Section: Orthopaedics

Original Research Article

TO STUDY THE INTER OBSERVER VARIATION IN THE ASSESSMENT OF SEVERITY IN OSTEOARTHRITIS OF KNEE ON PLAIN RADIOGRAPH USING KELLGREN-LAWRENCE GRADING

Sanjai Baskeran[1], T. Sathish Kumar[2], S. Natarajan[3]

[1,2]*Assistant Professor, Department of Orthopaedics, Panimalar Medical College Hospital and Research Institute, Chennai, India*
[3]*Professor, Department of Orthopaedics, Panimalar Medical College Hospital and Research Institute, Chennai, India.*

Received *: 30/06/2025*
Received in revised form *: 16/08/2025*
Accepted *: 02/09/2025*

Corresponding Author:
Dr. Sanjai Baskeran,
Assistant Professor, Department of Orthopaedic, Panimalar Medical College Hospital and Research Institute, Chennai, India.
Email: dr.sanjai.b@gmail.com

DOI: 10.70034/ijmedph.2025.3.519

Source of Support: Nil,
Conflict of Interest: None declared

Int J Med Pub Health
2025; 15 (3); 2830-2834

ABSTRACT

Background: Knee osteoarthritis (OA) is a common degenerative joint disease. Radiographic evaluation remains a cornerstone in diagnosis, staging, and monitoring disease progression, with the Kellgren-Lawrence (K-L) grading system widely used for severity assessment. However, variability among observers, especially between experienced and less experienced orthopaedic surgeons, may influence diagnostic consistency. The objective is to evaluate interobserver variation in the assessment of knee OA severity on plain radiographs using the K-L grading system between experienced and inexperienced orthopaedic surgeons.

Materials and Methods: A total of 128 anteroposterior (AP) standing knee radiographs of patients aged above 40 years were assessed. Twenty orthopaedic surgeons (8 with >5 years experience and 12 with <5 years experience) participated in the study. Radiographs were graded using the K-L system via an online platform. Data were analysed for interobserver agreement using intraclass correlation coefficients (ICC).

Results: The mean score of observers with less experience was 26.38 ± 8.741, while that of experienced observers was 16.26 ± 6.455. The interobserver agreement showed a high ICC value of 0.938 (95% CI: 0.912–0.956) for average measures and 0.883 (95% CI: 0.838–0.916) for single measures ($p<0.001$).

Conclusion: The study demonstrated excellent interobserver reliability in K-L grading of knee OA between experienced and inexperienced orthopaedic surgeons. While experienced surgeons tended towards slightly more consistent grading, statistical analysis revealed no significant variation between the groups.

Keywords: Osteoarthritis, Interobserver variability, Kellgren-Lawrence grading, Knee radiographs, Orthopaedic surgeons.

INTRODUCTION

One of the leading causes of long-term impairment in people worldwide is osteoarthritis (OA) of the knee, which is particularly common among the elderly Hunter DJ et al. (2019).[1] Gradual articular cartilage degradation, subchondral bone remodelling, and variable levels of synovial inflammation are characteristics of the syndrome Allen KD et al. (2015).[2] Plain radiography is still widely used in clinical and research settings, even though more recent imaging modalities, such magnetic resonance imaging (MRI), provide better visualisation of soft tissues, bone marrow oedema, and cartilage. Its widespread availability, affordability, and the relative simplicity of picture interpretation account for its continued usage Roemer FW et al.(2022); Braun HJ et al. (2011); KELLGREN JH et al.[3-5]

The Kellgren Lawrence (K L) grading scale was first used in 1957 and is now one of the most often used methods for radiological evaluation of OA Quinn L et al. (2023).[6] By considering the presence of osteophytes, joint space narrowing, subchondral sclerosis, and changes in bone morphology, the

approach ranks the severity of OA on a continuum from grade 0 (no symptoms of OA) to grade 4 (severe OA) Guermazi A et al. (2013).[7] The K L grading system has been around for a while, although it has its detractors. One persistent issue is its intrinsic subjectivity, which raises concerns regarding interobserver variability as different doctors may award different grades based on the same radiographic picture Steenkamp W et al. (2022).[8]

The degree of agreement or disagreement between several radiography interpreters while evaluating identical pictures is known as interobserver variability. Differences in clinical training, years of experience, knowledge with imaging criteria, and interpretative style can all lead to discrepancies in interpretation when it comes to knee OA Zhang W et al. (2008).[9] For several reasons, it is crucial to guarantee uniform and reproducible grading. First, proper diagnosis is essential to good patient care; incorrect categorisation might result in unsuitable treatment options. Second, grading is essential to research because it supports longitudinal monitoring, outcome evaluation, stratification, and patient inclusion criteria Javaid MK et al. (2012).[10]

Interobserver agreement, however, varies widely, with reported levels varying from modest to high agreement, depending on reader skill and methodological background, according to previous studies on K-L grading Knights AJ et al. (2023).[11] In routine practice, orthopaedic surgeons with varying degrees of expertise frequently examine knee radiographs. Whether these disparities in experience directly result in variations in grading accuracy and dependability is yet unknown. Eckersley T et al. (2021),[12] Decisions on patient treatment may suffer if non-expert readers consistently overestimate or underestimate K L grades. Determining the necessary experience, calibration sessions, or standardised training to lower variability can be guided by an understanding of this possible difference. Jiang T et al. (2024).[13]

Even though radiographic results are frequently used in osteoarthritis research, interobserver reliability data are frequently left out of studies. This disparity contrasts sharply with the assumption that radiographic grading would serve as the foundation for uniform, trustworthy assessment in open access research. Relatively few studies offer such information, especially when it comes to radiological characteristics like osteophytes or joint space narrowing, according to a systematic review looking at interobserver reliability in hip and knee OA Vina ER et al. (2018).[14] Larger, more meticulously planned experiments with a clear evaluation of interobserver reliability across grades and attributes were suggested by the authors.

Recognising these shortcomings, the current study uses the K L scale to assess the degree of interobserver agreement in knee OA diagnosis and grading. Orthopaedic surgeons from two different cohorts were compared: the "experienced group," which included those with a lot of clinical experience, and the "inexperienced group," which included those who were just training or had just begun their clinical careers. The objective is to shed light on any flaws in the present grading procedures by assessing agreement rates—using statistical techniques like Cohen's kappa coefficient—and pinpointing regions of consistent or inconsistent interpretation.

Objectives

1. Calculate the total interobserver agreement in K L grading for readers with and without expertise for a variety of knee radiographs.
2. Determine whether radiological characteristics lend themselves to more or lesser agreement by assessing consistency across individual aspects, such as the presence or absence of osteophytes, the degree of joint space constriction, subchondral sclerosis, and deformity.
3. Examine if interpretative consistency and experience level are correlated, offering empirical support for the idea that more training or calibration may be required to standardise grading among readers.
4. Based on the reliability data that has been seen, offer useful suggestions for standardising radiography scoring procedures and training frameworks in clinical and research settings.

Ensuring dependability is crucial because radiographic grading plays a crucial role in the therapy of knee OA, including clinical trial design, therapeutic decision-making, surgical planning, and diagnostic confirmation. It is anticipated that the results of this study will affect academic training programs, multi-center research trials, and routine orthopaedic practice. Techniques like standardised picture reference sets, calibration workshops, and digital training modules might be used to increase consistency if experience-dependent grading variability is verified. On the other hand, relying on radiographic grading across experience groups might be justified if there are little disparities.

MATERIALS AND METHODS

This observational study was conducted to evaluate interobserver variability in the radiographic grading of knee osteoarthritis (OA) using the Kellgren and Lawrence (K-L) system among orthopedic surgeons with varying levels of experience.

Study Design and Population: The study included anteroposterior (AP) radiographs of the knee joint obtained in the standing position from patients aged over 40 years. Only radiographs with adequate positioning and image quality were included; poorly taken or suboptimal radiographs were excluded to ensure consistent interpretation. A total of 128 radiographs meeting the inclusion criteria were selected for analysis.

Study Duration: 3 months

Study Location: Panimalar Medical College Hospital & Research Institute, Varadharajapuram, Poonamallee, Chennai – 600123.

Sample Size: 128 AP knee radiographs.

Inclusion Criteria

Radiographs with proper positioning and image quality.

Exclusion Criteria

Radiographs with poor quality or improper positioning.

Participants: 20 orthopaedic surgeons.

- Group 1: 8 surgeons with >5 years of experience.
- Group 2: 12 surgeons with <5 years of experience.

Procedure:

- Radiographs were uploaded into a Google Form with 128 questions, each containing one radiograph.
- For each radiograph, participants selected one of five choices corresponding to K-L grades (0 to 4).
- Participants could revise their responses before final submission.

Data Collection: Responses were collected online and analyzed anonymously.

Statistical Analysis: Data analyzed using SPSS. Intraclass correlation coefficient (ICC) calculated to assess interobserver agreement. Statistical significance set at $p<0.05$.

RESULTS

[Table 1] shows the observer ratings' mean scores and standard deviation. Compared to surgeons with more than five years of experience (16.26 ± 6.455), those with less than five years of experience had a higher mean score (26.38 ± 8.741). A total of 128 radiographs were included in the sample for both groups.

Table 1: Mean and Standard Deviation of Observer Scores

Observer Group	Mean Score	Standard Deviation	N
<5 years experience	26.38	8.741	128
>5 years experience	16.26	6.455	128

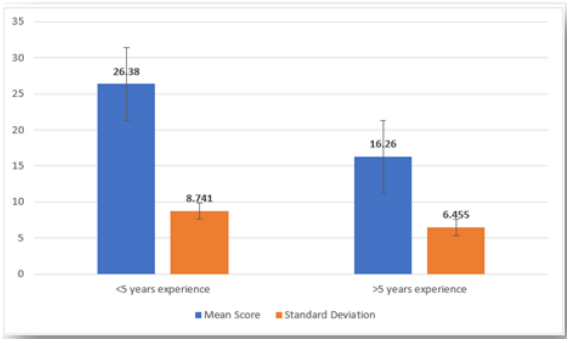

Figure 1: Mean and Standard Deviation of Observer Scores

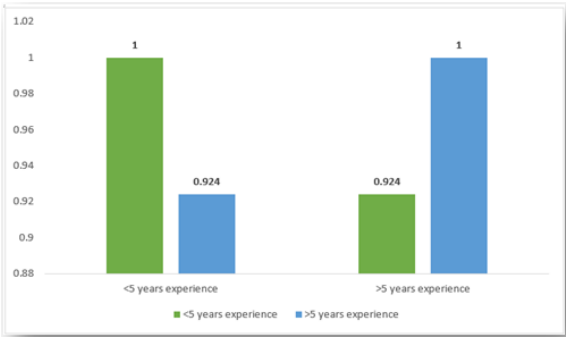

Figure 2: Inter-Item Correlation Matrix

Table 2: Inter-Item Correlation Matrix

	<5 years experience	>5 years experience
<5 years experience	1.000	0.924
>5 years experience	0.924	1.000

Displays the observer groups' inter-item correlation matrix. Strong agreement was indicated by the high correlation value (0.924) between less experienced and more experienced surgeons.

Table 3: Intraclass Correlation Coefficient (Single Measures)

ICC Value	95% CI (Lower)	95% CI (Upper)	F Value	df1	df2	p-value
0.883	0.838	0.916	16.088	127	127	<0.001

The intraclass correlation coefficient (ICC) for individual metrics is shown in Table 3. Excellent interobserver reliability was shown by the ICC value of 0.883, which had a 95% confidence range of 0.838 to 0.916. The F value was statistically significant ($p<0.001$) and was 16.088 (df1=127, df2=127).

Table 4: Intraclass Correlation Coefficient (Average Measures)

ICC Value	95% CI (Lower)	95% CI (Upper)	F Value	df1	df2	p-value
0.938	0.912	0.956	16.088	127	127	<0.001

The ICC for average metrics is shown in Table 4. The ICC score showed significantly greater agreement among observers, rising to 0.938 (95% CI: 0.912–0.956). The same statistical significance and F value were observed.

DISCUSSION

According to the Kellgren-Lawrence (K-L) grading system, orthopaedic surgeons in this study showed a high degree of agreement when determining the

severity of osteoarthritis (OA) in the knee. Hunter DJ et al. (2019),[1] Compared to surgeons with more than five years of experience (16.26 ± 6.455), those with less than five years of experience showed a higher mean grading score (26.38 ± 8.741). Nevertheless, overall dependability was unaffected by this change, which was not statistically significant [Table 1]. Consistent grading patterns across experience levels were shown by the inter-item correlation matrix, which revealed a substantial positive correlation (0.924) between the two groups [Table 2]. Steenkamp W et al. (2022).[8]

These results were further confirmed by the intraclass correlation coefficient (ICC) study. Both measures' ICCs were statistically significant ($p<0.001$), with 0.883 (95% CI: 0.838–0.916) for single measurements and 0.938 (95% CI: 0.912–0.956) for average measures [Tables 3 and 4]. These findings are consistent with other research that found significant interobserver reliability in K-L grading across a range of clinical experience levels Quinn L et al. (2023),[6] Jiang T et al. (2024).[13]

A discrepancy between radiological results and clinical symptoms in knee OA has been shown in several investigations Hunter DJ et al. (2019).[1] Radiographs frequently don't match up exactly with what patients say about their pain and functional limitations Guermazi A et al. (2013).[7] Researchers have suggested improving grading schemes, utilising quantitative joint space measuring methods, and utilising tilted radiography views for increased sensitivity in order to improve the diagnostic value of radiographs Javaid MK et al. (2012),[10] Hayashi D et al. (2014).[15]

Our results demonstrate the K-L grading system's strength as a trustworthy instrument, especially for novice surgeons. Variability, however, may be impacted by elements including observer bias and subjective interpretation of minute radiographic characteristics Teoh YX et al. (2022).[16] Consensus meetings, training programs, and reference atlases may all contribute to increased consistency among observers Tillett W et al. (2014) and Maricar N et al. (2016),[17,18] Knights AJ et al. (2023).[11] The expense and restricted availability of MRI in many contexts highlight the ongoing value of plain radiography for OA assessment, despite the fact that it offers superior soft tissue visualisation and can identify early cartilage changes Hayashi D et al. (2019).[19]

In conclusion, this study shows that the K-L method has great interobserver agreement when it comes to knee OA grading, indicating that it may be used to different degrees of clinical skill. For multicenter investigations and epidemiological research, periodic calibration and validation of observer reliability are still crucial Plotz B et al (2021).[20]

CONCLUSION

The results of this study show that experienced and novice orthopaedic surgeons do not significantly differ in their evaluation of the degree of knee osteoarthritis on plain radiographs using the Kellgren-Lawrence grading system. The statistical analysis revealed no significant difference between the two groups, despite the fact that experienced surgeons exhibited somewhat greater accuracy in grading. This implies that regardless of the observer's degree of expertise, the K-L grading system is a valid and consistent method for assessing knee OA.

Additionally, the strong intraclass correlation coefficients support the broad usage of this grading system in clinical and research contexts by confirming great agreement between observers. Grading consistency should be further improved, especially in multicenter research, by implementing frequent calibration exercises, standardised techniques, and systematic training. Even though sophisticated imaging techniques like MRI provide better soft tissue evaluation, plain radiography is still a useful and useful way to assess OA, particularly in environments with limited resources. In summary, this study highlights the Kellgren-Lawrence grading system's durability and dependability as well as its suitability for routine clinical usage by orthopaedic surgeons with different degrees of competence.

Limitations of the Study

1. The study was conducted in a single center, which may limit the generalizability of the findings to other settings.
2. The sample size of 128 radiographs, though adequate for analysis, may not capture the full variability present in larger populations.
3. To avoid using additional radiographic views that might affect grading, only anteroposterior (AP) standing knee radiographs were utilised.
4. MRI results and clinical complaints were not correlated with radiographic grading in this investigation, which may have yielded a more thorough evaluation.
5. Responses may have been impacted by observer tiredness and differing attention levels during the grading process.
6. Participants' grading behaviour may have been influenced by potential bias because they were aware that they were a part of a research project.

REFERENCES

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet. 2019 Apr 27;393(10182):1745-1759. doi: 10.1016/S0140-6736(19)30417-9. PMID: 31034380.
2. Allen KD, Golightly YM. State of the evidence. Curr Opin Rheumatol. 2015 May;27(3):276-83. doi: 10.1097/BOR.0000000000000161.
3. Roemer FW, Guermazi A, Demehri S, Wirth W, Kijowski R. Imaging in Osteoarthritis. Osteoarthritis Cartilage. 2022 Jul;30(7):913-934. doi: 10.1016/j.joca.2021.04.018. Epub 2021 Sep 22.
4. Braun HJ, Gold GE. Diagnosis of osteoarthritis: imaging. Bone. 2012 Aug;51(2):278-88. doi: 10.1016/j.bone.2011.11.019. Epub 2011 Dec 3. PMID: 22155587;
5. KELLGREN JH, LAWRENCE JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957 Dec;16(4):494-502.

doi: 10.1136/ard.16.4.494. PMID: 13498604; PMCID: PMC1006995.
6. Quinn L, Tryposkiadis K, Deeks J, De Vet HCW, Mallett S, Mokkink LB, Takwoingi Y, Taylor-Phillips S, Sitch A. Interobserver variability studies in diagnostic imaging: a methodological systematic review. Br J Radiol. 2023 Aug;96(1148):20220972. doi: 10.1259/bjr.20220972. Epub 2023 Jun 29. PMID: 37399082; PMCID: PMC10392644.
7. Guermazi A, Hayashi D, Eckstein F, Hunter DJ, Duryea J, Roemer FW. Imaging of osteoarthritis. Rheum Dis Clin North Am. 2013 Feb;39(1):67-105. doi: 10.1016/j.rdc.2012.10.003.
8. Steenkamp W, Rachuene PA, Dey R, Mzayiya NL, Ramasuvha BE. The correlation between clinical and radiological severity of osteoarthritis of the knee. SICOT J. 2022;8:14. doi: 10.1051/sicotj/2022014. Epub 2022 Apr 6.
9. Zhang W, Moskowitz RW, Nuki G, Abramson S, Altman RD, Arden N, Bierma-Zeinstra S, Brandt KD, Croft P, Doherty M, Dougados M, Hochberg M, Hunter DJ, Kwoh K, Lohmander LS, Tugwell P. OARSI recommendations for the management of hip and knee osteoarthritis, Part II: OARSI evidence-based, expert consensus guidelines. Osteoarthritis Cartilage. 2008 Feb;16(2):137-62. doi: 10.1016/j.joca.2007.12.013.
10. Javaid MK, Kiran A, Guermazi A, Kwoh CK, Zaim S, Carbone L, Harris T, McCulloch CE, Arden NK, Lane NE, Felson D, Nevitt M; Health ABC Study. Individual magnetic resonance imaging and radiographic features of knee osteoarthritis in subjects with unilateral knee pain: the health, aging, and body composition study. Arthritis Rheum. 2012 Oct;64(10):3246-55. doi: 10.1002/art.34594.
11. Knights AJ, Redding SJ, Maerz T. Inflammation in osteoarthritis: the latest progress and ongoing challenges. Curr Opin Rheumatol. 2023 Mar 1;35(2):128-134. doi: 10.1097/BOR.0000000000000923.
12. Eckersley T, Faulkner J, Al-Dadah O. Inter- and intra-observer reliability of radiological grading systems for knee osteoarthritis. Skeletal Radiol. 2021 Oct;50(10):2069-2078. doi: 10.1007/s00256-021-03767-y. Epub 2021 Apr 15.
13. Jiang T, Lau SH, Zhang J, Chan LC, Wang W, Chan PK, Cai J, Wen C. Radiomics signature of osteoarthritis: Current status and perspective. J Orthop Translat. 2024 Mar 16;45:100-106. doi: 10.1016/j.jot.2023.10.003.
14. Vina ER, Kwoh CK. Epidemiology of osteoarthritis: literature update. Curr Opin Rheumatol. 2018 Mar;30(2):160-167. doi: 10.1097/BOR.0000000000000479.
15. Hayashi D, Roemer FW, Guermazi A. Imaging of osteoarthritis-recent research developments and future perspective. Br J Radiol. 2018 May;91(1085):20170349. doi: 10.1259/bjr.20170349. Epub 2018 Jan 19.
16. Teoh YX, Lai KW, Usman J, Goh SL, Mohafez H, Hasikin K, Qian P, Jiang Y, Zhang Y, Dhanalakshmi S. Discovering Knee Osteoarthritis Imaging Features for Diagnosis and Prognosis: Review of Manual Imaging Grading and Machine Learning Approaches. J Healthc Eng. 2022 Feb 18;2022:4138666. doi: 10.1155/2022/4138666. Retraction in: J Healthc Eng. 2023 Oct 11;2023:9765742. doi: 10.1155/2023/9765742. PMID: 35222885; PMCID: PMC8881170.
17. Tillett W, Jadon D, Shaddick G, Robinson G, Sengupta R, Korendowych E, de Vries CS, McHugh NJ. Feasibility, reliability, and sensitivity to change of four radiographic scoring methods in patients with psoriatic arthritis. Arthritis Care Res (Hoboken). 2014 Feb;66(2):311-7. doi: 10.1002/acr.22104. PMID: 23925955.
18. Maricar N, Callaghan MJ, Parkes MJ, Felson DT, O'Neill TW. Interobserver and Intraobserver Reliability of Clinical Assessments in Knee Osteoarthritis. J Rheumatol. 2016 Dec;43(12):2171-2178. doi: 10.3899/jrheum.150835. Epub 2016 Oct 1.
19. Hayashi D, Roemer FW, Guermazi A. Magnetic resonance imaging assessment of knee osteoarthritis: current and developing new concepts and techniques. Clin Exp Rheumatol. 2019 Sep-Oct;37 Suppl 120(5):88-95. Epub 2019 Oct 15.
20. Plotz B, Bomfim F, Sohail MA, Samuels J. Current Epidemiology and Risk Factors for the Development of Hand Osteoarthritis. Curr Rheumatol Rep. 2021 Jul 3;23(8):61. doi: 10.1007/s11926-021-01025-7.