HHS Public Access

Author manuscript

J Rheumatol. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

J Rheumatol. 2016 December ; 43(12): 2171–2178. doi:10.3899/jrheum.150835.

Inter-Observer and Intra-Observer Reliability of Clinical Assessments in Knee Osteoarthritis

Nasimah Maricar[1,2,3], **Michael J Callaghan**[1,2], **Matthew J Parkes**[1,2], **David T Felson**[1,2,4], and **Terence W O'Neill**[1,2,5]

[1]Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

[2]NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

[3]Department of Physiotherapy, Salford Royal NHS Foundation Trust, Salford, UK

[4]Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, USA

[5]Department of Rheumatology, Salford Royal NHS Foundation Trust, Salford, UK

Abstract

Background—Clinical examination of the knee is subject to measurement error. The aim of this analysis was to determine inter- and intra-observer 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 intra-articular steroid therapy. Following standardisation of the clinical test procedures, two 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). Intra-class correlation coefficients(ICC), estimated kappa(κ), weighted kappa(κ^{ω}) and Bland and Altman plots were used to determine inter- and intra-observer levels of agreement.

Results—Using Landis and Koch criteria, inter-observer kappa 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

Address correspondence: Nasimah Maricar, NIHR Clinical Doctoral Fellow, Research in Osteoarthritis Manchester (ROAM), Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UNITED KINGDOM, M13 9PT (nasimah.maricar@postgrad.manchester.ac.uk; nasimah.maricar@srft.nhs.uk).

Data Sharing and Integrity

The corresponding author (NM) had full access to all the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Declarations of interest

None

tenderness(κ=0.76), and effusion assessed by ballottement(κ=0.73) and bulge sign(κ^{ω} =0.78); and excellent for lateral tibiofemoral joint tenderness(κ=1.00), flexion(ICC=0.97) and extension(ICC=0.87) ROM. Intra-observer kappa 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(κ=0.73) tenderness and excellent for effusion assessed by bulge sign(κ^{ω} =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.

Keywords

knee osteoarthritis; clinical tests; inter-observer reliability; intra-observer reliability

Introduction

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[1–3], presence of effusion[4–8] 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 which have formally assessed reliability in the assessment of common clinical signs for knee OA and in those studies that have reported reliability, findings have been somewhat inconsistent[2–4,9–12]. Some contributing factors to the inconsistency include lack of clarity and uniformity in the assessment procedures and also the grading criteria[2–4,9–12]. Reliable clinical assessment is important, as 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 this study was to determine intra- and inter-observer reliability for commonly used clinical tests in the assessment of knee OA.

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 intra-articular 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 hours in the previous 2 weeks or scored greater than 7 out of 32 on the Knee Injury and Osteoarthritis Outcome Score (KOOS), 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 standardised assessment was developed to provide clarity and consistency on the examination procedure. Several patients with knee OA were examined to test the standardised assessment procedure and to resolve issues about the procedure and outcome categorisation. 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 standardised 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 down stroke = 0, a small wave on medial side with down stroke = trace, larger bulge on medial side with down stroke = 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], and assessment of effusion using the ballottement test (absent = 0, present without click = 1, present with click (tap) = 2), patellofemoral joint tenderness (absent = 0, present = 1), pes anserine tenderness (absent = 0, present = 1), medial tibiofemoral joint tenderness (absent = 0, present = 1), lateral tibiofemoral joint tenderness (absent = 0, present = 1) and goniometric knee range of movement (ROM) including 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 can be found in the Appendix.

a) Inter-observer Reliability Assessment—An opportunity sample of twenty-five unselected participants who presented at the screening visit of TASK study was assessed independently by two observers (TON, NM), typically within a 30 to 60 minute interval period 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' 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 for a few times as needed for a consistent recording. For instance, during the performance of bulge sign, the upstroke on the medial aspect of the knee followed by the down stroke on the lateral aspect of the knee, the sequence could be repeated a few times when attempting to observe reappearance of fluid.

b) Intra-observer Reliability Assessment—An opportunity sample of 88 unselected subjects who attended the screening and baseline visits of TASK study was assessed for intra-observer reliability. One assessor (NM) undertook a single repeat clinical assessment of the 88 subjects separated by an interval period 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 inter-observer reliability (compared with intra-observer reliability) that the prevalence of individual examination features may differ.

Analysis

Intra- and inter-observer reliability were assessed using intra-class correlation coefficients (ICC) for continuous variables ICC (2,1) (two-way random effect with rater as random effect)[15], estimated kappa (κ) for dichotomous variables where 2 × 2 contingency tables were used, and weighted kappa (κ^{ω}) (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 them as random or fixed effects made very little difference to the ICC values or their confidence intervals (CI). 2 × 2 tables were used for the determination of estimated kappa values of items scored absent/present such as patellofemoral joint, pes anserine, medial and lateral tibiofemoral joint tenderness, and also for 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 vs absent/unsure while for knee joint crepitus, we dichotomized as either present palpatory/audible crepitus vs absent/unsure. For quadriceps wasting, we dichotomized as present vs 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 either. For the assessment of effusion using the bulge sign where there were 5 possible categories, a weighted kappa was used. For ICC and kappa, values of less than 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 agreement[16]. For continuous data (goniometric knee ROM), Bland and Altman plots were used to determine the limits of agreement (LoA), 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 inter-observer reliability assessment was 63 years (Standard Deviation, SD 10) and 14 (56%) were female. Among these subjects 14% had KL grade 2, 67% had KL grades 3 and 19% KL grade 4. Mean age of the 88 subjects included in the intra-observer 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.

Inter-observer Reliability

Inter-observer kappa scores as assessed by estimated kappa were excellent for the assessment of lateral tibiofemoral joint tenderness ($\kappa = 1.00$), and good for a number of other clinical signs including assessment of bony enlargement, quadriceps wasting, crepitus, medial tibiofemoral joint tenderness, and also the presence of effusion assessed using the bulge sign and ballottement test ($\kappa = 0.66 – 0.78$), see Table 1. Inter-observer estimated kappa scores were moderate for the assessment of patellofemoral joint tenderness and pes

anserine tenderness (κ = 0.48 – 0.53). Intra-class correlations were excellent for the assessment of the degrees of knee flexion and extension (ICC = 0.87 – 0.97) ROM, see 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, −0.12), see 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, see Figure 2. The percentage of raw agreement for all tests was high (> 80%).

Intra-observer Reliability

Intra-observer estimated kappa scores were excellent for bony enlargement, quadriceps wasting, the presence of effusion assessed using the bulge sign, knee flexion and extension ROM (κ = 0.83 – 0.98; ICC = 0.96 – 0.99) and good for the other clinical tests, knee joint crepitus, patellofemoral joint, medial tibiofemoral joint and pes anserine tenderness and the assessment of effusion assessed using ballottement test (κ = 0.64 – 0.78), see Tables 1 and 2. Intra-observer estimated kappa score was moderate for lateral tibiofemoral joint tenderness (κ = 0.60). The intra-observer estimated kappa scores for the clinical tests for knee OA were higher than their respective inter-observer kappa 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, see Figures 3 and 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 intra-observer than inter-observers.

Discussion

In this study, we have shown using a standardized assessment the reliability of commonly used clinical tests for the assessment of knee OA was mostly at least good. As expected, intra-observer reliability of the clinical tests was higher than the inter-observer reliability.

A variety of clinical tests have been used to assess the presence of knee effusion[5,8,18] including both static and dynamic tests though the terminology used in the literature to describe the tests is inconsistent[4–8]. In this study we looked at the reliability of two tests; the bulge sign and also 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 kappa score for inter-observer agreement for the assessment of effusion using the bulge sign (κ^{ω} = 0.78) was higher in magnitude than that reported by Sturgill et al.[4] (κ^{ω} = 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 intra-observer estimated kappa scores, we could only compare the value observed in this analysis with one study that used a 4-point scale (κ^{ω} = 0.35)[3] to assess effusion, in which the kappa 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 kappa value for inter-observer agreement was observed ($\kappa = 0.78$) in comparison to other studies that achieved kappa scores varying from $\kappa = 0.22 – 0.64$[1–3]. Two of these studies[1,2] used a similar grading system (absent, present) while one 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 intra-observer estimated kappa scores for the assessment of knee joint crepitus, we achieved a higher score ($\kappa = 0.78$) than one study[1] ($\kappa = 0.68$ for tibiofemoral crepitus and 0.50 for patellofemoral crepitus) that used a similar grading system (absent, present) and another study[3] ($\kappa = 0.53$) that assessed knee crepitus during sitting to standing movement, though comparable with one other study[2] ($\kappa = 0.78$ for tibiofemoral crepitus and 0.75 for patellofemoral crepitus) in which crepitus was categorized as absent or present.

For the assessment of patellofemoral joint tenderness, the estimated kappa scores for intra- ($\kappa = 0.66$) and inter-observer ($\kappa = 0.53$) were higher than that found in other studies[1,2] who used similar grading of tenderness (absent, present) where their intra-observer and inter-observer estimated kappa scores varied from $\kappa = 0.41 – 0.61$ and $\kappa = 0.27 – 0.35$, respectively. It is possible that the experience or skill of the assessors in the current study may have contributed to the better observer estimated kappa scores. For the assessment of quadriceps wasting and pes anserine tenderness, we reported lower inter-observer estimated kappa scores than that 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 intra- and inter-observer estimated kappa 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. Kappa 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 kappa and so caution is required when comparing kappa values from different studies[20]. Our inter-observer estimated kappa score for bony enlargement ($\kappa = 0.66$) was also higher than two other studies[1,21] ($\kappa = 0.55$ and 0.10, respectively) but lower than Cibere et al.[9] ($R_c = 0.97$) who used a different assessment scale (none, mild, moderate, severe) and assessed bony swelling through palpation rather than in our study through palpation and also visual inspection.

In our analysis there was a high estimated kappa score for inter-observer 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 kappa scores ($\kappa = 0.40 – 0.43$)[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 musculoskeletal disorders of the knee seen in physiotherapy clinics[23,24]. There was no evidence for any statistical 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 post-arthroscopic knee within four 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 inter-observer estimated kappa 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)[25–29] and MRI[30–35]; though estimated kappa scores reported in other US and MRI studies were higher, $\kappa > 0.90$[36–38]. The intra-observer estimated kappa score for bulge sign was also higher than the assessment with US ($\kappa = 0.78$) when repeat examinations were performed on the same day[29]. Similarly, a higher intra-observer estimated kappa score was observed when compared with MRI in some ($\kappa^{\omega} = 0.60 – 0.72$)[30,39] though not all studies[31,33,34].

For most tests, intra-observer estimated kappa scores were higher than inter-observer estimated kappa scores; however, intra-observer estimated kappa scores were lower than inter-observer estimated kappa 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 period between assessments of 32 days for the evaluation of intra-observer estimated kappa scores compared to the same day assessment for inter-observer estimated kappa scores. When data for medial and lateral tibiofemoral joint tenderness were re-analyzed before and after a threshold of 32 days, the intra-observer estimated kappa score for medial tibiofemoral joint tenderness was higher when assessments were made 32 days or less ($\kappa = 0.80$) than when the assessments were more than 32 days apart ($\kappa = 0.71$). For lateral tibiofemoral joint tenderness, no improvement in estimated kappa score was found – though the overall prevalence of lateral tibiofemoral joint tenderness was relatively low and so the results perhaps less reliable.

There are a number of limitations to be considered in interpreting these data. The clinical assessment reported here comprised ten 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 of 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, two experienced assessors examined the subjects; it is unclear if similar findings would be observed with different observers and with different levels of training and experience. In the analysis of intra-observer reliability, subjects were reassessed after an interval period of up to 10 weeks and it is possible true change in disease characteristics may have occurred during this time. The effect of such true change would be if anything, however, to worsen the degree of observer variability. We cannot exclude recall bias in the assessment of intra-observer kappa scores; however, such bias seems unlikely given the interval period between the assessments of 32 days apart [mean 32 days (SD 16.8); min1 to max 75 days], though this cannot be excluded. 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 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 two categories to make them more clinically meaningful though some caution is needed in interpreting the results. Generally there were few instances of uncertainty in findings; for example, in the inter-observer assessment of crepitus, there was only one case of an 'unsure'. We repeated the inter- and intra-observer 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 as specific clinical tests that can differentiate against effusion, muscle atrophy or bony enlargement. Whilst girth or knee circumferential measures may be useful in monitoring changes in knee effusion[40], for instance, during post-operative knee swelling, we do not consider them useful as a one off assessment measure. Further comparison against a 'normal' measure, that is, against a normal knee is required which was not always possible as we included people with bilateral knee OA. Some caution should also be taken due to the small sample size for the inter-observer reliability evaluation with suggestion that future reliability studies to include larger samples. In relation to inter-observer reliability, the order which assessors examined the participants was not randomized or recorded and 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.

In conclusion clinical examination of knee OA is reliable if a standardised 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.

Acknowledgments

This study was 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 decision to submit the manuscript for publication.

This report includes independent research supported by (or funded by) the National Institute for Health Research 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 National Institute for Health Research or the Department of Health. The Research in Osteoarthritis Manchester (ROAM) group is supported by the Manchester Academic Health Sciences Centre (MAHSC). Nasimah Maricar is supported by an NIHR Allied Health Professional Clinical Doctoral Fellowship. The authors would like to acknowledge the equipment and facilities provided by Salford Royal NHS Foundation Trust.

List of Abbreviations

ACR	American College of Rheumatology
CI	Confidence Interval
ICC	Intraclass Correlation Coefficient

κ	Kappa
κ^{ω}	Weighted Kappa
KL	Kellgren Lawrence
KOOS	Knee Injury and Osteoarthritis Outcome Score
LoA	Limits of Agreement
MRI	Magnetic Resonance Imaging
MSK	Musculoskeletal
OA	Osteoarthritis
R_c	Reliability Coefficient
ROM	Range of Movement
SD	Standard Deviation
US	Ultrasound

Reference List

1. Cushnaghan J, Cooper C, Dieppe P, Kirwan J, McAlindon T, McCrae F. Clinical assessment of osteoarthritis of the knee. Ann Rheum Dis. 1990; 49:768–770. [PubMed: 2241265]
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; 51:243–245. [PubMed: 1550411]
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; 59:512–520. [PubMed: 16632140]
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; 39:845–849. [PubMed: 20032559]
5. Davies GJ, Malone T, Bassett FH III. Knee examination. Phys Ther. 1980; 60:1565–1574. [PubMed: 7454781]
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. 3. 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; 50:458–468. [PubMed: 14872488]
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; 44:267–274. [PubMed: 11504260]
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; 33:711–717. [PubMed: 22562715]
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; 33:1139–1143. [PubMed: 23942728]

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. 2015 In Press.
14. Clarkson, HM. Musculoskeletal Assessment: Joint Range of Motion and Manual Muscle Strength. 2. Baltimore: Williams & Wilkins; 1999.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86:420–428. [PubMed: 18839484]
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–310. [PubMed: 2868172]
18. Altman RD. Criteria for classification of clinical osteoarthritis. J Rheumatol Suppl. 1991; 27:10–12. [PubMed: 2027107]
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; 29:1039–1049. [PubMed: 3741515]
20. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993; 46:423–429. [PubMed: 8501467]
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; 50:467–470. [PubMed: 1877852]
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. [PubMed: 17705860]
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. [PubMed: 11245764]
24. Rothstein JM, Miller PJ, Roettger RF. Goniometric reliability in a clinical setting. Elbow and knee measurements Phys Ther. 1983; 63:1611–1615. [PubMed: 6622536]
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. [PubMed: 21470410]
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; 30:673–678. [PubMed: 22765952]
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; 33:879–885. [PubMed: 22811011]
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; 51:2013–2019. [PubMed: 22843774]
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; 67:585–594. [PubMed: 23679910]
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. [PubMed: 23522337]
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; 28:1330–1337. [PubMed: 11409127]
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; 67:206–211. [PubMed: 17472995]

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. [PubMed: 21645627]
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; 31:1347–1357. [PubMed: 22729470]
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; 17:748–753. [PubMed: 19008123]
36. Hauzeur JP, Mathy L, De MV. Comparison between clinical evaluation and ultrasonography in detecting hydrarthrosis of the knee. J Rheumatol. 1999; 26:2681–2683. [PubMed: 10606382]
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; 31:1243–1246. [PubMed: 22526480]
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; 63:2983–2991. [PubMed: 21647860]
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; 17:1562–1569. [PubMed: 19583959]
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; 15:126–134. [PubMed: 20024893]

Appendix: Description of Assessment and Outcome Categories of Clinical Assessment/Tests

a) Bony enlargement

With the patient's knees extended, observation and palpation of the distal end of femur and the proximal end of tibia was made for the presence of enlargement, assessed as either present, absent or unsure. 'Present' was defined as obvious palpatory or visual bony joint enlargement in comparison to the opposite knee or both; 'absent' as no obvious palpatory and no visual bony joint enlargement in comparison to the opposite knee; and 'unsure' when assessors were uncertain or comparison to opposite knee was not possible (example as in the case of bilateral knee OA).

b) Quadriceps Wasting

With the patient's knee extended, observation was made by comparing it with the opposite leg for any apparent reduced muscle bulk of the quadriceps over the anterior aspect of the thigh proximal to the base of the patella, assessed as either present, unsure and absent. 'Present' was defined as obvious reduced quadriceps bulk with the anterior thigh looking flatter or the thigh circumference just proximal to the base of patella appearing smaller; 'possible' when there was a lack of certainty if reduced quadriceps muscle bulk was present; and 'absent' when there was no obvious flattening of the anterior thigh or the circumference of the thigh of the affected limb looking similar to the opposite side.

c) Knee Joint Crepitus

For this test the patient's knee was flexed and extended with the examiner's hand over the anterior aspect of the knee joint and feeling for the presence of any palpatory/audible crepitus anywhere within the knee joint, assessed as present (palpable), present (audible), absent or unsure. 'Present, palpable' was defined as obvious crepitus felt; 'present, audible' as obvious crepitus heard while the knee was moving; 'absent' when there was no crepitus felt or heard as the knee was moving; and 'unsure' when assessors were uncertain if crepitus was present during knee movement. The knee could be extended and flexed for a few times to elicit any crepitus.

d) Tibiofemoral Joint Tenderness

With the knee flexed to about 90°, firm thumb pressure was used to palpate for any tenderness along the tibiofemoral joint line, differentiating tenderness on the medial and lateral side of the joint, assessed as present or absent medial tenderness and present or absent lateral tenderness; repeated as necessary to obtain consistent scoring. For medial joint tenderness, 'present' was defined as obvious tenderness when palpating the medial aspect of the joint. For lateral joint tenderness, 'present' was defined as obvious tenderness when palpating the lateral aspect of the joint line. 'Absent' was defined as no tenderness reported when palpating the medial and lateral joint lines, respectively.

e) Patellofemoral Joint Tenderness

With the knee extended, firm thumb pressure was used to palpate along the medial, lateral, superior and inferior borders of the patella for any tenderness, assessed as present or absent; repeated as necessary to obtain consistent scoring. 'Present' was defined as obvious tenderness when palpating any aspect of the borders of the patella; and 'absent' when the patient reported no tenderness along all borders of patella.

f) Anserine Tenderness

With the knee flexed to about 90°, firm thumb pressure was used to palpate the area of the pes anserine bursa over the anteromedial superior aspect of tibia, about 3–4 fingers distal to the medial joint line, assessed as present or absent; repeated as necessary to obtain consistent scoring. 'Present' was defined as obvious tenderness when palpating around the pes anserine area; 'absent' when patient reported no tenderness around the pes anserine bursa area.

g) Bulge Sign

With the knee extended, starting at the medial gutter, the examiner stroked upwards 2 to 3 times towards the suprapatellar pouch and then stroked downwards on the lateral aspect of the knee joint from the suprapatellar pouch towards the lateral joint-line and observed for any wave of fluid reappearing on the medial side of the knee. The test was repeated for a few times to observe reappearance of fluid. An ordinal scale grading from 0 to 3 was used where 0 was defined as no wave produced on down stroke; "trace" as a small wave on medial side with down stroke; 1 as larger bulge on medial side with down stroke; 2 spontaneously

returned to medial side after upstroke (no down stroke necessary); and 3 as so much fluid that it was not possible to move the effusion out of the medial aspect of the knee[4].

h) Ballottement Test

With the knee extended, using one hand to apply pressure over the suprapatellar pouch squeezing fluid downwards while the thumb and index finger of the opposite hand applied anteroposterior pressure onto the patella, assessed as present without click, present with click (tap) or absent; repeated as necessary to obtain consistent scoring. ‘Present without click’ was defined as balloting of patella, that is, patella moving downwardly and then rebounded upon removing pressure on the patella; ‘present with click’ when palpable click was felt as the patella hit the femur underneath; and ‘absent’ when no rebounding or balloting of the patella occurred and no click was felt.

i) Knee Range of Motion (ROM)

Extension ROM

With the subject lying supine and the knee flexed, the axis of the goniometer was aligned on the lateral aspect of the knee joint with one arm of the goniometer in line with the femur and the other in line with the tibia. Keeping the goniometer in place and the clinician supporting the weight of the limb, the knee was extended as fully as possible with recording of the angles in degrees. The highest of three readings was recorded.

Flexion ROM

With the subject lying supine and the knee extended, the axis of the goniometer was aligned on the lateral aspect of the knee joint with one arm of the goniometer in line with the femur and the other in line with the tibia. Keeping the goniometer in place and the clinician supporting the weight of the limb, the knee was flexed fully with recording of the angles in degrees. The highest of three readings was recorded.

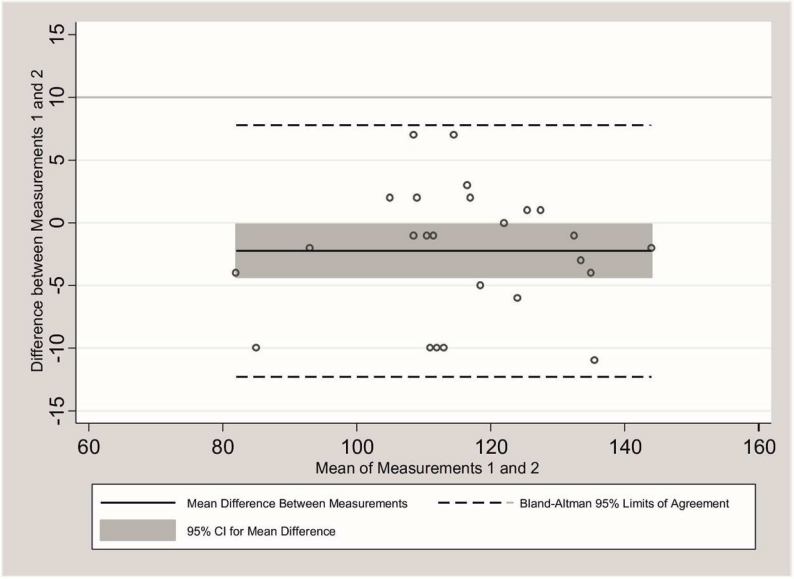

Figure 1.
Bland and Altman Plot – Inter-observer Agreement for Knee Flexion Range of Movement

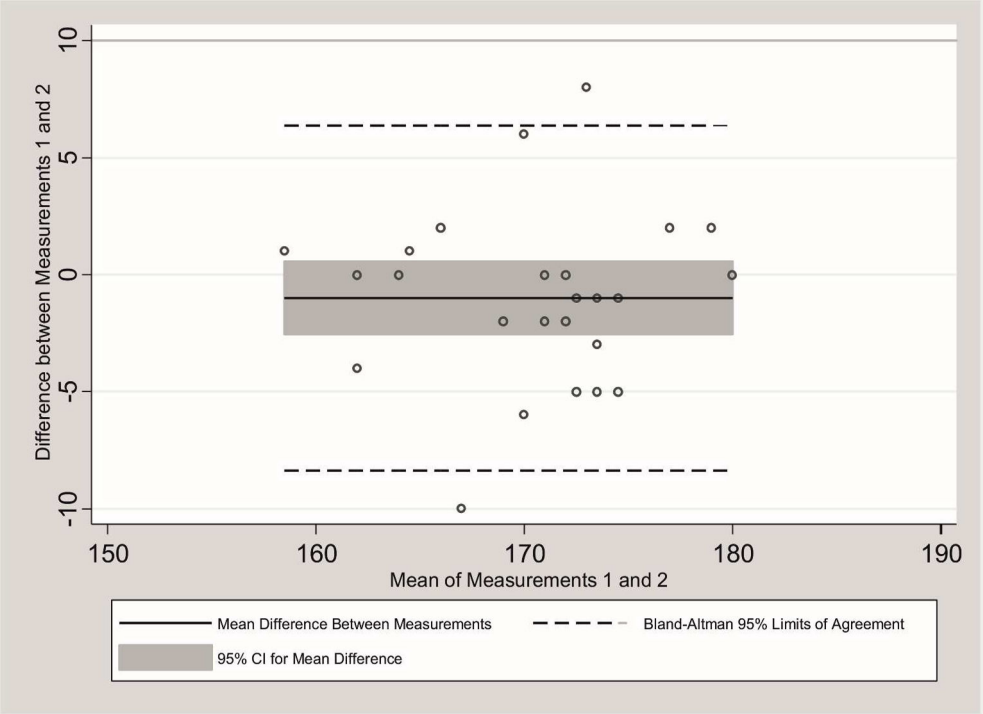

Figure 2.
Bland and Altman Plot – Inter-observer Agreement for Knee Extension Range of Movement

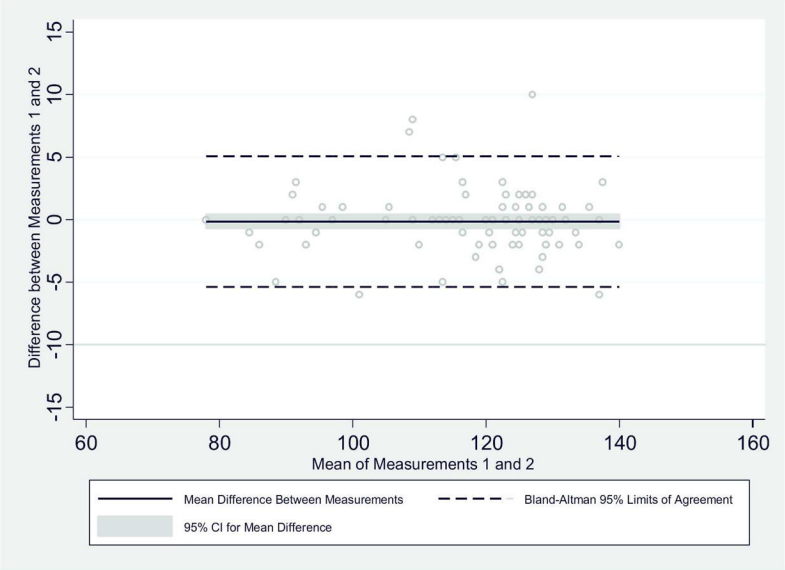

Figure 3.
Bland and Altman Plot – Intra-observer Agreement for Knee Flexion Range of Movement

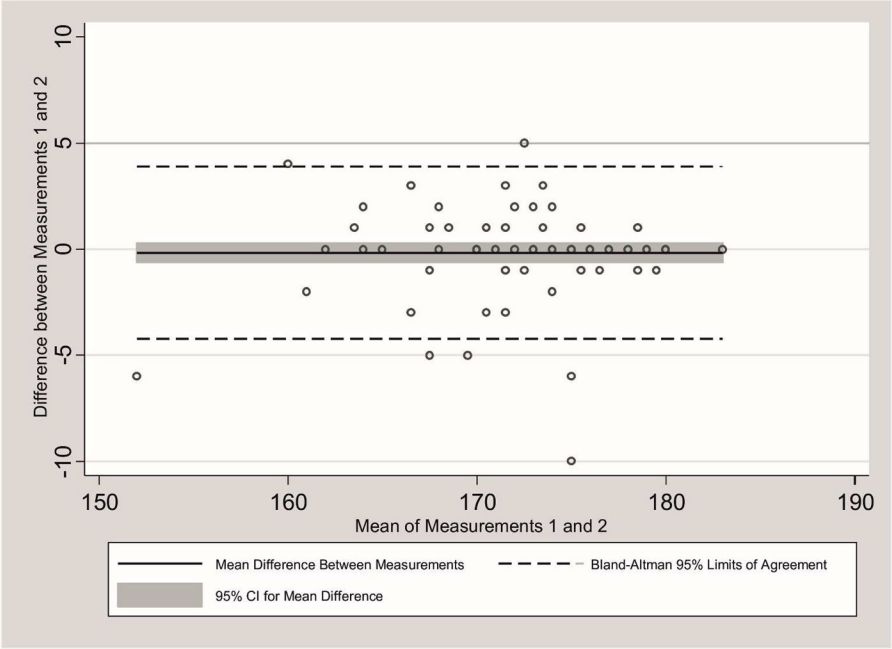

Figure 4.
Bland and Altman Plot – Intra-observer Agreement for Knee Extension Range of Movement

Table 1

Inter-Observer and Intra-Observer Reliability of Clinical Tests for Knee OA

	Inter-observer Reliability					Intra-observer Reliability			
Clinical Evaluation	**Outcome Category**	**Prevalence^**	**Kappa Values**	**95% CI**	**% Raw Agreement**	**Prevalence^**	**Kappa Values**	**95% CI**	**% Raw Agreement**
Bony Enlargement	present vs absent/unsure	0.22	0.66	0.32 to 1.00	88.0	0.37	0.98	0.93 to 1.00	98.9
Quadriceps Wasting	present vs absent/ possible present palpable/audible vs	0.24	0.78	0.40 to 1.00	84.0	0.62	0.83	0.72 to 0.95	90.9
Knee Joint Crepitus	absent/unsure	0.90	0.78	0.36 to 1.00	96.0	0.91	0.78	0.55 to 1.00	96.6
Medial Tibiofemoral Joint Tenderness	present vs absent	0.54	0.76	0.50 to 1.00	88.0	0.48	0.64	0.49 to 0.80	81.8
Lateral Tibiofemoral Joint Tenderness	present vs absent	0.28	1.00	1.00 to 1.00	100.0	0.22	0.60	0.39 to 0.80	86.4
Patellofemoral Joint Tenderness	present vs absent	0.30	0.53	0.16 to 0.89	80.0	0.36	0.66	0.60 to 0.92	84.1
Anserine Tenderness	present vs absent	0.26	0.48	0.09 to 0.87	80.0	0.22	0.73	0.61 to 0.99	90.9
Effusion : Bulge Sign	5-point Likert scale	0.96	0.78*	0.55 to 1.00	80.0	0.64	0.83*	0.73 to 0.94	85.2
Effusion : Ballottement Test▰	present with/without click vs absent	0.66	0.73	0.45 to 1.00	88.0	0.18	0.77	0.60 to 0.95	93.2

Legends: OA – osteoarthritis; CI – confidence interval;

*Weighted kappa;

^Prevalence was calculated as the average of positive findings between the two rated scores;

▰Ballottement test was defined as positive click/tap or downward movement of the patella on pressure and rebounding of patella upon removal of pressure

Table 2

Inter-Observer and Intra-Observer Reliability for Measurement of Passive Knee Range of Movement

	Inter-observer Agreement				Intra-observer Agreement			
Knee PROM	**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, −0.12)	−12.29, 7.81	1.03	0.99 (0.99, 0.99)	−0.14 (−0.43, 0.70)	−5.37, 5.10	0.28
Extension	0.87 (0.72, 0.94)	−1.00 (−2.55, 0.55)	−8.38, 6.38	0.75	0.96 (0.94, 0.98)	−0.17 (−0.61, 0.27)	−4.23, 3.89	0.22

Legends: ICC – intra-class correlation coefficient; CI – confidence interval; PROM –passive range of movement; LoA – limits of agreement; SEM – standard error of measurements