The reliability of the American Knee Society Score

Raymond Y L Liow, Karen Walker, Mohammad A Wajid, Gurinder Bedi and Catherine M E Lennox

Department of Trauma and Orthopaedics, Hartlepool General Hospital, Hartlepool, Cleveland TS24 9AH, U.K.
Correspondence: Mr. Raymond Y L Liow, 39 Bradman Drive, Riverside, Chester-le-Street, County Durham, DH3 3QS, U.K. Tel +44 191 389 2496. E-mail: Rliow@aol.com
Submitted 99-09-18. Accepted 00-04-26

ABSTRACT – We have investigated the reliability of the American Knee Society Score (AKS), a widely used functional outcome score for knee arthroplasty. 29 subjects were assessed in random order by 6 observers, each patient seen twice by each observer. The interobserver reference intervals were 16 points for the Knee Score (0–100 points) and 21 points for the Function Score (0–100 points). The intraobserver reference intervals were smaller, 11 points and 13 points for the Knee Score and the Function Score, respectively. The presence of arthritis in other joints and general debility did not affect the reliability of the scores. The more experienced observers had greater intraobserver reproducibility. Notably, we found moderate agreement between observers in the subjective variables, while the objective variables produced lower levels of agreement.

The high inter- and intraobserver variations of the AKS makes estimation of score change questionable. Reliable use of the AKS would necessitate repeated evaluation by an experienced observer.

■

The American Knee Society Score (AKS) was developed by consensus in the Knee Society (Insall et al. 1989). The AKS consists of two parts. The first is the Knee Score which considers pain, stability and range of motion as the main parameters, with deductions for flexion contractures, extension lag and malalignment. A maximum of 100 points will be obtained by a well-aligned knee with no pain, 125 degrees of motion and negligible anteroposterior or mediolateral instability. The second part, the Function Score, utiliszes walking distance and stair climbing as the main parameters, with deduction for the use of a walking aid. The maximum Function Score is 100, given to an individual who is able to walk an unlimited distance and can ascend and descend stairs normally.

A patient categorization system is included in the AKS to identify patients whose function may be undermined by factors other than the knee in question. Category A patients do not have any substantial disease in the contralateral knee, in contrast with category B patients who have substantial arthrosis. Category C patients are those with multiple joint involvement or generalized debility. The AKS is designed so that the Knee Score is independent of these factors, while the Function Score can decline with multiple joint arthrosis and generalized debility.

Our purpose was to determine the reliability of the AKS, specifically the degree to which the AKS produces consistent results between different observers (i.e., the interobserver reliability), as well as with repeated examination by the same individual (intraobserver reproducibility).

Patients and methods

29 subjects were evaluated by 6 observers with the AKS in random order. In order to represent a wide range of knee functions, the subjects selected consisted of patients awaiting primary or revision knee arthroplasty and patients who have received knee replacements at least 6 months prior to the study.

The 6 observers consisted of 1 consultant, 2 registrars (residents), 1 senior house officer (junior doctor) and 2 arthroplasty nurse practitioners. The observers varied in their experience with examination of the knee joint. The consultant had an interest in knee surgery, the 2 registrars had at least 3 years' experience in orthopedics while the senior house officer had 3 months' experience in this field. The 2 nurses provide domiciliary support for arthroplasty patients and their responsibilities occasionally include clinical assessment of recipients of joint replacements. The senior house officer and the nurses have been instructed on the examination of the knee joint and were familiar with the AKS prior to the study.

Each observer scored the same patient twice, at least 2 hours apart. The observers' scores were compared with each other to estimate the interobserver reliability, and a single observers' 2 scores for each subject were compared for intraobserver reproducibility.

Scores were recorded in a standardized chart. Appropriately-sized goniometers were provided to estimate the ranges of knee motion and deformities. Range of motion was recorded in the chart by each observer and the entry converted into points by RL.

The Knee Score and Function Score have a range of 1–100 and thus, the data were statistically analyzed as continuous variables. To determine the interobserver variation, the method described by Morris was employed (1997). Briefly, for each of the subjects, the standard deviation of the observers' scores were determined and squared to obtain a set of variances. The average of the variances gave the interobserver variance and the square root of this, the interobserver standard deviation. Multiplying the standard deviation by 1.96 gave the "95% confidence intervals". In statistical terminology, "95% reference interval" is more appropriate than "confidence interval" and this term is used throughout this paper. The same technique was employed to quantify the intraobserver variances, standard deviation and 95% reference interval. Single factor ANOVA was also used to verify the calculations, by defining the "within group mean square" or the "residual mean square" (Morris, 1997) . Differences between the variances were tested for significance, using the variance ratio test (F test).

To assess intraobserver agreement for each component of the AKS, unweighted Kappa statistic was applied (Cohen 1960). Kappa values for multiple observers were calculated, as described by Siegel and Castellan (1988), to estimate the interobserver agreement. The level of acceptability was based on the recommendation of Landis and Koch (1977).

Table 1. Interobserver variances by patient category

Category	Knee Score	Function Score
A	35	131
B	76	104
C	73	108

No significant difference between categories (F test)

Results

The interobserver variance for the Knee Score was 64 (SD 8.0). This gave a 95% reference interval that a single observer measuring a single subject will obtain a score within 16 points of the true score. Analysis of variance (ANOVA) showed that the observers did not differ in their scores ($p = 0.9$). The Function Score yielded an interobserver variance of 112 (SD 11) and a 95% reference interval of 21 points. The observers did not differ ($p = 0.3$) in the scores that they obtained.

The variances of the intraobserver Knee Score and Function Score were smaller compared to the interobserver variances. For the Knee Score, the intraobserver variance was 33 (SD 5.7). A single observation from a single assessor is likely to lie within 11 points of the true value. For the Function Score, the intraobserver variance was 41 (SD 6.4), giving a 95% reference limit of 13 points.

The interobserver variances were analyzed with the subjects stratified according to the multiplicity of joints affected by arthrosis and general infirmity (Table 1). Interobserver variance for the Knee Score was smallest in category A patients. Similar magnitudes of variances were obtained for Function Score in all three categories. The variance ratio test (F test) revealed no significant differences between the categories in both the Knee Score and the Function Score.

For the Knee Score, the senior house officer had the highest intraobserver variance which differed significantly from the other observers. A signifi-

Table 2. Intraobserver variances (reproducibility) by grade of observer

	Knee Score	Function Score
Nurse practitioners	29	53 [c]
Senior house officer (SHO)	74 [a]	34 [c]
Registrars	17	35 [c]
Consultant	31 [b]	39 [c]

[a] The senior house officer differed significantly from all other grades ($p < 0.05$, F test)
[b] The consultant differed from registrars ($p < 0.05$, F test)
[c] NS

cant difference was also seen when comparing the variances of the registrars with the consultant. The nurse practitioners had the highest intraobserver variance for the Function Score, although the intergroup differences in variance were not significant (Table 2).

The 6 observers agreed moderately on Pain and Walking Distance categories, while Stairs, Flexion Contracture and Alignment produced fair agreement. The agreement for Mediolateral Stability was no greater than that expected by chance (Table 3). When we excluded the nurses and considered the interobserver agreement for the consultants the registrars only, the Kappa values for Flexion Contracture, Alignment and Mediolateral Stability were higher at 0.19, 0.32 and 0.36, respectively. This implies that the assessments of the nurses and senior house officer were quite different from those of the more experienced individuals in the study.

We found moderate to excellent intraobserver agreement for Pain, and moderate to fair levels for Walking Distance and Stairs. Although Mediolateral Stability produced the least agreement between the 6 observers, the intraobserver agreement was higher. Of note, the 2 nurses obtained perfect intraobserver agreement while the other individuals obtained fair to moderate agreement. Similarly, the nurses were perfectly consistent in the assessment of Alignment. Fair to good levels of intraobserver agreement were found for Flexion Contracture. We noted that the nurses failed to detect significant instability and alignment in the same patients, suggesting that their classification was consistently wrong.

Variation in the range of motion is expressed by the variances of the points, each point representing 5° of motion. A single observer taking a single reading is 95% likely to be within 3.9 points or 20° of the true value. The intraobserver variance of the group of assessors was smaller, with a 95% reference limit of 2.7 points or 13°. The nurses were the least consistent with their goniometric measurement while the Registrars, the most consistent (Table 4).

Discussion

Unlike many of the outcome measure instruments for total knee arthroplasty, the AKS has been the subject of some validation studies. Kreibich et al. (1996) have evaluated its sensitivity to change and

Table 3. Interobserver reliability and intraobserver reproducibility of component variables of the AKS, Kappa values

	Interobserver reliability	Intraobserver reproducibility					
		Nurse 1	Nurse 2	SHO	Reg. 1	Reg. 2	Cons.
Pain	0.46	0.84	0.61	0.51	0.85	0.65	0.68
Walking distance	0.43	0.77	0.51	0.64	0.60	0.65	0.62
Stairs	0.39	0.40	0.63	0.58	0.64	0.74	0.74
Mediolateral stability	0	1.00	1.00	0.49	0.50	0.50	0.35
Flexion contractures	0.25	0.26	0.64	0.66	0.37	0.52	0.40
Alignment	0.25	1.00	1.00	-0.13	0.83	1.00	0.67

Strength of agreement: < 0.20: poor; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: good; 0.81–1.00: very good (Landis and Koch 1977).

Table 4. Observer variation of the goniometric measurements in AKS points (range 0–25), where each point represents 5° of motion

	Interobserver		Intraobserver	
	Variance	95% ref. interval	Variance	95% ref. interval
All observers	3.8	3.9	1.8	2.7
Nurses			2.5 [a]	3.1
SHO			1.5	2.4
Registrars			1.0	2.0
Consultant			1.8	2.6

[a] The nurses differed significantly from the registrars ($p < 0.03$)

found the AKS to be the most responsive compared to other functional scores. König et al. (1997) have described the sequential course of the AKS in patients after knee arthroplasty. The demographic biases that may influence the reporting of results, using the AKS, have also been reported (Brinker et al. 1997).

The reliability of this instrument has been addressed in only one study in which the interobserver variability was compared with other knee scores (Ryd et al. 1997). The interobserver SD for the Knee Score and the Function Score were 26 and 20 points, respectively, more than twice the magnitude of interobserver error that we obtained. It is difficult to explain this difference but the case-mix of subjects may be a factor (Brinker et al. 1997).

The observations of “trained assessors” are deemed necessary when reliability of scoring systems are evaluated. Should a poor interobserver agreement be obtained, this can be confidently attributed to the poor reliability of the scoring instrument and not to the heterogeneity of the observers (Sun et al. 1997). As greater numbers of knee arthroplasties are performed, the surveillance of joint replacement patients becomes more time-consuming (Bankes et al. 1999). As a consequence, the role of patient surveillance is increasingly undertaken by arthroplasty outcome teams which encompass allied health professionals and junior orthopedic staff (O’Brien and Beverland 1998). It is important to determine the reliability of outcome scores such as the AKS when employed by these individuals. For this reason, we have included relatively inexperienced assessors in our study. We have shown that observer experience influences the consistency of scoring as the lesser experienced observers had larger intraobserver variations. Despite the presence of inexperienced assessors in our study, it is gratifying to find the interobserver variation among this group of observers to be markedly smaller than in the study by Ryd et al. (1997) in which 10 experienced orthopedic surgeons were the assessors.

The assessment of Mediolateral Stability and Alignment by the nurses differed greatly from that made by the registrars and consultant. Interobserver agreement when all 6 assessors were considered was low, but when we excluded the nurses and the senior house officer, the interobserver agreement between the registrars and the consultant improved . We cannot conclusively determine who is correct, but the less experienced individuals failed to detect the instability and malalignment in 2 patients who were awaiting total knee replacements with constrained implants. The fact that the nurses obtained perfect agreement in 2 of these categories meant that their assessments were consistently erroneous. This reinforces our belief that familiarity in the examination of patients is necessary for better scoring accuracy. Even then, when the experienced observers detected abnormalities in these parameters, they, at best, only moderately agreed on the severity. This makes the estimation of severity of instability and malalignment rather artificial.

Data for outcome measurement are generally classified as objective or “hard”, and subjective or “soft”. The traditional belief that “hard” data is more reliable than “soft” data is beginning to be dispelled. In a study of the interobserver reliability of the Jones Score (Jones et al. 1991), a rating system for arthrosis of the knee, the objective components of the score (clinical signs) registered Kappa values of 0.09–0.35 while the subjective variables (pain and stiffness) produced Kappa values of 0.46–0.72. Evidence from our study mirrored their findings. The degree of interobserver agreement for Pain, Walking Distance and Stairs were superior to that of Stability, Contractures and Alignment. Research in many areas of medicine and surgery has shown that patients can give reliable information about their health status and the benefits of treatment (Fitzpatrick et al. 1992, Johanson et al. 1992). The greater reliability of the subjective components adds weight to the use of patient-administered questionnaires to assess function after joint arthroplasties where "noise" from unreliable, so-called "hard" data can be eliminated.

The reference limit that we found for range of motion assessment was quite considerable, about 20°. This is higher than the 12° level of interobserver reliability quoted in other studies of goniometric measurement of the knee (Boone et al. 1978). The inclusion of unreliable recording of motion range by the nurses can partly explain the poor interobserver reliability. Furthermore, there may be a tendency for the assessors to round their recording of motion range to the nearest 5°, the equivalent of a single point in this component of the AKS.

Ambiguity in the options given for some sections of the AKS may be responsible for the interobserver variation. Pain is assessed in 2 dimensions, namely its severity and frequency. Similarly, with walking distance, every observer differed in their opinion as to what constitutes the distance covered by walking "a block". This must partly explain the interobserver variation.

The AKS attempts to isolate deteriorating function when the patient has any other condition that may undermine his walking ability. The Knee Score is based on variables that should remain unaffected by patient debility, in contrast to the Function Score. Our study has shown that the reliability of the Knee Score was not affected by categorization of the patient according to debility, vindicating the proposal of the original authors. Similarly, interobserver reliability for the Function Score did not deteriorate with increasing disability in patients.

The 16 points and 21 points 95% reference intervals for the Knee and Function Scores were quite large. Designation of the numerical score into a category (such as excellent, good, fair, poor) have frequently been used by authors to report their results. If these categories are assigned narrow intervals, such as 10 points per category (Wright et al. 1990, König et al. 1997), it is conceivable that a single reading by a single observer may produce an error that spans 2 categories.

We found considerable interobserver and intraobserver variations in the AKS. The clinical importance of this has been reported by Morris (1997). Using the coefficient of reproducibility (British Standard Institution 1979, Bland and Altman 1986), we can estimate that for the Knee Score (1) if a patient is seen by 2 different observers at 2 visits, we cannot be confident that a change is real unless it exceeds 23 points (interobserver standard deviation multiplied by the coefficient of reproducibility: $8.0 \times 2.9 = 23$), (2) if a patient is seen by the same observer on 2 occasions, the corresponding value is 16 points (intraobserver standard deviation multiplied by the coefficient of reproducibility: $5.7 \times 2.8 = 16$) and (3) if the same observer makes 2 assessments at 2 visits, a change of 12 points would be significant (intraobserver variance halved and square-rooted, multiplied by the coefficient of reproducibility: $\sqrt{33.0/2} \times 2.8 = 12$). Corresponding figures for the Function Score would be (1) 30 points, (2) 19 points and (3) 13 points. Thus, for reliable use of the AKS, evaluation should be made by one experienced observer who is familiar with it, preferably using repeated measurements.

We wish to thank Sister W Bradshaw and Dr L Howe for organizing and participating in the study, and the patients for giving their valuable time. We also acknowledge Prof. Peter Kelly for providing advice on statistical methods.

Bankes M J K, Coull R, Ferris B D. How long should patients be followed-up after total hip replacement? Ann Royal Coll Surg Eng. 1999; 81: 348-51.

Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986: i: 307-10.

Boone D C, Azen S P, Lin C-M, Spence C et al. Reliability of goniometric measurements. Phys Ther 1978; 11: 1355-60.

Brinker M R, Lund P J, Barrack R L. Demographic biases of scoring instruments for the results of total knee arthroplasty. J Bone Joint Surg (Am) 1997; 79: 858-65.

British Standard Institution. Precision test methods I: Guide for the determination and reproducibility for a standard test method. (BS5497, Part I). BSI, 1979, London.

Cohen J. A coefficient of agreement for nominal scales. Edu.Psychol Measurement 1960; 20: 37-46.

Fitzpatrick R, Fletcher A, Gore S et al. Quality of life measures in Health care. I: Applications and issues in assessment. BMJ 1992; 305: 1074-7.

Insall J N, Dorr L D, Scott R D, Scott W N. Rationale of the Knee Society Clinical Rating System. Clin Orthop 1989; 248: 13-4.

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

Jones A, Hopkinson N, Pattrick M, Berman P, Doherty M. Evaluation of a method for clinically assessing osteoarthritis of the knee. Ann Rheum Dis 1991; 51: 243-5.

Kreibich D N, Vaz M, Bourne R B, Rorabeck C H et al. What is the best way of assessing outcome after total knee replacement? Clin Orthop 1996; 331: 221-5.

König A, Scheidler M, Rader C, Eulert J. The need for a dual rating system in total knee arthroplasty. Clin Orthop 1997; 345: 161-7.

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

Morris R W. Assessing reliability in clinical measurement. In: Assessment methodology in orthopaedics (Eds. Pynsent P, Fairbank J, Carr A J). Butterworth Heinemann, Oxford 1997: 1-17.

O'Brien S, Beverland D E. Setting up a hip and knee replacement outcomes team. J Orthop Nursing. 1998; 2: 196-201.

Ryd L, Kärrholm J, Ahlvin P et al. Knee scoring systems in gonarthrosis: evaluation of interobserver variability and the envelope of bias. Acta Orthop Scand 1997; 68: 41-6.

Siegel S, Castellan N J. Nonparametric statistics for the behavioural sciences. Second ed. Singapore, McGraw-Hill, 1988: 285-90.

Sun Y, Sturmer T, Gunther K P, Brenner H. Reliability and validity of clinical outcome measurement of osteoarthritis of the hip and knee - a review of the literature. Clin Rheumatol 1997; 16: 185-98.

Wright R J, Lima ChB J, Scott R D, Thornhill T S. Two to four-year results of posterior cruciate- sparing condylar total knee arthroplasty with an uncemented femoral component. Clin Orthop 1990; 260: 80-6.