Interobserver and Intraobserver Reliability of Clinical Assessments in Knee Osteoarthritis

Nasimah Maricar, Michael J. Callaghan, Matthew J. Parkes, David T. Felson, and Terence W. O'Neill

ABSTRACT. ***Objective.*** Clinical examination of the knee is subject to measurement error. The aim of this analysis was to determine interobserver and intraobserver reliability of commonly used clinical tests in patients with knee osteoarthritis (OA).

Methods. We studied subjects with symptomatic knee OA who were participants in an open-label clinical trial of intraarticular steroid therapy. Following standardization of the clinical test procedures, 2 clinicians assessed 25 subjects independently at the same visit, and the same clinician assessed 88 subjects over an interval period of 2–10 weeks; in both cases prior to the steroid intervention. Clinical examination included assessment of bony enlargement, crepitus, quadriceps wasting, knee effusion, joint-line and anserine tenderness, and knee range of movement (ROM). Intraclass correlation coefficients (ICC), estimated kappa (κ), weighted kappa (κω), and Bland-Altman plots were used to determine interobserver and intraobserver levels of agreement.

Results. Using Landis and Koch criteria, interobserver κ scores were moderate for patellofemoral joint (κ = 0.53) and anserine tenderness (κ = 0.48); good for bony enlargement (κ = 0.66), quadriceps wasting (κ = 0.78), crepitus (κ = 0.78), medial tibiofemoral joint tenderness (κ = 0.76), and effusion assessed by ballottement (κ = 0.73) and bulge sign (κω = 0.78); and excellent for lateral tibiofemoral joint tenderness (κ = 1.00), flexion (ICC = 0.97), and extension (ICC = 0.87) ROM. Intraobserver κ scores were moderate for lateral tibiofemoral joint tenderness (κ = 0.60); good for crepitus (κ = 0.78), effusion assessed by ballottement test (κ = 0.77), patellofemoral joint (κ = 0.66), medial tibiofemoral joint (κ = 0.64), and anserine tenderness (κ = 0.73); and excellent for effusion assessed by bulge sign (κω = 0.83), bony enlargement (κ = 0.98), quadriceps wasting (κ = 0.83), flexion (ICC = 0.99), and extension (ICC = 0.96) ROM.

Conclusion. Among individuals with symptomatic knee OA, the reliability of clinical examination of the knee was at least good for the majority of clinical signs of knee OA. (First Release October 1 2016; J Rheumatol 2016;43:2171–8; doi:10.3899/jrheum.150835)

Key Indexing Terms:
KNEE OSTEOARTHRITIS
INTEROBSERVER RELIABILITY
CLINICAL TESTS
INTRAOBSERVER RELIABILITY

From the Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre (MAHSC), University of Manchester; UK National Institute for Health Research (NIHR) Manchester Musculoskeletal Biomedical Research Unit, Central Manchester National Health Service (NHS) Foundation Trust, MAHSC, Manchester; Department of Physiotherapy, and Department of Rheumatology, Salford Royal NHS Foundation Trust, Salford, UK; Clinical Epidemiology Unit, Boston University School of Medicine, Boston, Massachusetts, USA.

Funded by Arthritis Research UK grant 20380, and special strategic award grant 18676. The funding agency had no role in any of the following: design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; and the decision to submit the manuscript for publication. This report includes independent research supported by (or funded by) the NIHR Biomedical Research Unit Funding Scheme. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. The Research in Osteoarthritis Manchester group is supported by the MAHSC. N. Maricar is supported by an NIHR Allied Health Professional Clinical Doctoral Fellowship.

N. Maricar, BSc (Hons), MSc, PhD, Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, MAHSC, University of Manchester, NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, MAHSC, and Department of Physiotherapy, Salford Royal NHS Foundation Trust; M.J. Callaghan, PhD, Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, MAHSC, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, MAHSC; M.J. Parkes, BSc (Hons), Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, MAHSC, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, MAHSC; D.T. Felson, MD, MPH, Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, MAHSC, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, MAHSC, and Clinical Epidemiology Unit, Boston University School of Medicine; T.W. O'Neill, MB, BCh, BAO, Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, MAHSC, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, MAHSC, and Department of Rheumatology, Salford Royal NHS Foundation Trust.

Address correspondence to Dr. N. Maricar, University of Manchester, Arthritis Research UK Centre for Epidemiology, Institute of Inflammation

and Repair, Faculty of Medical and Human Sciences, MAHSC, Manchester M13 9PT, UK. E-mail: nasimah.maricar@postgrad.manchester.ac.uk; nasimah.maricar@srft.nhs.uk

Full Release Article. For details see Reprints/Permissions at jrheum.org

Accepted for publication August 31, 2016.

Clinical assessment of the knee forms an integral part of any joint examination in osteoarthritis (OA) and includes a variety of specific clinical tests including assessment of tenderness[1,2,3], presence of effusion[4,5,6,7,8] or bony enlargement[1,3,9], muscle atrophy[9], and crepitus[2,9]. As with any clinical test, clinical examination of the knee is subject to measurement error. There are, however, few studies that have formally measured reliability in the assessment of common clinical signs for knee OA and in those studies that have reported reliability, findings have been somewhat inconsistent[2,3,4,9,10,11,12]. Some contributing factors to the inconsistency include lack of clarity and uniformity in the assessment procedures and the grading criteria[2,3,4,9,10,11,12]. Reliable clinical assessment is important, because poor reliability may result in misclassification in clinical and research studies of knee OA and reduce the chance of finding clinically important biological associations between clinical features of the disease and outcome or response to therapy. The aim of our study was to determine intraobserver and interobserver reliability for commonly used clinical tests in the assessment of knee OA.

MATERIALS AND METHODS

Subjects. Men and women aged 40 years and over were recruited from primary and secondary care clinics for participation in an open-label study (TASK)[13] looking at the efficacy of intraarticular steroid therapy in symptomatic knee OA (ISRCTN: 07329370). Subjects were included in the trial if they met the American College of Rheumatology (ACR) criteria including moderate knee pain for more than 48 h in the previous 2 weeks or scored greater than 7 out of 32 on the Knee Injury and Osteoarthritis Outcome Score, questions P2–P9. Other inclusion criteria included imaging confirmation of definite OA on radiograph [Kellgren-Lawrence (KL) score ≥ 2] by an expert musculoskeletal radiologist or typical changes of OA with at least cartilage loss on magnetic resonance imaging (MRI) scan or at arthroscopy. The exclusion criteria were the presence of gout, previous septic arthritis, or inflammatory arthritis, injection with hyaluronic acid or steroid injection within the previous 3 months, history of knee surgery within the previous 6 months, concurrent life-threatening illness, and any contraindication to MRI scanning. Ethics approval was obtained from the Leicestershire Multicentre Research Ethics Committee, reference 09/H0402/107.

Assessment of reliability. A standardized assessment was developed to provide clarity and consistency on the examination procedure. Several patients with knee OA were examined to test the standardized assessment procedure and to resolve issues about the procedure and outcome categorization. An "unsure/possible" category was included in some of the outcome assessment of the clinical tests for indeterminate cases where assessors were uncertain or comparison to the opposite knee was not possible because of bilateral knee OA. The final standardized examination included assessment of bony enlargement (absent = 0, unsure = 1, present = 2), joint crepitus (absent = 0, unsure = 1, present palpable = 2, present audible = 3), quadriceps muscle wasting (absent = 0, possible = 1, present = 2), assessment of effusion using the bulge sign (no wave produced on downstroke = 0, a small wave on medial side with downstroke = trace, larger bulge on medial side with downstroke = 1, spontaneously returned to medial side after upstroke = 2, so much fluid that it was not possible to move the effusion out of the medial aspect of the knee = 3)[4]. The examination also included assessment of effusion using the ballottement test [absent = 0, present without click = 1, present with click (tap) = 2], and the following, all scored absent = 0, present = 1: patellofemoral joint tenderness, pes anserine tenderness, medial tibiofemoral joint tenderness, and lateral tibiofemoral joint tenderness. Goniometric knee range of movement (ROM) assessment included flexion and extension, measured to the nearest degrees[14]. Assessments were undertaken prior to the participants having their steroid injections. Description of the assessment and outcome categories is available from the authors on request.

Interobserver reliability assessment. An opportunity sample of 25 unselected participants who presented at the screening visit of the TASK study was assessed independently by 2 observers (TON, NM), typically within a 30-min to 60-min interval between each other's assessment. One was an experienced rheumatologist (TON) and the other (NM) was an Advanced Musculoskeletal (MSK) Practitioner (senior physiotherapist) with more than 15 years of experience in MSK. The assessors were blinded to each other's assessments, and the examination findings were recorded on different summary sheets. During the clinical examination, the individual clinicians performed each test a few times as needed for a consistent recording. For instance, during the performance of bulge sign, the sequence (the upstroke on the medial aspect of the knee followed by the downstroke on the lateral aspect of the knee) could be repeated a few times when attempting to observe reappearance of fluid.

Intraobserver reliability assessment. An opportunity sample of 88 unselected subjects who attended the screening and baseline visits of the TASK study was assessed for intraobserver reliability. One assessor (NM) undertook a single repeat clinical assessment of the 88 subjects separated by an interval of between 2 to 10 weeks, prior to their steroid injections.

It was anticipated that because of the different number of subjects in the assessment of interobserver reliability (compared with intraobserver reliability) that the prevalence of individual examination features may differ.

Analysis. Intraobserver and interobserver reliability were assessed using intraclass correlation coefficients (ICC) for continuous variables ICC (2,1; 2-way random effect with rater as random effect)[15], estimated κ for dichotomous variables where 2 × 2 contingency tables were used, and weighted kappa [κω; linear weights were used, i.e., $w_i = 1 - (i / (k-1))$ for ordinal variables using Stata version 13.1. For the determination of ICC, in the model "assessor" was treated as a random effect; in our analysis, however, treating the assessors as random or fixed effects made very little difference to the ICC values or their CI. For the determination of estimated κ values of items scored absent/present, 2 × 2 tables were used. The items included patellofemoral joint, pes anserine, medial and lateral tibiofemoral joint tenderness, and clinical tests of bony enlargement, knee crepitus, quadriceps wasting, and effusion assessed using the ballottement test. For bony enlargement, we dichotomized the variable as present versus absent/unsure while for knee joint crepitus, we dichotomized as either present palpatory/audible crepitus versus absent/unsure. For quadriceps wasting, we dichotomized as present versus absent/possible. For assessment of effusion using ballottement, we looked at those with a positive test (either ballottement or patella tap/click) compared to those without. For the assessment of effusion using the bulge sign, where there were 5 possible categories, a weighted κ was used. For ICC and κ, values of < 0.2 were considered as indicating poor agreement, between 0.21 and 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 as good, and values above 0.80 as excellent[16]. For continuous data (goniometric knee ROM), Bland-Altman plots were used to determine the limits of agreement, and 95% CI about the mean difference both within and between observers were constructed to test for bias between assessors[17].

RESULTS

Subjects. The mean age of the 25 subjects included in the interobserver reliability assessment was 63 years (SD 10) and

14 (56%) were female. Among these subjects, 14% had KL grade 2, 67% had KL grade 3, and 19% had KL grade 4. Mean age of the 88 subjects included in the intraobserver reliability assessment was 64 years (SD 10), and 46 (52%) were female. Of these, 34% were KL grade 2, 55% KL grade 3, and 11% KL grade 4.

Interobserver reliability. Interobserver κ scores as assessed by estimated κ were excellent for the assessment of lateral tibiofemoral joint tenderness (κ = 1.00), and good for a number of other clinical signs including assessment of bony enlargement, quadriceps wasting, crepitus, medial tibiofemoral joint tenderness, and the presence of effusion assessed using the bulge sign and ballottement test (κ = 0.66–0.78; Table 1). Interobserver estimated κ scores were moderate for the assessment of patellofemoral joint tenderness and pes anserine tenderness (κ = 0.48–0.53). ICC were excellent for the assessment of the degrees of knee flexion and extension ROM (ICC = 0.87–0.97; Table 2). For knee flexion, the limits of agreement between observers were –12.29° to 7.81°. There was evidence of a relatively small difference in the assessment between observers (mean difference = –2.24°; 95% CI –4.36 to –0.12; Figure 1 and Table 2). For knee extension, the limits of agreement between observers were –8.38° to 6.38°. There was no evidence of a significant difference between observers with the 95% CI around the mean difference including zero (Figure 2). The percentage of raw agreement for all tests was high (≥ 80%).

Intraobserver reliability. Intraobserver estimated κ scores were excellent for bony enlargement, quadriceps wasting, the presence of effusion assessed using the bulge sign, and knee flexion and extension ROM (κ = 0.83–0.98; ICC = 0.96–0.99) and good for the other clinical tests such as knee joint crepitus, patellofemoral joint, medial tibiofemoral joint, and pes anserine tenderness, and the assessment of effusion using ballottement test (κ = 0.64–0.78; Table 1 and Table 2). Intraobserver estimated κ score was moderate for lateral tibiofemoral joint tenderness (κ = 0.60). The intraobserver estimated κ scores for the clinical tests for knee OA were higher than their respective interobserver κ scores apart from medial and lateral tibiofemoral joint tenderness. In the assessment of both knee flexion and extension, the 95% CI around the mean difference included zero, suggesting no detectable evidence of bias (Figure 3 and Figure 4). The percentage of raw agreement for the clinical tests was high (81.8%–98.9%). With the exception of medial and lateral tibiofemoral joint tenderness, the percentage of raw agreement for all tests was higher for intraobservers than interobservers.

Table 1. Interobserver and intraobserver reliability of clinical tests for knee osteoarthritis.

Clinical Evaluation	Outcome Category	Interobserver Reliability				Intraobserver Reliability			
		Prevalence^	K Values	95% CI	% Raw Agreement	Prevalence^	K Values	95% CI	% Raw Agreement
Bony enlargement	Present vs absent/unsure	0.22	0.66	0.32–1.00	88.0	0.37	0.98	0.93–1.00	98.9
Quadriceps wasting	Present vs absent/possible	0.24	0.78	0.40–1.00	84.0	0.62	0.83	0.72–0.95	90.9
Knee joint crepitus	Present palpable/ audible vs absent/unsure	0.90	0.78	0.36–1.00	96.0	0.91	0.78	0.55–1.00	96.6
Medial tibiofemoral joint tenderness	Present vs absent	0.54	0.76	0.50–1.00	88.0	0.48	0.64	0.49–0.80	81.8
Lateral tibiofemoral joint tenderness	Present vs absent	0.28	1.00	1.00–1.00	100.0	0.22	0.60	0.39–0.80	86.4
Patellofemoral joint tenderness	Present vs absent	0.30	0.53	0.16–0.89	80.0	0.36	0.66	0.60–0.92	84.1
Anserine tenderness	Present vs absent	0.26	0.48	0.09–0.87	80.0	0.22	0.73	0.61–0.99	90.9
Effusion: bulge sign	5-point Likert scale	0.96	0.78*	0.55–1.00	80.0	0.64	0.83*	0.73–0.94	85.2
Effusion: ballottement test**	Present with/ without click vs absent	0.66	0.73	0.45–1.00	88.0	0.18	0.77	0.60–0.95	93.2

^ Prevalence calculated as the average of positive findings between the 2 rated scores. * Weighted κ. ** Ballottement test defined as positive click/tap or downward movement of the patella on pressure and rebounding of patella upon removal of pressure.

Table 2. Interobserver and intraobserver reliability for measurement of passive knee range of movement.

Knee PROM	Interobserver Agreement				Intraobserver Agreement			
	ICC (95% CI)	Mean Difference (95% CI)	95% LoA (°)	SEM	ICC (95% CI)	Mean Difference (95% CI)	95% LoA (°)	SEM
Flexion	0.97 (0.92–0.99)	–2.24 (–4.36 to –0.12)	–12.29, 7.81	1.03	0.99 (0.99–0.99)	–0.14 (–0.43 to 0.70)	–5.37, 5.10	0.28
Extension	0.87 (0.72–0.94)	–1.00 (–2.55 to 0.55)	–8.38, 6.38	0.75	0.96 (0.94–0.98)	–0.17 (–0.61 to 0.27)	–4.23, 3.89	0.22

ICC: intraclass correlation coefficient; PROM: passive range of movement; LoA: limits of agreement; SEM: standard error of measurement.

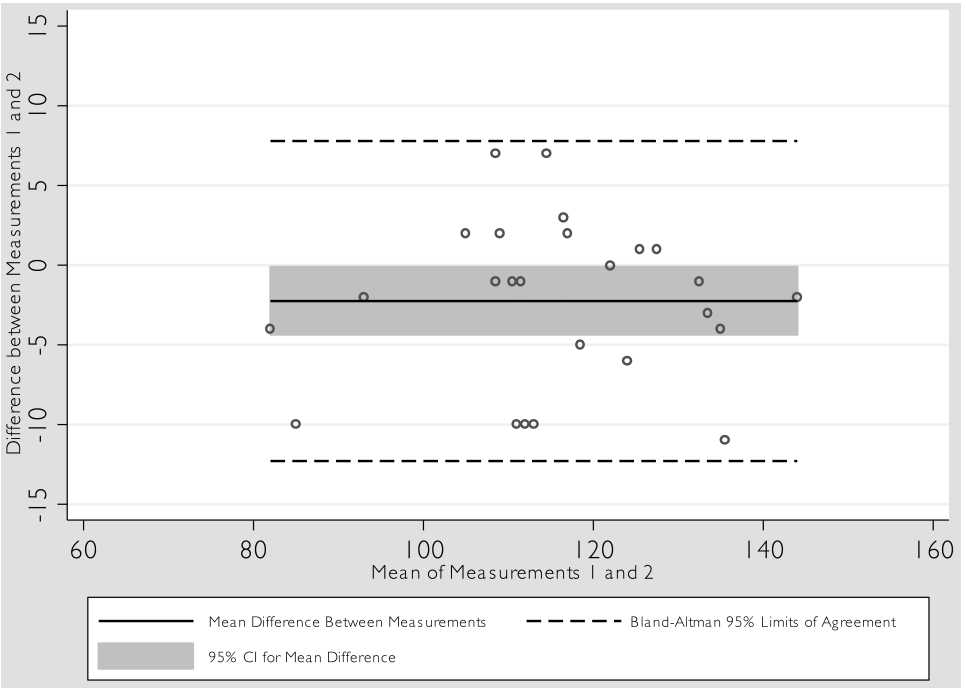

Figure 1. Bland-Altman plot for interobserver agreement for knee flexion range of movement.

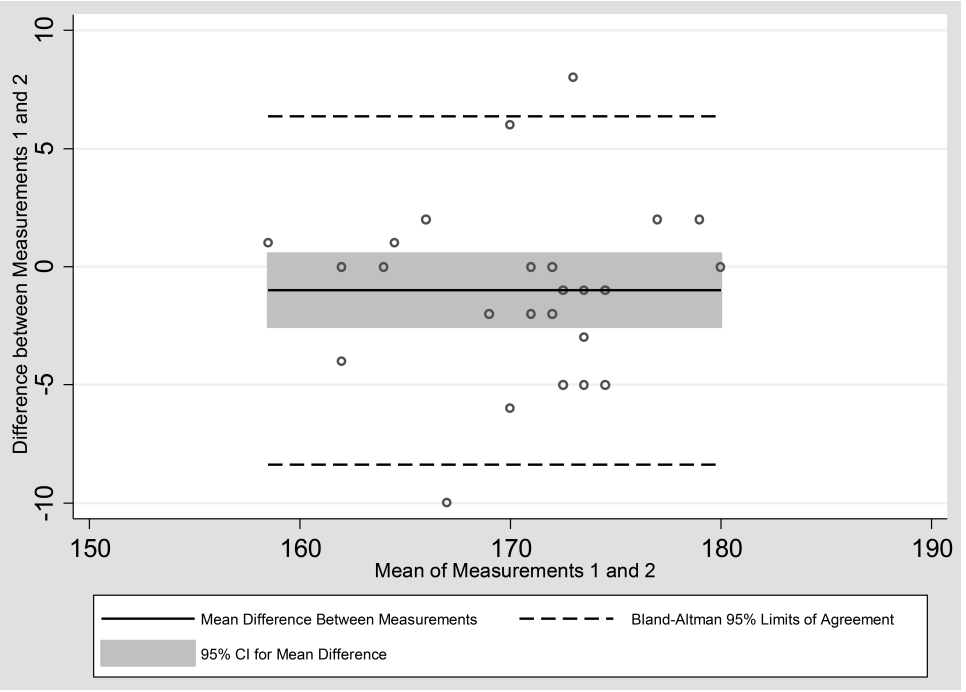

Figure 2. Bland-Altman plot for interobserver agreement for knee extension range of movement.

DISCUSSION

In our study we have shown, using a standardized assessment, at least good reliability for commonly used clinical tests for the assessment of knee OA. As expected, intraobserver reliability of the clinical tests was higher than interobserver reliability.

A variety of clinical tests has been used to assess the presence of knee effusion[5,8,18] including both static and

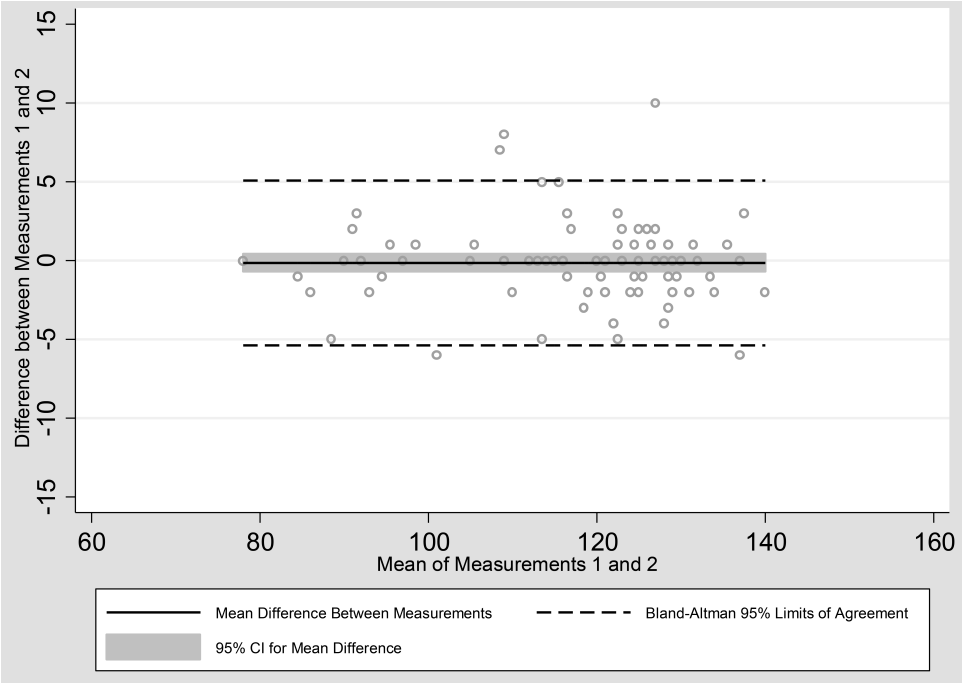

Figure 3. Bland-Altman plot for intraobserver agreement for knee flexion range of movement.

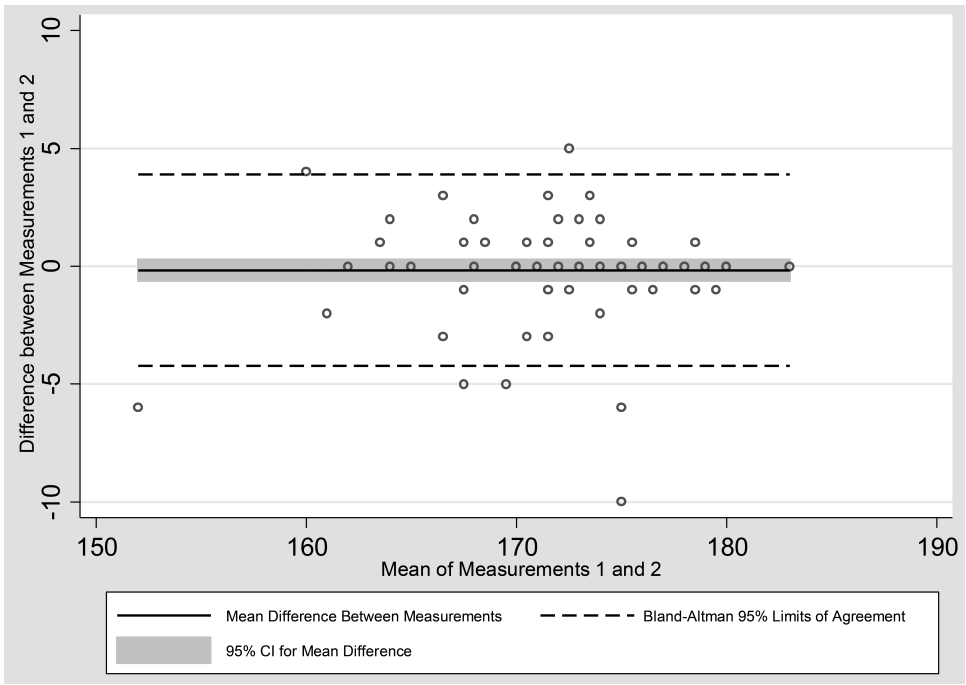

Figure 4. Bland-Altman plot for intraobserver agreement for knee extension range of movement.

dynamic tests, although the terminology used in the literature to describe the tests is inconsistent[4,5,6,7,8]. We looked at the reliability of 2 tests, the bulge sign and ballottement of the patella, with a positive test defined as either rebounding movement of the patella or a patella click (or "tap"). For bulge sign, the 5-point scale described by Sturgill, *et al*[4] was used. The estimated κ score for interobserver agreement for the assessment of effusion using the bulge sign ($\kappa\omega = 0.78$) was higher in magnitude than that reported by Sturgill, *et al*[4] ($\kappa\omega = 0.68$) and several other studies in which effusion was categorized as present or absent or not defined[3,10], but lower than that reported by Cibere, *et al* [reliability coefficient (R_c)

= 0.97][9], though the latter study used a different method of assessment of reliability. For intraobserver estimated κ scores, we could only compare the value observed in this analysis with 1 study that used a 4-point scale (κω = 0.35)[3] to assess effusion, in which the κ score was lower. Differences in the sample and assessment scale are possible reasons for the apparent differences.

For the assessment of knee crepitus, a higher estimated κ value for interobserver agreement was observed (0.78) in comparison to other studies that achieved κ scores varying from 0.22 to 0.64[1,2,3]. Two of these studies[1,2] used a similar grading system (absent, present) while 1 study[3] looked for coarse crepitus during the movement of sitting to standing. Cibere, *et al*[9], who used a different scale (none, fine, coarse) to assess knee crepitus, achieved R_c = 0.67 during the assessment of active knee movement and R_c = 0.96 with passive knee movement. For intraobserver estimated κ scores for the assessment of knee joint crepitus, we achieved a higher score (0.78) than 1 study[1] (0.68 for tibiofemoral crepitus and 0.50 for patellofemoral crepitus) that used a similar grading system (absent, present) and another study[3] (0.53) that assessed knee crepitus during sitting to standing movement. That study was comparable with 1 other study[2] (κ = 0.78 for tibiofemoral crepitus and 0.75 for patello-femoral crepitus), in which crepitus was categorized as absent or present.

For the assessment of patellofemoral joint tenderness, the estimated κ scores for intraobserver (0.66) and interobserver (0.53) were higher than those found in other studies[1,2] that used similar grading of tenderness (absent, present). Their intraobserver and interobserver estimated κ scores varied from 0.41–0.61 and 0.27–0.35, respectively. It is possible that the experience or skill of the assessors in the current study may have contributed to the better observer estimated κ scores. For the assessment of quadriceps wasting and pes anserine tenderness, we reported lower interobserver estimated κ scores than those found by Cibere, *et al*[9], though the latter used a different grading scale (none, mild, severe) for the assessment of quadriceps muscle wasting and a different method of assessment of reliability (R_c).

Bony enlargement in the knee is also often consequential to more advanced degeneration of the joint[19] and our higher intraobserver and interobserver estimated κ scores when compared to another study[3] could be due to a higher prevalence of patients with OA in our study, and the latter categorizing bony enlargement as either medial or lateral. The κ values are affected by prevalence of the exposure or baseline frequency, with a high or low prevalence in a sample tending to lower the value of κ, so caution is required when comparing κ values from different studies[20]. Our interobserver estimated κ score for bony enlargement (0.66) was also higher than that of 2 other studies[1,21] (0.55 and 0.10, respectively) but lower than that of Cibere, *et al*[9] (R_c = 0.97). The Cibere study used a different assessment scale (none, mild, moderate, severe) and assessed bony swelling through palpation rather than through palpation and visual inspection, as in our study.

In our analysis there was a high estimated κ score for interobserver reliability of lateral tibiofemoral joint tenderness. Two other studies used similar nominal grading for lateral and medial knee joint tenderness; one[9] also found a high reliability coefficient (R_c = 0.85–0.94), though another reported lower estimated κ scores (κ = 0.40–0.43)[1]. The discrepancy in the findings could be due to less-experienced assessors (3 trainees out of 5 assessors) included in the latter study[1].

We found that the reliability of knee ROM measurement was excellent for both flexion and extension. These findings are consistent with other studies that used different cohorts such as individuals who just had total knee arthroplasties[22] and MSK disorders of the knee seen in physiotherapy clinics[23,24]. There was no evidence for any statistically significant bias in the assessment of knee extension, though there was a small significant difference between observers in the assessment of flexion ROM. The minimal detectable change for goniometric knee measurement in knee OA is not known, though in a different population sample and clinical setting such as postarthroscopic knee within 4 days of surgery[22] it could vary between 8.2° for active extension and 17.6° for passive flexion.

Of all the clinical tests, assessment of effusion using the bulge sign appeared the most reliable. The interobserver estimated κ score for the bulge sign was comparable if not slightly better than those obtained when knee effusion was assessed in some studies using ultrasound (US)[25,26,27,28,29] and MRI[30,31,32,33,34,35]; though estimated κ scores reported in other US and MRI studies were higher (> 0.90)[36,37,38]. The intraobserver estimated κ score for bulge sign was also higher than the assessment with US (0.78) when repeat examinations were performed on the same day[29]. Similarly, a higher intraobserver estimated κ score was observed when compared with MRI in some (κω = 0.60–0.72)[30,39] though not all studies[31,33,34].

For most tests, intraobserver estimated κ scores were higher than interobserver estimated κ scores; however, intraobserver estimated κ scores were lower than interobserver estimated κ scores in the assessment of medial and lateral tibiofemoral joint tenderness. It is possible that this is due to real biological change, with the mean interval between assessments of 32 days for the evaluation of intraobserver estimated κ scores compared to the same-day assessment for interobserver estimated κ scores. When data for medial and lateral tibiofemoral joint tenderness were reanalyzed before and after a threshold of 32 days, the intraobserver estimated κ score for medial tibiofemoral joint tenderness was higher when assessments were made 32 days or sooner (0.80) than when the assessments were more than 32 days apart (0.71). For lateral tibiofemoral joint tenderness, no improvement in

estimated κ score was found, though the overall prevalence of lateral tibiofemoral joint tenderness was relatively low, making the results perhaps less reliable.

There are a number of limitations to be considered in interpreting these data. The clinical assessment reported here comprised 10 common clinical tests; other tests used in clinical practice were not assessed. The reason was pragmatic — to focus on frequently used tests. With the sample comprising those with symptomatic knee OA of KL grade 2 to 4, the findings may not be generalizable to those without OA or those with early radiographic knee OA, or in a different clinical setting. In our study, 2 experienced assessors examined the subjects; it is unclear whether similar findings would be observed with different observers and with different levels of training and experience. In the analysis of intraobserver reliability, subjects were reassessed after an interval period of up to 10 weeks and it is possible that true change in disease characteristics may have occurred during this time. The effect of such true change would be, if anything, to worsen the degree of observer variability. We cannot exclude recall bias in the assessment of intraobserver κ scores; however, such bias seems unlikely given the interval period between the assessments of 32 days [mean 32 days (SD 16.8); min 1 to max 75 days]. The lower reliability for the palpation of tenderness might also be due to difficulty in standardizing the pressure exerted during the assessment of tenderness. Future studies should consider standardizing assessment possibly with the use of a pressure algometer. The use of binary-choice tests in some of the clinical tests could present further limitation because of their low information content. For some of the clinical tests, assessment categories have been collapsed into 2 categories to make them more clinically meaningful, but some caution is needed in interpreting the results.

Generally there were few instances of uncertainty in findings; for example, in the interobserver assessment of crepitus, there was only 1 case of an "unsure." We repeated the interobserver and intraobserver reliability assessment of the clinical tests using all categories within their respective scales and found no overall change in the moderate/good/excellent grading of the tests. We have considered girth or knee circumferential measures; however, we do not consider them specific clinical tests that can differentiate against effusion, muscle atrophy, or bony enlargement. While girth or knee circumferential measures may be useful in monitoring changes in knee effusion[40], for instance during postoperative knee swelling, we do not consider them useful as a 1-time assessment measure. Further comparison against a "normal" measure, that is, against a normal knee is required, but was not always possible because we included people with bilateral knee OA. Some caution should also be taken owing to the small sample size for the interobserver reliability evaluation, with the suggestion that future reliability studies include larger samples. In relation to interobserver reliability, the order in which the assessors examined the participants was not randomized or recorded, so it was not possible to determine whether there was any order effect. Future studies should include provision for assessment of an order effect. Finally, we did not look separately at reliability in men and women.

Clinical examination of knee OA is reliable if a standardized approach to assessment is used. Among subjects with symptomatic knee OA, the reliability of the majority of clinical tests was good. Assessment of effusion using the bulge sign and assessment of quadriceps wasting were among the more reliable clinical tests.

ACKNOWLEDGMENT

The authors acknowledge the equipment and facilities provided by Salford Royal NHS Foundation Trust.

REFERENCES

1. Cushnaghan J, Cooper C, Dieppe P, Kirwan J, McAlindon T, McCrae F. Clinical assessment of osteoarthritis of the knee. Ann Rheum Dis 1990;49:768-70.
2. Jones A, Hopkinson N, Pattrick M, Berman P, Doherty M. Evaluation of a method for clinically assessing osteoarthritis of the knee. Ann Rheum Dis 1992;51:243-5.
3. Wood L, Peat G, Wilkie R, Hay E, Thomas E, Sim J. A study of the noninstrumented physical examination of the knee found high observer variability. J Clin Epidemiol 2006;59:512-20.
4. Sturgill LP, Snyder-Mackler L, Manal TJ, Axe MJ. Interrater reliability of a clinical scale to assess knee joint effusion. J Orthop Sports Phys Ther 2009;39:845-9.
5. Davies GJ, Malone T, Bassett FH III. Knee examination. Phys Ther 1980;60:1565-74.
6. Currey H, Hull S. Rheumatology for general practitioners. Oxford: Oxford Medical Publications; 1987.
7. Doherty M, Hazleman B, Hutton C, Maddison P, Perry J. Rheumatology examination and injection techniques. London: W.B. Saunders; 1992.
8. Isenberg O, Maddison P, Woo P, Glass D, Breedveld F. Oxford textbook of rheumatology. Third ed. New York: Oxford University Press; 2004.
9. Cibere J, Bellamy N, Thorne A, Esdaile JM, McGorm KJ, Chalmers A, et al. Reliability of the knee examination in osteoarthritis: effect of standardization. Arthritis Rheum 2004;50:458-68.
10. Dervin GF, Stiell IG, Wells GA, Rody K, Grabowski J. Physicians' accuracy and interrator reliability for the diagnosis of unstable meniscal tears in patients having osteoarthritis of the knee. Can J Surg 2001;44:267-74.
11. Esen S, Akarirmak U, Aydin FY, Unalan H. Clinical evaluation during the acute exacerbation of knee osteoarthritis: the impact of diagnostic ultrasonography. Rheumatol Int 2013;33:711-7.
12. Ulasli AM, Yaman F, Dikici O, Karaman A, Kacar E, Demirdal US. Accuracy in detecting knee effusion with clinical examination and the effect of effusion, the patient's body mass index, and the clinician's experience. Clin Rheumatol 2014;33:1139-43.
13. O'Neill TW, Parkes MJ, Maricar N, Marjanovic EJ, Hodgson R, Gait AD, et al. Synovial tissue volume: a treatment target in knee osteoarthritis (OA). Ann Rheum Dis 2016;75:84-90.
14. Clarkson HM. Musculoskeletal assessment: joint range of motion and manual muscle strength. 2nd ed. Baltimore: Williams & Wilkins; 1999.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.

16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
18. Altman RD. Criteria for classification of clinical osteoarthritis. J Rheumatol Suppl. 1991 Feb;27:10-2.
19. Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, et al. Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. Arthritis Rheum 1986;29:1039-49.
20. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423-9.
21. Hart DJ, Spector TD, Brown P, Wilson P, Doyle DV, Silman AJ. Clinical signs of early osteoarthritis: reproducibility and relation to x ray changes in 541 women in the general population. Ann Rheum Dis 1991;50:467-70.
22. Lenssen AF, van Dam EM, Crijns YH, Verhey M, Geesink RJ, van den Brandt PA, et al. Reproducibility of goniometric measurement of the knee in the in-hospital phase following total knee arthroplasty. BMC Musculoskelet Disord 2007;8:83.
23. Brosseau L, Balmer S, Tousignant M, O'Sullivan JP, Goudreault C, Goudreault M, et al. Intra- and intertester reliability and criterion validity of the parallelogram and universal goniometers for measuring maximum active knee flexion and extension of patients with knee restrictions. Arch Phys Med Rehabil 2001;82:396-402.
24. Rothstein JM, Miller PJ, Roettger RF. Goniometric reliability in a clinical setting. Elbow and knee measurements. Phys Ther 1983;63:1611-5.
25. Abraham AM, Goff I, Pearce MS, Francis RM, Birrell F. Reliability and validity of ultrasound imaging of features of knee osteoarthritis in the community. BMC Musculoskelet Disord 2011;12:70.
26. Bevers K, Zweers MC, van den Ende CH, Martens HA, Mahler E, Bijlsma JW, et al. Ultrasonographic analysis in knee osteoarthritis: evaluation of inter-observer reliability. Clin Exp Rheumatol 2012;30:673-8.
27. Gok M, Erdem H, Gogus F, Yilmaz S, Karadag O, Simsek I, et al. Relationship of ultrasonographic findings with synovial angiogenesis modulators in different forms of knee arthritides. Rheumatol Int 2013;33:879-85.
28. Iagnocco A, Perricone C, Scirocco C, Ceccarelli F, Modesti M, Gattamelata A, et al. The interobserver reliability of ultrasound in knee osteoarthritis. Rheumatology 2012;51:2013-9.
29. Wu D, Huang Y, Gu Y, Fan W. Efficacies of different preparations of glucosamine for the treatment of osteoarthritis: a meta-analysis of randomised, double-blind, placebo-controlled trials. Int J Clin Pract 2013;67:585-94.
30. Gudbergsen H, Boesen M, Christensen R, Bartels EM, Henriksen M, Danneskiold-Samsoe B, et al. Changes in bone marrow lesions in response to weight-loss in obese knee osteoarthritis patients: a prospective cohort study. BMC Musculoskelet Disord 2013;14:106.
31. Hill CL, Gale DG, Chaisson CE, Skinner K, Kazis L, Gale ME, et al. Knee effusions, popliteal cysts, and synovial thickening: association with knee pain in osteoarthritis. J Rheumatol 2001;28:1330-7.
32. Hunter DJ, Lo GH, Gale D, Grainger AJ, Guermazi A, Conaghan PG. The reliability of a new scoring system for knee osteoarthritis MRI and the validity of bone marrow lesion assessment: BLOKS (Boston Leeds Osteoarthritis Knee Score). Ann Rheum Dis 2008;67:206-11.
33. Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthritis Cartilage 2011;19:990-1002.
34. Railhac JJ, Zaim M, Saurel AS, Vial J, Fournie B. Effect of 12 months treatment with chondroitin sulfate on cartilage volume in knee osteoarthritis patients: a randomized, double-blind, placebo-controlled pilot study using MRI. Clin Rheumatol 2012;31:1347-57.
35. Roemer FW, Guermazi A, Hunter DJ, Niu J, Zhang Y, Englund M, et al. The association of meniscal damage with joint effusion in persons without radiographic osteoarthritis: the Framingham and MOST osteoarthritis studies. Osteoarthritis Cartilage 2009; 17:748-53.
36. Hauzeur JP, Mathy L, De Maertelaer V. Comparison between clinical evaluation and ultrasonography in detecting hydrarthrosis of the knee. J Rheumatol 1999;26:2681-3.
37. Hirsch G, O'Neill T, Kitas G, Klocke R. Distribution of effusion in knee arthritis as measured by high-resolution ultrasound. Clin Rheumatol 2012;31:1243-6.
38. Krasnokutsky S, Belitskaya-Levy I, Bencardino J, Samuels J, Attur M, Regatte R, et al. Quantitative magnetic resonance imaging evidence of synovial proliferation is associated with radiographic severity of knee osteoarthritis. Arthritis Rheum 2011;63:2983-91.
39. Lo GH, McAlindon TE, Niu J, Zhang Y, Beals C, Dabrowski C, et al. Bone marrow lesions and joint effusion are strongly and independently associated with weight-bearing pain in knee osteoarthritis: data from the osteoarthritis initiative. Osteoarthritis Cartilage 2009;17:1562-9.
40. Jakobsen TL, Christensen M, Christensen SS, Olsen M, Bandholm T. Reliability of knee joint range of motion and circumference measurements after total knee arthroplasty: does tester experience matter? Physiother Res Int 2010;15:126-34.