	0.18 [a]
Extension defect	0.43		4.7	13.3	7.6	0.62 [a]
Alignment	0.74		9.3	26.3	180.6	0.05 [a]
VAS pain at rest	0.93		25.0	70.8	14.0	1.78
VAS-pain on exertion	0.88		29.8	84.3	31.6	0.94
Venn		0.84				
HSS	0.24		15.8	44.7	71.7	0.22
HSS-classification		0.78				
KSS knee	0.83		25.6	72.4	62.2	0.41
KSS patient	0.71		20.0	56.3	58.8	0.34
Maximum walking distance	0.97		1 101	–	973	1.13
Painfree walking distance	0.91		697	–	551	1.26

[a] Indicates coefficients of correlations skewed by small or big mean values.

Table 3. Individual VAS pain (ache) registrations and mean (SD) for case 2 (successfully operated) and for case 10 (on the waiting list for kne arthroplasty) by ten experienced orthopedic surgeons

	Case 2	Case 10
	0	48
	0	48
	0	44
	7	9
	0	59
	0	27
	0	27
	0	6
	2	40
	0	69
Mean (SD)	0.9 (2.2)	38 (20)

Table 4. Mean (range) effect size as computed for 10 assessors

	Effect size, mean	Range
VAS pain at rest	0.8	(0.6–1.4)
VAS-pain on exertion	2.2	(1.2–4.6)
HSS	1.5	(0.9–2.6)
KSS knee	2.4	(0.7–4.7)
KSS patient	0.7	(0.3–1.5)

Discussion

Fundamentally, any measuring instrument should be evaluated regarding its ability to yield reproducible results, which, in turn, should reflect the changing course of events, such as symptoms of a disease or the outcome of its treatment (Wright et al. 1992). A number of scores have been devised to register the outcome of joint replacement procedures. Few of them have been properly validated and they have been found to give inconsistent results (Andersson 1972, Jónsson 1981). We assessed the 3 scores most frequently used in Sweden and internationally.

Our study shows that the scores are reasonably valid, i.e., they reflect the same changes of the status of the underlying disease; the agreement of the mean values with the only validated score for knee arthrosis, the Lequesne ISK, was acceptable. In this context, concordance values of between 0.6 and 0.8 have been suggested to reflect "substantial agreement" for kappa statistics (Landis et al. 1977), and may give some guidance also here. It is, however, sobering that an informal conversation with a nurse is equally valid. Furthermore, the finding that the consumption of analgesics during the last month also reflects the status of the knee is interesting. Counts of tablets can in some respects perhaps be considered as "hard data".

A number of ways to assess the reliability, repeatability or reproducibility of the instruments have been used here. Simple correlation to assess conformity of data has the disadvantage of disregarding observer bias; one observer may consistently assign higher values. If so high correlation coefficients can be reached despite poor agreement between the two observers. This obstacle is overcome by ICC, which neglects the individual assessments and focuses on the agreement between assessors. The weighted kappa used for ordinal data has the advantage of compensating for the agreement occurring purely by chance (Kramer et al. 1981). These data show the VAS measurements and some of the scores to be reliable while the goniometric measurements were not. The learning effects for the walking distance was reflected by a very high reliability; once estimated by the patients, these values were easy to remember. None of these figures, however, offer a simple way by which one can assess if two measurements, on different occasions, reflect a true "difference". A simple way to arrive at this information is to compute the coefficient of repeatability. This value indicates the magnitude by which one measure must differ from a previous one (by another assessor) to represent a significant change (at the 95% limit); a most useful figure.

The coefficient of variation is derived by dividing the (mean) SD by the (mean) mean. This can be looked upon as a way to relate the strength of the signal to the magnitude of the noise involved. For some variables the mean is very small or very large resulting in numbers that are hard to interpret. For a number of variables, though, the means were similar thus allowing comparisons as indicated in Table 2. The effect size concept, finally, is an interesting way to address the problem that while some effects of an intervention may be statistically significant the magnitude of the improvement may be so small as to be clinically irrelevant. An intervention is effective only if both statistical significance and a large effect size is recorded. In this context an effect size of less than 0.2 is small, 0.5 is moderate and above 0.8 is large (Cohen 1977).

It is known that also simple measurements like range of motion show SDs of 5–10 degrees when performed by different assessors on the same subject (Solgaard et al. 1986, Resch et al. 1995). The standard deviations of the individual variables in this study were of the same range. The composite scores are liable to combine the variations of all the individual variables included in them, thus making the scores

even more uncertain. The results of our study support this notion; the composite scores showed larger standard deviations than the individual variables. The finding that the effect size of TKA as measured by, for example, the KSS patient score, for some assessors, was very small points in the same direction. In the light that TKA is regarded as one of the most successful orthopedic interventions ever invented, the inability to register a significant improvement lies in inadequate measuring instruments. Finally, any score with a coefficient of repeatability of 75 out of 100 is not reliable enough for practical use.

The mean values found in our study may be regarded as reflections of the signal to measure and the standard deviations, the coefficients of variations and coefficients of repeatability represent the noise in the system. Statistically, the noise usually represents the random (stochastic) variation while the third constituent of any measurement, the systematic (bias) variation, is regarded as a part of the signal. Clinically, one is interested in the true signal and a more appealing approach is to let the noise represent any distortion of the signal, systematic or random. Moreover, it appears reasonable to assume that the less robust an instrument is, i.e., the larger the noise data, the more leeway, or latitude, is given for systematic distortion. The results show that the scores need to differ by about 30 (HSS) to 70 (KSS patient) points to reflect a true difference at the 95% confidence limit. Despite the nominally high reliability of many of our variables, the latitude for the action of systematic errors (bias) is thus considerable. This represents the envelop of bias.

The argument that the recruitment of a large number of patients tends to even out the noise may be fallacious. Materials are compiled of individual cases where the action of bias usually acts in the same direction for every patient. The expectations by the patients and the surgeons surely act strongly in the same (positive) direction, thus possibly skewing the means of the entire material to the same extent as for the individual patient (Ryd et al. 1994). Incidentally, the improvement observed after an arthroplasty in the individual patient often is in the range of 30–40 points (Insall et al. 1983, Nilsson et al. 1993). Thus the noise-to-signal ratio is unfavorable.

The commonest reason for any patient to seek orthopedic advice is probably pain. While it appears that the registration of little pain can be performed with a reasonable consistency, the individual registrations of case 10, with more pain, lend themselves to some contemplation; it is disquieting that 10 experienced orthopedic surgeons assess one and the same patient so differently (Table 3). For comparison, coefficients of variation in the engineering world are orders of magnitudes smaller. When assessing the measuring accuracy of a digitizing table for RSA a coefficient of variation of 0.0015 was found (Önsten 1994).

In conclusion, our study has shown that the scores, by correlation analyses, are reasonably valid and reliable. In depth analysis, however, show that the magnitude of the noise is conspicuous. We suggest that this noise allows also considerable amounts of systematic distortion (bias) to be misinterpreted as a true signal. The remedy may lie in using controlled and randomized designs in any clinical research project as well as independent observers. Moreover, the studies could include more objective evaluations like scintimetry, DEXA, RSA, etc., systems that are readily accessible in orthopedics today. Also, in this context, this study suggests that simple long term counts of analgesic consumption could be feasible. The use of patient administered questionnaires enables the patient to answer standard questions and to score, for example, pain without the interpretation of an assessor with his own biases (Johanson et al. 1992, Levine et al. 1993).

Acknowledgements

The authors are grateful to Dr. Jonas Ranstam for statistical suggestions and fruitful discussions. Financial support for the study was given by Alfred Österlund Foundation, Konung Gustav V's Jubileumsfond, Greta och Johan Kocks Foundation, Stiftelsen Bistånd åt Vanföra i Skåne, the Medical Research Council (09509) and the Medical Faculty of Lund University.

The Score Assessment Group consisted of Peter Ahlvin, Maria Hilding, Ingemar Ivarsson, Johan Kärrholm, Arne Lundberg, Hans Mallmin, Kjell-Gunnar Nilsson, Leif Ryd, Anders Wykman and Ingemar Önsten.

References

Andersson G. Hip assessment: A comparison of nine different methods. J Bone Joint Surg (Br) 1972; 54: 621-5.

Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinicla measuremnt. Lancet 1986; (February 8): 307-10.

Cohen J. Statistical power analysis for the behavioral sciences. Academic Press, New York 1977.

Conboy V B, Morris R W, Kiss J, Carr A J. An evalusation of the Constant-Murley shoulder assessment. J Bone Joint Surg (Br) 1996; 78 (2): 229-32.

Ewald F C. The knee society total knee arthroplasty roentgenographic evaluation and scoring system. Clin Orthop 1989; 248: 9-12.

Fleiss J L, Cohen J. The equivalent of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Physiol Measurement 1973; 33: 613-9.

Insall J N, Hood R W, Flawn L B, Sullivan D B. The total condylar proshtesis in gonarthrosis. A 5 to 9 year follow-up of the first 100 consecutive knee replacement. J Bone Joint Surg (Am) 1983; 65: 619-28.

Johanson N A, Charlson M E, Szatrowski T P, Ranawat C S. A self-administred hip-rating questionnaire for teh assessment of outcome after total hip replacement. J Bone Joint Surg (Am) 1992; 74 (4): 587-97.

Jónsson G T. Compartmental arthroplasty for gonarthrosis. Acta Orthop Scand (Suppl 193) 1981; 52: 3-109.

Kazis L E, Anderson J J, Meenan R F. Effect sizes for interpretating changes in health status. Med Care (Suppl 3) 1989; 27: 178-89.

Kramer M S, Feinstein A R. Clinical biostatistics. The biostatistics of concordance. Clin Pharnacol Ther 1981; 29 (1): 111-23.

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

Lequesne M G, Mery C, Samson M, Gerard P. Indexes of severity for osteoarthritis of the hip and knee. Validation–value in comparison with other assessment tests. Scand J Rheumatol (Suppl 65) 1987: 85-9.

Levine D W, Simmons B P, Koris M J, Daltroy L H, Hohl G G, Fossel A H, Katz J N. A self-administred questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel surgery. J Bone Joint Surg (Am) 1993; 75 (11): 1585-92.

Nilsson K G, Kärrholm J. Increased varus-valgus tilting of screw fixated knee prostheses. J Arthroplasty 1993; 8 (5): 529-40.

Ranawat C S, Insall J N. Duocondylar knee arthroplasty. Clin Orthop 1976; 120: 76-92.

Resch S, Ryd L, Stenström A, Johnsson K, Reynisson K. Measuring hallux valgus. A comparison of conventional radiography and clinical parameters with regard to measurements accuracy. Foot Ankle 1995; 16 (5): 267-70.

Ryd L, Dahlberg L. On bias. Acta Orthop Scand 1994; 65 (5): 499-504.

Solgaard S, Carlsen A, Kramhøft M, Petersen V S. Reproducibility of goniometry of the wrist. Scand J Rehabil Med 1986; 18: 5-7.

Wright J G, Feinstein A R. Improving the reliability of orthopaedic measurements. J Bone Joint Surg (Am) 1992; 74 (2): 287-91.

Önsten I. Fixation of total hip components in rheumatoid arthritis and arthrosis, Thesis, Lund University, Sweden 1994.