Published in final edited form as:

J Rheumatol. 2016 July ; 43(7): 1421–1426. doi:10.3899/jrheum.151300.

Reliability and accuracy of cross-sectional radiographic assessment of severe knee osteoarthritis: role of training and experience

Kristina Klara, BS¹, Jamie E. Collins, PhD^{1,3}, Ellen Gurary, MS^{1,2}, Scott A. Elman, BA^{1,3}, Derek S. Stenquist, BA^{1,3}, Elena Losina, PhD^{1,2,3}, and Jeffry N. Katz, MD, MS^{1,3,4}

¹Orthopaedic and Arthritis Center for Outcomes Research, Department of Orthopedic Surgery, Brigham and Women's Hospital, Boston, MA

²Department of Biostatistics, Boston University School of Public Health, Boston, MA

³Harvard Medical School, Boston, MA

⁴Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA

Abstract

Objective—To determine the reliability of radiographic assessment of knee osteoarthritis (OA) by non-clinician readers compared to an experienced radiologist.

Methods—The radiologist trained three non-clinicians to evaluate radiographic characteristics of knee OA. The radiologist and non-clinicians read preoperative films of 36 patients prior to total knee replacement. Intra- and inter-reader reliability was measured using the weighted kappa statistic and intra-class correlation coefficient (ICC). Kappa <0.20 indicated slight agreement, 0.21 – 0.40 fair, 0.41 – 0.60 moderate, 0.61 – 0.80 substantial, and 0.81 – 1.0 almost perfect agreement.

Results—Intra-reader reliability among non-clinicians (kappa) ranged from 0.40 to 1.0 for individual radiographic features and 0.72 to 1.0 for Kellgren-Lawrence (K-L) grade. ICCs ranged from 0.89 to 0.98 for the Osteoarthritis Research Society International (OARSI) Summary Score. Inter-reader agreement among non-clinicians ranged from kappa of 0.45 to 0.94 for individual features, and 0.66 to 0.97 for K-L grade. ICCs ranged from 0.87 to 0.96 for the OARSI Summary Score. Inter-reader reliability between non-clinicians and the radiologist ranged from kappa of 0.56 to 0.85 for KL grade. ICCs ranged from 0.79 to 0.88 for the OARSI Summary Score.

Conclusion—Intra- and inter-rater agreement was variable for individual radiograph features but substantial for summary KL grade and OARSI Summary Score. Investigators face trade-offs between cost and reader experience. These data suggest that in settings where costs are constrained, trained non-clinicians may be suitable readers of radiographic knee OA, particularly if a summary score (K-L grade or OARSI Score) is used to capture radiographic severity.

Correspondence: Jeffrey N. Katz, M.D., M.S. Director, Orthopaedic and Arthritis Center for Outcomes Research (OrACORe), Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, 75 Francis Street, BC 4-016, Boston, MA 02115, Telephone: 617-732-5338, Fax: 617-525-7900, jnkatz@partners.org.

No other financial interests or disclosures.

Keywords

knee osteoarthritis; radiography; reliability; inter-reader reliability

Introduction

Knee osteoarthritis (OA) is characterized by degenerative changes in cartilage, bone, meniscus and other joint structures, in concert with pain, stiffness, and functional loss. The severity of structural damage in knee OA can be assessed by radiographic evidence of joint space narrowing (JSN) and osteophyte formation ^{1,2}. In clinical research, knee OA is often staged by radiologists or other physicians using ordinal scales such as the Kellgren-Lawrence (K-L) grade or the Osteoarthritis Research Society International (OARSI) Score. Employing experienced clinicians to grade radiographic OA in research settings can be expensive, raising the question of whether non-radiologist readers can be trained to provide reliable, valid readings.

Several studies have measured variability in radiographic assessment of knee osteoarthritis by clinicians ³⁻⁶. Intra- and inter-reader reliability of the K-L score varies widely across studies, with weighted kappas ranging from 0.26 to 0.88 and 0.56 to 0.80, respectively ⁴⁻⁶. One prior study examined the inter-reader reliability of radiographic assessment of knee osteoarthritis between non-clinician readers and an experienced clinical reader. This study documented kappa statistics for radiographic features of tibiofemoral OA ranging from 0.12 to 0.80, suggesting this approach merits further research ⁷.

In the current study, we aimed to determine the inter- and intra-reader reliability of radiographic assessment of severe knee OA among three junior non-clinician readers and to assess the agreement between their readings and those of an experienced radiologist.

Methods

Study Population

The data presented in this report were collected as part of the Adding Value in Knee Arthroplasty (AViKA) Postoperative Care Navigation study, a randomized controlled trial conducted at Brigham and Women's Hospital in Boston, Massachusetts. The trial prospectively evaluated a behavioral intervention to optimize postoperative outcomes following primary total knee replacement (TKR). Three hundred and nine patients 40 years old with a primary diagnosis of osteoarthritis who underwent primary total knee replacement (TKR) were enrolled ⁸. We chose 39 AViKA participant x-rays at random for this study. Demographic information for the subjects analyzed is presented in Table 1. This study was approved by the Partners Healthcare Institutional Review Board (protocol 2010P002597).

Training

Two non-clinician readers (medical students) studied the OARSI Atlas of Individual Radiographic Features in Osteoarthritis ² to learn to grade osteophytes and JSN on standing

bilateral radiographs. They spent five hours reading films before attending 2 hour-long training sessions with an experienced radiologist who has graded knee OA for several cohort studies and trials. The training sessions for the medical students took place before a research assistant was included in the study. Due to feasibility concerns (limited study funding and time constraints of the radiologist), the medical students provided the initial training for the research assistant to assess radiographic features of osteoarthritis in 3 hour-long sessions. The three non-clinician readers (2 medical students and 1 research assistant together) had a final training session with the radiologist before reading films for reliability analyses.

Data Collection Procedures

The non-clinician readers and the radiologist viewed 39 standing bilateral preoperative radiographs in Centricity Web Version 3.0 and graded individual features of knee osteoarthritis, blinded to the ratings of other readers. Three subjects were excluded from the analysis because readers inadvertently graded different x-ray views. For each of the 36 subjects included in the analysis, we examined both the left and right knees; however, we were unable to grade radiographic features of OA in knees with implants. Six out of 36 subjects analyzed had one knee replaced before enrolling in AViKA. Thus, out of 72 knees (36×2), we were able to analyze 66 knees ($72 - 6$) for inter-reader reliability analyses. Standing bilateral films were used to assess tibiofemoral features of OA (preferably the posterioranterior (PA) view or the anteroposterior view if PA was unavailable). These clinical radiographic protocols did not use a standardized positioning device. Sunrise views were used to grade patellofemoral features. Radiology technicians routinely assess image quality and repeat images that are inadequate; therefore, all images used for this reliability analysis were acceptable.

In order to assess intra-reader reliability, 2-3 weeks following their initial reading, two non-clinician readers re-graded 17 radiographs (31 knees), and one non-clinician reader re-graded 36 radiographs (66 knees). We analyzed left and right knees separately, with one film per subject. Examining all knees in the same analysis would have created the potential for clustering of observations between 2 knees of the same subject, requiring a less transparent analysis.

Radiographic Measures

Anatomic alignment (in degrees) was defined as the angle formed by the intersection of a line drawn from the intercondylar notch to the center of the femoral shaft and another line drawn from the center of the tibial spines to the center of the tibial shaft. The angle of anatomic alignment and direction of deformity (varus or valgus) were documented. The deformity was considered varus when the tibia was angled inward with respect to the femur and valgus when the tibia was angled outward with respect to the femur. The length of tibia and femur available to measure on each study was not standardized. Osteophytes and JSN were graded on a 4-point scale (0-3) as per OARSI guidelines². Grade 0 was considered normal. Grade 1 indicated mild osteophytes or narrowing, grade 2 moderate osteophytes or narrowing, and grade 3 marked and severe osteophytes or narrowing.

Osteophytes were assessed in the lateral femoral, lateral tibial, medial femoral, and medial tibial compartments. Medial and lateral patellofemoral osteophytes were assessed, but only the largest score was recorded. Lateral and medial tibiofemoral JSN were graded in addition to patellofemoral JSN. Joint space width (JSW) was measured in millimeters (mm) at the narrowest point in each of these compartments. The radiologist graded osteophytes and JSN only.

We used individual tibiofemoral osteophyte and JSN scores to generate a Kellgren-Lawrence grade and OARSI Summary Score for each knee; patellofemoral osteophytes and JSN are not included in these summary scores. If the highest osteophyte score was 0 and the highest JSN score was 0 or 1, we considered the knee K-L 0. The knee was K-L 1 if the highest osteophyte score was 1 and the highest JSN score was 0 or 1. The knee was K-L 2 if the highest JSN score was 0 or 1 and the highest osteophyte score was 2. The knee was K-L 3 if the highest JSN score was 2, regardless of osteophyte scores. The knee was K-L 4 if the highest JSN score was 3, regardless of osteophyte scores. We determined the OARSI Summary Score by adding all tibiofemoral osteophyte and JSN scores.

Statistical Analysis

We define *intra-reader* agreement as a measurement of how consistent a reader is with respect to him-or herself. *Inter-reader* agreement is a measure of how consistent a reader is with respect to other readers. To calculate intra-reader agreement, we compared the first and second reads of the same x-rays by the same reader. To calculate inter-reader agreement, we compared each reader's first read of the same x-rays. To assess agreement on categorical variables, we used weighted kappa coefficients. Weighted kappa scores take into account the ordering of the categorical levels of a variable⁹. A weighted kappa score that evaluates agreement between readers *A* and *B* is calculated through a matrix in which reader *A*'s ratings are arrayed in the rows, and reader *B*'s ratings in the columns. The diagonal shows the completely correct agreement. A weight is assigned to the agreements of the off-diagonals reflecting the severity of disagreement, which increases as the levels chosen are further apart. These Cicchetti-Allison weights are linear and calculated as follows:

$w_{ij} = 1 - \frac{|i - j|}{K - 1}$, where *i* is the level of Rater 1 and *j* is the level of Rater 2, and *K* is the number of levels of the variable⁹. Kappas range between 0 and 1; weighted kappa <0.20 indicates slight agreement; 0.21 – 0.40 fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 substantial agreement; and 0.81 – 1.0 almost perfect agreement¹⁰.

To assess agreement on continuous variables, we used intraclass correlation coefficients (ICCs) which quantify the similarity of grouped measures between observers. We used the

fixed Shrout-Fleiss ICC. The ICC is estimated by $\frac{BMS - EMS}{BMS}$ where BMS is between-subject mean square or the variation explained by the differences between subjects and EMS is residual mean square, or the variance leftover¹¹. The ICC ranges from -1 to 1; values closer to 0 indicate weaker reliability, while values closer to -1 and 1 indicate stronger reliability. The sample size was chosen to provide reasonable precision as reflected in 95% confidence intervals around the estimates of kappa.

Results

Gold Standard Comparison

The agreement in readings of the gold standard radiologist and the non-clinician readers for individual radiographic features of tibiofemoral OA was fair to substantial, with kappa statistics ranging from 0.39 to 0.76 (Table 2). Agreement was generally higher for joint space narrowing than for osteophytes and for the tibiofemoral joint structures than for the patellofemoral joint. Inter-reader reliability for K-L scores was moderate to almost perfect, with kappa statistics ranging from 0.56 to 0.85 (Table 2). The OARSI Summary Score showed excellent inter-reader agreement between the radiologist and non-clinician readers, with ICCs ranging from 0.79 to 0.88 (Table 3).

Intra-reader Reliability among Non-clinician Readers

The intra-reader reliability among non-clinicians for individual radiographic features of tibiofemoral OA was fair to almost perfect, with kappa statistics ranging from 0.40 to 1.0. Here too, agreement was generally higher for joint space narrowing than for osteophytes and for tibiofemoral than for patellofemoral joints. The K-L score showed substantial to almost perfect intra-reader agreement, with kappa statistics ranging from 0.72 to 1.0. Intra-reader reliability for alignment (varus or valgus) was substantial to almost perfect, with kappa statistics ranging from 0.87 to 1.0 (Table 2).

The OARSI Summary Score showed excellent intra-reader agreement, with ICCs ranging from 0.89 to 0.98 (Table 3). Intra-reader reliability of knee angle measurement (in degrees) was also excellent, with ICCs ranging from 0.92 to 0.98. The ICCs for lateral JSW ranged from 0.82 to 0.96 and for medial JSW, ranged from 0.94 to 1.0 (Table 3).

Inter-reader Reliability among Non-clinician Readers

The inter-reader reliability among non-clinician readers for individual radiographic features of tibiofemoral OA was moderate to almost perfect, with kappa statistics ranging from 0.45 to 0.94. Inter-reader agreement for K-L scores was substantial to almost perfect, with kappa statistics ranging from 0.66 to 0.97 (Table 2). Kappa statistics for inter-reader reliability of alignment (varus or valgus) ranged from 0.76 to 0.93 (Table 2).

The OARSI Summary Score showed excellent inter-reader reliability, with ICCs ranging from 0.87 to 0.96 (Table 3). Inter-reader agreement of anatomic alignment angle (in degrees) was also excellent, with the ICCs ranging from 0.88 to 0.94. ICCs ranged from 0.83 to 0.92 for lateral JSW and from 0.93 to 0.96 for medial JSW (Table 3).

Patellofemoral Findings

Patellar osteophytes and patellofemoral JSN are not addressed in the OARSI atlas ². As such, our analyses of these features are exploratory in nature. The inter-reader reliability between the radiologist and non-clinician readers for patellofemoral features of OA was slight to substantial, with kappa statistics ranging from 0.17 to 0.79. The intra- and inter-reader agreement among non-clinician readers for patellofemoral features was better, with

kappa statistics ranging from 0.43 to 0.84 for intra-reader agreement and 0.28 to 0.83 for inter-rater agreement (Table 2).

Discussion

The current study examines the intra- and inter-reader reliability of radiographic assessment of knee OA between three non-clinicians and measures the agreement between their readings and those of an experienced radiologist. Inter-reader agreement between non-clinician readers and the gold standard radiologist was moderate to almost perfect for the K-L score and excellent for the OARSI Summary Score. There was substantial to almost perfect intra-reader reliability among non-clinician readers for the K-L and OARSI Summary scores, and inter-reader agreement between them was substantial to almost perfect for these summary measures.

Of the individual radiographic features of OA, osteophytes showed lower agreement than JSN. The agreement for patellar osteophytes and patellofemoral JSN was lowest, possibly because patellar features are not addressed in the OARSI atlas ². While agreement between the radiologist and non-clinicians was moderate to almost perfect for the K-L and OARSI Summary scores, reliability varied widely for individual radiographic characteristics, depending on the feature considered.

Agreement for K-L grade was more variable than agreement for the OARSI Summary Score. The K-L classification has been criticized for its sensitivity to osteophyte size ¹², which we found to be less reliable than JSN. The OARSI Summary Score, which sums osteophyte and JSN scores in all but the patellofemoral compartment, is less affected by one-point differences in individual radiologic features. Therefore, we suggest that the OARSI Summary Score may be a more robust and reliable measure of OA severity than the K-L score.

Several prior studies have examined the reliability of radiographic measurement of OA. Spector et al. reported on variability in the assessment of knee osteoarthritis in a longitudinal female cohort study. Kappa coefficients for intra- and inter- reader agreement for K-L scores ranged from 0.66 to 0.88 and from 0.56 to 0.80, respectively ⁵. In a study conducted by Gossec et al., three rheumatologists graded 50 standing radiographs to assess inter-reader reliability. Kappa statistics for inter- and intra-reader agreement for K-L grade were 0.56 and 0.61, respectively ⁶. Agreement between rheumatologists in Gossec et al. and between the readers in Spector et al. is comparable to the agreement observed between non-clinicians and the gold standard read in the current study.

Riddle et al. examined the reliability of radiographic assessment of knee OA between two experienced and two inexperienced orthopedic surgeons for 116 patients in the Osteoarthritis Initiative (OAI), a multi-center study of patients who have or are at risk for osteoarthritis. They assessed the validity of their readings by comparison to the gold standard, an adjudicated reading by experienced radiologists in the OAI. Two central readers at the OAI evaluated radiographic knee OA at baseline, and if they disagreed on a particular score, a third reader reviewed the film. If the third reviewer agreed with either of the original readers,

that score was final. If the third reviewer did not agree with either of the original readers, the three readers came to a consensus score together. Comparison to the gold standard K-L grade was fair to substantial, with weighted kappa statistics ranging from 0.36 to 0.80⁴. In another study from the OAI, Guermazi et al. assessed the reliability between the central and site-specific readings and reported kappa statistics for inter-reader agreement for lateral and medial JSN were 0.65 and 0.71, respectively and 0.37 for osteophytes. Inter-reader agreement for K-L grade was moderate, with kappa equaling 0.52.¹³ The findings of these two OAI reliability studies resemble ours and provide further evidence that reliability even among experienced readers is generally modest.

Many reliability studies are conducted using an experienced clinician reader as the gold standard. In contrast, Sheehy et al. compared assessment of radiographic knee OA to assessment via MRI as the gold standard. They found that K-L grading, OARSI JSN scoring, and the compartmental grading scale for OA correlated well with MRI findings, with correlation coefficients equaling 0.836, 0.840, and 0.773, respectively.¹⁴

To our knowledge, only one study measuring the reliability of radiographic assessment of knee OA among non-clinician readers has been conducted. In a cohort of patients with early symptomatic knee OA (K-L 0 or K-L 1 at baseline), Damen et al. assessed the inter-reader reliability between four research assistants and a general practitioner (GP) experienced in grading knee OA. The average agreement for K-L grade 1 between the non-clinicians and GP was moderate, with kappa equal to 0.58. Average kappa statistics for individual radiographic features, graded based on the OARSI atlas, ranged from 0.12 to 0.80. Agreement between the GP reader and non-clinician readers is similar to that observed between non-clinicians and the radiologist in this analysis. In contrast to the current study, Damen et al. did not report intra- or inter-reader reliability between the non-clinician readers, nor do they report agreement for the OARSI Summary Score⁷.

The current study has certain limitations. First, the sample size was low. Additionally, all study subjects underwent TKR and had moderate to advanced radiographic OA. Fifty-five of 66 knees were classified as K-L 4 by the experienced radiologist, and our results should be generalized cautiously to a population-based sample or to a population with less severe, low-grade disease. In contrast, the study conducted by Damen et al. assessed the reliability of non-clinician readers to grade early knee OA (K-L 0 and K-L 1 at baseline) and found similar agreement to our analysis⁷; the study by Damen et al. may be useful in evaluating the reliability of non-clinician assessment of early knee OA.

Another limitation of the current study is that it was not longitudinal and did not take into account the ability of non-clinician readers to evaluate OA progression. Therefore, we were unable to evaluate the sensitivity and specificity of assessment of structural changes over time. Moreover, while the non-clinician readers participated in the same protocol, they trained in two waves, which may have led to minor departures from uniformity in training. Lastly, short view x-rays were read for this study, as full-length radiographs were unavailable. Thus, anatomic alignment and deformity are based on the anatomical and not mechanical axis. Similarly, the study radiographs were done for clinical and not research purposes. For example, this study used clinical protocols that did not involve standardized

positioning devices designed to give an MTP view. This limitation should not influence reliability since the different readers used a single image; all readers were exposed to the same image; and the readers did not assess longitudinal change.

These results have significant implications for research on populations with severe osteoarthritis. It can be expensive to employ radiologists to grade knee OA in research settings. Even among clinician readers, reliability varies³⁻⁶. The current study highlights trade-offs involved in using trained non-clinician readers. The findings show that non-clinician readers assess severe knee OA features with a level of reliability that may be acceptable for certain study settings. The balance between cost and reader experience must be weighed carefully, and our data will help in this regard. Because the current study focused on a population with advanced knee OA, further research on the reliability of non-clinician assessment of knee OA in a population of subjects with ranging severity is needed. Future studies may also focus on enhancing training for non-clinicians to improve their agreement with expert readers. Additional sessions with the radiologist reader in this study as well as more independent practice prior to assessing knee OA for the reliability analysis may have improved the accuracy of non-clinician assessment of OA. Still, our results are in line with reliability studies conducted by experienced clinicians, suggesting that radiographic characterization of knee OA is inherently subjective.

Acknowledgments

We thank Piran Aliabadi, MD for training the non-clinician readers and for reading x-ray films to provide the gold standard read in this analysis.

Supported by: NIH/NIAMS K24AR057827, T32AR055885

References

1. Felson DT, Lawrence RC, Dieppe PA, Hirsch R, Helmick CG, Jordan JM, et al. Osteoarthritis: new insights. Part 1: the disease and its risk factors. *Ann Intern Med.* 2000; 133:635–646. [PubMed: 11033593]
2. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis Cartilage.* 2007; 15(Suppl A):A1–56. [PubMed: 17320422]
3. Sun Y, Gunther KP, Brenner H. Reliability of radiographic grading of osteoarthritis of the hip and knee. *Scand J Rheumatol.* 1997; 26:155–165. [PubMed: 9225869]
4. Riddle DL, Jiranek WA, Hull JR. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons. *Orthopedics.* 2013; 36:e25–32. [PubMed: 23276348]
5. Spector TD, Hart DJ, Byrne J, Harris PA, Dacre JE, Doyle DV. Definition of osteoarthritis of the knee for epidemiological studies. *Ann Rheum Dis.* 1993; 52:790–794. [PubMed: 8250610]
6. Gossec L, Jordan JM, Mazucca SA, Lam MA, Suarez-Almazor ME, Renner JB, et al. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI-OMERACT task force. *Osteoarthritis Cartilage.* 2008; 16:742–748. [PubMed: 18417373]
7. Damen J, Schiphof D, Wolde ST, Cats HA, Bierma-Zeinstra SM, Oei EH. Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study. *Osteoarthritis Cartilage.* 2014; 22:969–974. [PubMed: 24857977]
8. Losina E, Collins JE, Daigle ME, Donnell-Fink LA, Prokopetz JJ, Strnad D, et al. The AViKA (Adding Value in Knee Arthroplasty) postoperative care navigation trial: rationale and design features. *BMC Musculoskelet Disord.* 2013; 14:290. [PubMed: 24119081]

9. Vanbelle S, Albert A. A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*. 2009; 6:157–163.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
11. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979; 86:420–428. [PubMed: 18839484]
12. Spector TD, Cooper C. Radiographic assessment of osteoarthritis in population studies: whither Kellgren and Lawrence? *Osteoarthritis Cartilage*. 1993; 1:203–206. [PubMed: 15449506]
13. Guermazi A, Hunter DJ, Li L, Benichou O, Eckstein F, Kwok CK, et al. Different thresholds for detecting osteophytes and joint space narrowing exist between the site investigators and the centralized reader in a multicenter knee osteoarthritis study--data from the Osteoarthritis Initiative. *Skeletal Radiol*. 2012; 41:179–186. [PubMed: 21479521]
14. Sheehy L, Culham E, McLean L, Niu J, Lynch J, Segal NA, et al. Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the Multicenter Osteoarthritis Study (MOST). *Osteoarthritis Cartilage*. 2015; 23:1491–1498. [PubMed: 26003948]

Table 1**Cohort Characteristics**

Characteristic	Mean (SD) / N (%)	
Mean age (years)	66.8 (7.4)	
Sex		
Female	16 (44.4%)	
Male	20 (55.6%)	
Index Knee		
Right	16 (44.4%)	
Left	20 (55.6%)	
K-L Score[/]	Left	Right
0	1 (3.0%)	1 (3.0%)
1	2 (6.1%)	4 (12.1%)
2	3 (9.1%)	0 (0%)
3	0 (0%)	0 (0%)
4	27 (81.8%)	28 (84.9%)

[/] K-L score is reported as determined by the gold standard read.

Table 2

Inter- and Intra-reader Reliability for Categorical Variables. Weighted kappa (95% Confidence Interval; % agreement).

Variable	Comparison	Reader One ^I		Reader Two ^I		Reader Three ^I	
		Left Knee	Right Knee	Left Knee	Right Knee	Left Knee	Right Knee
K-L Score	Gold Standard ²	0.71 (0.52, 0.9; 73%)	0.56 (0.39, 0.73; 55%)	0.85 (0.72, 0.97; 85%)	0.70 (0.56, 0.85; 73%)	0.81 (0.68, 0.95; 82%)	0.63 (0.49, 0.77; 64%)
	Reader One	0.72 (0.54, 0.89; 79%)	0.74 (0.58, 0.9; 76%)	0.78 (0.63, 0.94; 82%)	0.77 (0.62, 0.92; 76%)	0.75 (0.58, 0.93; 82%)	0.66 (0.49, 0.83; 67%)
	Reader Two	0.78 (0.63, 0.94; 82%)	0.77 (0.62, 0.92; 76%)	1 (1, 1; 100%)	1 (1, 1; 100%)	0.97 (0.91, 1; 97%)	0.83 (0.69, 0.97; 85%)
	Reader Three	0.75 (0.58, 0.93; 82%)	0.66 (0.49, 0.83; 67%)	0.97 (0.91, 1; 97%)	0.83 (0.69, 0.97; 85%)	1 (1, 1; 100%)	1 (1, 1; 100%)
Alignment (varus or valgus)	Reader One	0.94 (0.82, 1; 97%)	0.93 (0.8, 1; 97%)	0.76 (0.54, 0.97; 88%)	0.93 (0.8, 1; 97%)	0.88 (0.72, 1; 94%)	0.8 (0.59, 1; 91%)
	Reader Two	0.76 (0.54, 0.97; 88%)	0.93 (0.8, 1; 97%)	1 (1, 1; 100%)	0.87 (0.64, 1; 93%)	0.87 (0.71, 1; 94%)	0.87 (0.7, 1; 64%)
	Reader Three	0.88 (0.72, 1; 94%)	0.8 (0.59, 1; 91%)	0.87 (0.71, 1; 94%)	0.87 (0.7, 1; 64%)	1 (1, 1; 100%)	1 (1, 1; 100%)
Lateral Femoral Osteophyte	Gold Standard ²	0.5 (0.3, 0.7; 52%)	0.45 (0.26, 0.63; 52%)	0.49 (0.27, 0.71; 49%)	0.43 (0.2, 0.65; 55%)	0.53 (0.34, 0.73; 55%)	0.39 (0.19, 0.58; 55%)
	Reader One	0.58 (0.4, 0.76; 58%)	0.63 (0.43, 0.82; 70%)	0.59 (0.38, 0.81; 67%)	0.62 (0.42, 0.82; 67%)	0.61 (0.41, 0.8; 67%)	0.45 (0.24, 0.66; 55%)
	Reader Two	0.59 (0.38, 0.81; 67%)	0.62 (0.42, 0.82; 67%)	0.4 (-0.02, 0.82; 69%)	0.42 (0.12, 0.71; 53%)	0.72 (0.56, 0.88; 73%)	0.64 (0.46, 0.82; 67%)
	Reader Three	0.61 (0.41, 0.8; 67%)	0.45 (0.24, 0.66; 55%)	0.72 (0.56, 0.88; 73%)	0.64 (0.46, 0.82; 67%)	1 (1, 1; 100%)	1 (1, 1; 100%)
Lateral Joint Space Narrowing	Gold Standard ²	0.57 (0.35, 0.8; 61%)	0.65 (0.43, 0.88; 70%)	0.71 (0.51, 0.91; 76%)	0.63 (0.41, 0.85; 64%)	0.71 (0.53, 0.89; 73%)	0.75 (0.6, 0.91; 73%)
	Reader One	0.76 (0.57, 0.96; 82%)	0.69 (0.46, 0.93; 76%)	0.79 (0.61, 0.97; 82%)	0.65 (0.4, 0.9; 76%)	0.86 (0.72, 1; 88%)	0.73 (0.52, 0.93; 79%)
	Reader Two	0.79 (0.61, 0.97; 82%)	0.65 (0.4, 0.9; 76%)	0.81 (0.58, 1; 88%)	0.68 (0.37, 0.99; 73%)	0.93 (0.84, 1; 93%)	0.81 (0.62, 0.99; 85%)
	Reader Three	0.86 (0.72, 1; 88%)	0.73 (0.52, 0.93; 79%)	0.93 (0.84, 1; 93%)	0.81 (0.62, 0.99; 85%)	0.92 (0.76, 1; 94%)	0.92 (0.75, 1; 93%)
Lateral Tibial Osteophyte	Gold Standard ²	0.46 (0.24, 0.68; 55%)	0.57 (0.38, 0.76; 64%)	0.53 (0.34, 0.72; 58%)	0.62 (0.45, 0.79; 64%)	0.46 (0.27, 0.66; 52%)	0.67 (0.49, 0.85; 70%)

Variable	Comparison	Reader One ^J			Reader Two ^J			Reader Three ^J		
		Left Knee	Right Knee		Left Knee	Right Knee		Left Knee	Right Knee	
	Reader One	0.64 (0.47, 0.82; 67%)	0.65 (0.49, 0.81; 70%)		0.81 (0.65, 0.97; 85%)	0.63 (0.47, 0.79; 70%)		0.73 (0.54, 0.92; 79%)	0.61 (0.46, 0.77; 67%)	
	Reader Two	0.81 (0.65, 0.97; 85%)	0.63 (0.47, 0.79; 70%)		0.45 (0.13, 0.78; 69%)	0.73 (0.48, 0.98; 80%)		0.92 (0.8, 1; 94%)	0.74 (0.58, 0.9; 79%)	
	Reader Three	0.73 (0.54, 0.92; 79%)	0.61 (0.46, 0.77; 67%)		0.92 (0.8, 1; 94%)	0.74 (0.58, 0.9; 79%)		0.87 (0.65, 1; 94%)	0.9 (0.72, 1; 93%)	
Medial Femoral Osteophyte	Gold Standard ²	0.73 (0.56, 0.9; 73%)	0.53 (0.33, 0.73; 58%)		0.61 (0.45, 0.77; 61%)	0.54 (0.33, 0.74; 61%)		0.61 (0.41, 0.81; 67%)	0.44 (0.22, 0.65; 55%)	
	Reader One	0.56 (0.37, 0.74; 61%)	0.52 (0.31, 0.74; 64%)		0.65 (0.5, 0.81; 64%)	0.73 (0.56, 0.89; 76%)		0.55 (0.38, 0.71; 52%)	0.75 (0.56, 0.94; 82%)	
	Reader Two	0.65 (0.5, 0.81; 64%)	0.73 (0.56, 0.89; 76%)		1 (1, 1; 100%)	0.78 (0.48, 1; 93%)		0.87 (0.75, 0.99; 88%)	0.68 (0.49, 0.87; 73%)	
	Reader Three	0.55 (0.38, 0.71; 52%)	0.75 (0.56, 0.94; 82%)		0.87 (0.75, 0.99; 88%)	0.68 (0.49, 0.87; 73%)		0.81 (0.58, 1; 88%)	0.91 (0.76, 1; 93%)	
Medial Joint Space Narrowing	Gold Standard ²	0.55 (0.41, 0.7; 52%)	0.67 (0.51, 0.82; 56%)		0.73 (0.58, 0.89; 76%)	0.71 (0.53, 0.9; 73%)		0.64 (0.5, 0.79; 67%)	0.76 (0.62, 0.9; 73%)	
	Reader One	0.79 (0.67, 0.91; 76%)	0.80 (0.68, 0.93; 76%)		0.74 (0.62, 0.85; 70%)	0.67 (0.49, 0.86; 67%)		0.76 (0.62, 0.91; 76%)	0.73 (0.59, 0.87; 67%)	
	Reader Two	0.74 (0.62, 0.85; 70%)	0.67 (0.49, 0.86; 67%)		0.9 (0.77, 1; 94%)	1 (1, 1; 100%)		0.91 (0.82, 1; 91%)	0.84 (0.67, 1; 88%)	
	Reader Three	0.76 (0.62, 0.91; 76%)	0.73 (0.59, 0.87; 67%)		0.91 (0.82, 1; 91%)	0.84 (0.67, 1; 88%)		1 (1, 1; 100%)	1 (1, 1; 100%)	
Medial Tibial Osteophyte	Gold Standard ²	0.64 (0.48, 0.8; 61%)	0.49 (0.23, 0.75; 67%)		0.75 (0.6, 0.9; 73%)	0.45 (0.19, 0.72; 64%)		0.75 (0.6, 0.89; 73%)	0.48 (0.21, 0.75; 67%)	
	Reader One	0.63 (0.44, 0.81; 64%)	0.67 (0.47, 0.88; 76%)		0.7 (0.53, 0.87; 70%)	0.79 (0.61, 0.96; 85%)		0.64 (0.46, 0.82; 64%)	0.82 (0.66, 0.99; 88%)	
	Reader Two	0.7 (0.53, 0.87; 70%)	0.79 (0.61, 0.96; 85%)		0.78 (0.54, 1; 81%)	0.83 (0.61, 1; 87%)		0.94 (0.85, 1; 94%)	0.87 (0.72, 1; 91%)	
	Reader Three	0.64 (0.46, 0.82; 64%)	0.82 (0.66, 0.99; 88%)		0.94 (0.85, 1; 94%)	0.87 (0.72, 1; 91%)		1 (1, 1; 100%)	1 (1, 1; 100%)	
Patellofemoral Joint Space Narrowing	Gold Standard ²	0.25 (-0.12, 0.62; 46%)	0.7 (0.44, 0.95; 71%)		0.33 (0.08, 0.58; 54%)	0.79 (0.59, 0.99; 81%)		0.3 (0.08, 0.51; 46%)	0.79 (0.55, 1; 85%)	
	Reader One	0.69 (0.5, 0.89; 77%)	0.78 (0.59, 0.97; 77%)		0.55 (0.31, 0.78; 69%)	0.74 (0.51, 0.97; 77%)		0.32 (0.07, 0.57; 54%)	0.69 (0.38, 0.99; 76%)	

Variable	Comparison	Reader One ^I		Reader Two ^I		Reader Three ^I	
		Left Knee	Right Knee	Left Knee	Right Knee	Left Knee	Right Knee
Patellofemoral Osteophyte	Reader Two	0.55 (0.31, 0.78; 69%)	0.74 (0.51, 0.97; 77%)	0.66 (0.31, 1; 85%)	0.74 (0.35, 1; 78%)	0.81 (0.62, 1; 85%)	0.83 (0.64, 1; 86%)
	Reader Three	0.32 (0.07, 0.57; 54%)	0.69 (0.38, 0.99; 76%)	0.81 (0.62, 1; 85%)	0.83 (0.64, 1; 86%)	0.78 (0.63, 0.94; 85%)	0.79 (0.49, 1; 78%)
	Gold Standard ²	0.25 (-0.05, 0.55; 58%)	0.48 (0.26, 0.69; 62%)	0.3 (-0.07, 0.67; 58%)	0.53 (0.25, 0.81; 62%)	0.17 (-0.17, 0.5; 50%)	0.4 (0.08, 0.72; 50%)
	Reader One	0.43 (0.13, 0.74; 73%)	0.46 (0.19, 0.73; 64%)	0.44 (0.17, 0.71; 69%)	0.6 (0.39, 0.81; 73%)	0.28 (-0.07, 0.63; 62%)	0.37 (0.09, 0.66; 62%)
	Reader Two	0.44 (0.17, 0.71; 69%)	0.6 (0.39, 0.81; 73%)	0.7 (0.4, 1; 77%)	0.71 (0.37, 1; 78%)	0.67 (0.42, 0.91; 77%)	0.81 (0.6, 1; 86%)
	Reader Three	0.28 (-0.07, 0.63; 62%)	0.37 (0.09, 0.66; 62%)	0.67 (0.42, 0.91; 77%)	0.81 (0.6, 1; 86%)	0.51 (0.09, 0.92; 77%)	0.84 (0.52, 1; 89%)

^I Intra-reader reliability is found by comparing the same reader in the column and the row (for example, comparing the “Reader One” row with the “Reader One” column). Inter-reader reliability is found by comparing different readers in the column and row (for example, comparing the “Reader One” column with the “Gold Standard” row gives the inter-reader reliability between reader 1 and the radiologist).

² Gold standard is the reading of the experienced radiologist.

Table 3
Inter- and Intra-reader Reliability for Continuous Variables

Variable	Comparison	Reader One ^I		Reader Two ^I		Reader Three ^I	
		Left Knee	Right Knee	Left Knee	Right Knee	Left Knee	Right Knee
		ICC ² Mean Difference ³ (SD)					
OARS Score	Gold Standard ⁴	0.87 -0.88 (1.9)	0.85 -0.15 (1.84)	0.88 0.03 (1.78)	0.81 0.42 (2.11)	0.87 0.58 (1.84)	0.79 0.97 (2.05)
	Reader One	0.92 -1.58 (1.25)	0.9 -1.42 (1.25)	0.9 -0.91 (1.47)	0.87 -0.58 (1.52)	0.91 -1.45 (1.33)	0.87 -1.12 (1.43)
	Reader Two	0.9 -0.91 (1.47)	0.87 -0.58 (1.52)	0.93 0.81 (0.98)	0.89 1.33 (1.29)	0.96 -0.55 (0.83)	0.93 -0.55 (1.09)
	Reader Three	0.91 -1.45 (1.33)	0.87 -1.12 (1.43)	0.96 -0.55 (0.83)	0.93 -0.55 (1.09)	0.98 -0.13 (0.5)	0.97 0.07 (0.59)
Knee Alignment Angle (degrees)	Reader One	0.95 0.13 (1.19)	0.92 0.12 (1.65)	0.89 0.04 (1.68)	0.88 0.15 (1.88)	0.94 0.06 (1.25)	0.88 0.08 (1.98)
	Reader Two	0.89 0.04 (1.68)	0.88 0.15 (1.88)	0.97 0.36 (0.77)	0.96 0.2 (0.87)	0.91 0.02 (1.48)	0.93 -0.07 (1.44)
	Reader Three	0.94 0.06 (1.25)	0.88 0.08 (1.98)	0.91 0.02 (1.48)	0.93 -0.07 (1.44)	0.97 -0.02 (0.83)	0.98 -0.05 (0.64)
Lateral Joint Space Width (mm)	Reader One	0.91 0.2 (1.05)	0.87 -0.08 (1.57)	0.88 0.68 (1.09)	0.83 0.57 (1.69)	0.85 0.69 (1.24)	0.89 0.39 (1.39)
	Reader Two	0.88 0.68 (1.09)	0.83 0.57 (1.69)	0.87 -0.29 (0.88)	0.9 -0.79 (1.05)	0.92 0.01 (0.86)	0.92 -0.17 (1.17)
	Reader Three	0.85 0.69 (1.24)	0.89 0.39 (1.39)	0.92 0.01 (0.86)	0.92 -0.17 (1.17)	0.82 0.26 (1.11)	0.96 0.26 (0.61)
Medial Joint Space Width (mm)	Reader One	0.97 0.24 (0.52)	0.98 0.11 (0.41)	0.94 0.37 (0.75)	0.93 0.01 (0.93)	0.93 0.44 (0.82)	0.96 0.34 (0.69)
	Reader Two	0.94 0.37 (0.75)	0.93 0.01 (0.93)	0.97 -0.06 (0.37)	0.94 0.08 (0.72)	0.96 0.07 (0.58)	0.93 0.32 (0.87)
	Reader Three	0.93 0.44 (0.82)	0.96 0.34 (0.69)	0.96 0.07 (0.58)	0.93 0.32 (0.87)	1 0.01 (0.1)	0.99 0.05 (0.3)
Patellofemoral Joint Space Width (mm)	Reader One	0.95 0.13 (0.76)	0.91 0.6 (0.88)	0.92 0.11 (0.95)	0.91 -0.17 (0.83)	0.86 0.29 (1.27)	0.9 0.27 (1)
	Reader Two	0.92 0.11 (0.95)	0.91 -0.17 (0.83)	0.86 -0.45 (1.11)	0.89 0.4 (0.77)	0.91 0.17 (1)	0.91 0.48 (0.9)
	Reader Three	0.86 0.29 (1.27)	0.9 0.27 (1)	0.91 0.17 (1)	0.91 0.48 (0.9)	0.56 -0.58 (2.01)	0.94 -0.38 (0.68)

¹ Intra-reader reliability is found by comparing the same reader in the column and the row (for example, comparing the "Reader One" row with the "Reader One" column). Inter-reader reliability is found by comparing different readers in the column and row (for example, comparing the "Reader One" column with the "Gold Standard" row gives the inter-reader reliability between reader 1 and the radiologist).

² Pearson correlations similar to ICCs and available on request

³ Difference in means calculated by Column Reader – Row Reader. In the instances of intra-reader reliability, mean difference is calculated by First Read – Second Read.

⁴ Gold standard is the reading of the experienced radiologist.