Arthroplasty Today 15 (2022) 98–101

Contents lists available at ScienceDirect

Arthroplasty Today

journal homepage: http://www.arthroplastytoday.org/

Original research

Do Physicians Overestimate Radiographic Findings in Patients Undergoing Knee Arthroplasty?

James J. Gregory, MD [a, *], Parisa Ziarati, MPH [b], Paul M. Werth, PhD [a, c], David S. Jevsevar, MD, MBA [a, c]

[a] Department of Orthopaedics, Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA
[b] Geisel School of Medicine at Dartmouth, Hanover, NH, USA
[c] Department of Orthopaedics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

ARTICLE INFO

Article history:
Received 8 December 2021
Received in revised form
24 March 2022
Accepted 25 March 2022

Keywords:
Total joint arthroplasty
Total knee arthroplasty
Osteoarthritis
Radiographs
Kellgren-Lawrence
Inter-rater reliability

ABSTRACT

Background: Total knee arthroplasty (TKA) is 1 of the top 2 most common and expensive surgical procedures among Medicare beneficiaries. Due to the procedure's high annual cost, overdiagnosis and subsequent overutilization of TKA has substantial health-policy implications. Concerns regarding the overexaggeration of radiographic findings and overutilization of TKA have been expressed by medical insurers. Currently, the standard of care for assessing potential knee arthroplasty candidates includes assigning a Kellgren-Lawrence (KL) radiographic score. Our study investigated the accuracy of reported preoperative KL scores in patients undergoing TKA.

Material and methods: Records of 277 patients who had underwent TKA at our institution for knee osteoarthritis were randomly selected from a large patient data registry and retrospectively reviewed. Two blinded raters assigned KL scores to the radiographs obtained during the preoperative assessment, which were compared to the scores reported by the operative surgeon. An intraclass correlation coefficient (ICC) was calculated to determine inter-rater reliability.

Results: Between blinded raters, $ICC_{3k} = 0.88$ (95% confidence interval: 0.86-0.90, $P < .001$), demonstrating good reliability. Between all raters, $ICC_{2k} = 0.89$ (95% confidence interval: 0.86-0.90, $P < .001$), also demonstrating good agreement. Raters fully agreed on the KL classification for 196 patients (70.76%). Compared with blinded raters, the operative surgeon assigned lower KL scores.

Conclusion: Reporting of KL score is consistent between operative surgeons and independent reviewers. In cases of disagreement between reviewers, the operative surgeon was generally more conservative in their estimation of the extent of osteoarthritis present radiographically. Concerns regarding inflation of radiographic findings to support surgical preauthorization are unwarranted.

Introduction

Total knee arthroplasty (TKA) has been shown to be a beneficial treatment for moderate to severe knee osteoarthritis (OA) when compared with conservative treatment strategies and medical management of arthritis-related pain [1-3]. OA affects over 250 million individuals worldwide, of which 83% report knee OA [4]. To date, TKA is 1 of the 2 most common and expensive surgical procedures among Medicare beneficiaries with an estimated aggregate cost of $9.2 billion in 2007 [5]. OA accounted for $353 billion in health-care expenditure in 2005 alone, and TKA case volume is projected to increase to 3.48 million cases annually by 2030. The overdiagnosis and subsequent overutilization of TKA has significant health-policy ramifications [5,6]. A 2014 study by Riddle et al. estimated that as many as 34% of TKA procedures were inappropriate, most commonly due to inappropriate radiographic evidence of disease [7]. Following these findings, insurers have speculated that surgeons performing arthroplasty may overestimate the severity of OA during preoperative assessment to facilitate insurance preauthorization.

Currently, the most widely used classification scheme for radiographic evidence of knee OA is the Kellgren-Lawrence (KL) classification. Radiographs evaluated using the KL classification are

* Corresponding author. Dartmouth-Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH 03766, USA. Tel.: +1 603 653 3595.
E-mail address: james.j.gregory@hitchcock.org

https://doi.org/10.1016/j.artd.2022.03.022

Table 1
Kellgren-Lawrence classifications.

Kellgren-Lawrence classification	Description
Grade 0	No joint space narrowing or reactive changes
Grade 1	Doubtful joint space narrowing, possible osteophyte presence
Grade 2	Osteophytes present, possible joint space narrowing
Grade 3	Moderate osteophytes, definite joint space narrowing, sclerosis may be present, possible bone-end deformity
Grade 4	Large osteophytes present, marked joint space narrowing, severe sclerosis, bone-end deformity present

assigned a grade ranging from 0 to 4, with grade 0 demonstrating no signs of OA, and grade 4 being severe OA. Descriptions of all scoring are provided in Table 1 [8,9].

The decision to proceed with TKA is not based entirely upon radiographic findings but relies upon shared decision-making between patient and surgeon. Shortly following the study by Riddle et al., the American Academy of Orthopaedic Surgeons outlined appropriate use criteria for the surgical management of OA of the knee, which considers the clinical picture of the patient [7]. These criteria look at function-limiting pain, range of motion in both flexion and extension, functional instability, pattern of arthritic involvement, radiographic severity of disease, limb alignment, mechanical symptoms, and patient age [10]. For TKA to be deemed appropriate, mild to moderate or severe joint space narrowing must be present, thus demonstrating a KL classification of 2 to 4. The appropriate use criteria were developed using the modified RAND Delphi method, a system of determining appropriateness for a myriad of surgical procedures over the last 30 years [11]. It is important to note that the RAND method does not consider cost and instead focuses on the best prognostic evidence of risks and benefits of a procedure, along with the consensus of an expert clinical group [12]. Due to the substantial cost associated with TKA, it is imperative that surgeons are diligent in identifying when TKA is warranted. We sought to assess if orthopedic surgeons performing TKA are accurately assigning KL scores in patients with knee OA during preoperative assessments.

Material and methods

This retrospective study was submitted to and approved by our institutional review board and conducted in alignment with their policies. Two hundred seventy-seven patients scheduled for a TKA, between 2011 and 2020, with a diagnosis of knee OA, were randomly selected from a large patient data repository at a tertiary medical center in New England, representing 6.1% (n = 4509) of all cases performed during the study window. Patient demographics were recorded and included gender, race, age, and Charlson Comorbidity Index.

Two blinded independent raters assessed the radiographs of the patient cohort to determine KL classification. Classification levels include grade 2 ("Some osteophytes, some narrowing"), grade 3 ("Moderate osteophytes, significant narrowing"), or grade 4 ("Large osteophytes, marked narrowing") as described in Table 1. The KL classifications were first compared between blinded raters for agreement. Subsequent agreement levels were sought across both blinded raters and the treating surgeons (n = 14). Finally, individual rater bias was assessed to determine if the treating surgeon group tended toward greater or lesser severity classifications. The treating surgeon raters are all members of the arthroplasty clinic onsite. One blinded rater is a fellowship-trained arthroplasty surgeon, while the other is a medical student.

Analytic approach

To determine inter-rater reliability, intraclass correlation coefficients were populated, first for the 2 blinded raters (average of all ratings for fixed raters: ICC_{3k} [13]), followed by all 3 rater groups (average of all ratings for random raters: ICC_{2k} [13]). Both ICC values were calculated in the R environment v.4.0.3 [14] utilizing the "psych" package [15]. Subsequent individual rater bias was determined utilizing a coefficient of systematic rater bias [16] using the "irr" package [17]. The direction and distance from 0.50 paired with a significant χ^2 statistic can be interpreted as the direction and severity of bias between 2 raters. Classification of agreement was based upon the guidelines set forth by Koo and Li where ICC <0.50 = poor, 0.5-0.75 = moderate, 0.75-0.9 = good, and >0.90 = excellent [18].

Table 2
Patient cohort characteristics.

Total	277
Gender[a], male (%)	132 (47.7)
Race, Hawaiian Native/Pacific Islander (%)	1 (0.4)
Race, White (%)	276 (99.6)
Age, mean (SD)	68.95 (10.27)
CCI, mean (SD)	2.52 (2.81)
CCI category (%)	
0	95 (34.4)
1	25 (9.1)
2+	156 (56.5)

CCI, Charlson Comorbidity Index; SD, standard deviation.
[a] Referent: female.

Table 3
Number of cases by year.

Year	Number of cases
2011	11 (4.0%)
2012	26 (9.4%)
2013	41 (14.8%)
2014	30 (10.8%)
2015	26 (9.4%)
2016	31 (11.2%)
2017	38 (13.7%)
2018	48 (17.3%)
2019	19 (6.9%)
2020	7 (2.5%)

Results

Patient characteristics are presented in Table 2. Of the 277 cases evaluated, 132 (47.7%) were male, and 276 (99.6%) were Caucasian. The mean patient age was 68.95 ± 10.27 years. The mean Charlson Comorbidity Index score was 2.52 ± 2.81. The patient cohort included KL classification frequencies, as defined by the treating surgeon, including grade 2 (n = 46), grade 3 (n = 105), and grade 4 (n = 126). Cases by year are also presented in Table 3. The ICC_{3k} for the 2 blinded raters was 0.88 (95% CI: 0.86-0.90, $P < .001$) demonstrating good reliability. The ICC_{2k} for the 3 rater groups was 0.89 (95% CI: 0.86-0.90, $P < .001$), also demonstrating good reliability. Raters fully agreed on the KL classification for 196 patients (70.76%).

Sample radiographs where all raters were in agreement can be seen in Figure 1.

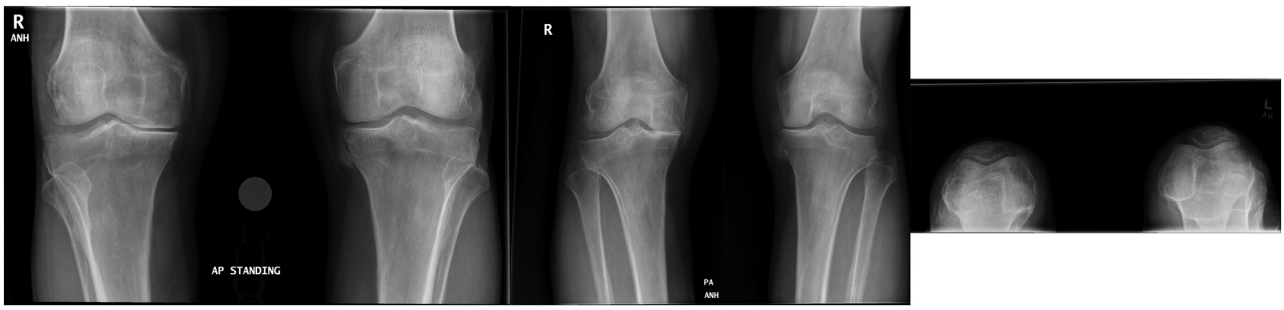

Figure 1. Standing anteroposterior, posteroanterior Rosenberg, and sunrise view radiographs of a patient who underwent right total knee arthroplasty. The blinded raters and the operative surgeon were in agreement on the level of osteoarthritis present. All raters assigned a Kellgren-Lawrence grade 4.

Systematic bias was negligible between the 2 blinded raters ($P = .60$, $\chi^2 = 2.08$, $P = .15$). A significant systematic bias was found between the surgeon blinded rater and the treating surgeon group ($P = .34$, $\chi^2 = 8.91$, $P = .003$) demonstrating that when disagreement was discovered, the treating surgeon tended to classify the level of OA as less severe than the blinded rater. A similar direction and level of bias were discovered between the medical student blinded rater and the treating surgeon group ($P = .32$, $\chi^2 = 6.48$, $P = .01$). Table 4 demonstrates the frequency and direction of disagreements of blinded rates compared with all treating surgeons. Sample radiographs of a patient where the blinded raters agreed but the operative surgeon was in disagreement can be seen in Figure 2.

Discussion

Our study sought to investigate whether orthopedic surgeons accurately report KL classifications of patients scheduled to undergo TKA during the preoperative assessment. Between blinded evaluators, there was good agreement on KL classification of radiographs as demonstrated by an ICC_{3k} value of 0.88. That is to say, the KL classification assigned to a given radiograph was consistent between both reviewers. When comparing operative surgeons to blinded evaluators, good agreement was demonstrated once more with an ICC_{2k} of 0.89. The ICC values observed in this study align with, and even exceeded, ICC values reported in current literature evaluating interrater reliability of KL classification of knee radiographs [19]. These findings demonstrate that the radiographic severity of OA present at patients' preoperative assessments was consistently evaluated to be the same by both the operative surgeon and retrospective evaluators. In instances where the KL score differed between the preoperative assessment and blinded retrospective review, the operative surgeon tended to report less severe OA, suggesting the decision to proceed with TKA was not skewed by overestimation of disease severity present on imaging. These findings show that payer claims of inflation of radiographic findings to support surgical preauthorization for TKA at our institution may be unwarranted.

While all images in this study were scored as KL classification grade 2 or greater, the decision to proceed with joint arthroplasty is not based on imaging alone. It has been well documented that patients with radiographs demonstrating a KL grade 2 and below have increased risk of poor outcomes following TKA [20-23]. One could speculate that orthopedic surgeons with a focus on providing value-based care and working toward the best possible outcomes for their patients would have more conservative scoring of disease state as opposed to overestimating the KL classification.

Table 4
Agreement and disagreement frequencies and direction as compared to treating surgeons.

Evaluator	Less severe	Agreement	More severe
Surgeon blinded rater	28 (10.1%)	196 (70.8%)	53 (19.1%)
Medical student blinded rater	15 (5.4%)	232 (83.8%)	30 (10.8%)

This study is limited by several factors. Notably, the ability to generalize these findings is difficult given that the data used in the analysis were from a single academic institution. This is partially mitigated by the fact that the patients reviewed were initially evaluated by several treating clinicians over the course of an 8-year period. Furthermore, the fellowship-trained rater and medical student demonstrated agreement, suggesting the broad utility of this classification system. While the patient demographics included in this study are representative of the population served by our institution, the homogenous patient population does not allow for evaluation of the impact of other demographic factors beyond sex to be considered. Future work should expand the study to involve multiple centers across a variety of geographic locations and environments and include a more diverse patient population to improve the generalizability of findings.

Conclusion

While health-care cost reduction remains a leading topic across all medical fields, the notion that orthopedic surgeons overestimate the severity of OA in preoperative assessments of patients may be overstated. Operative surgeons were consistently in agreement with 2 blinded evaluators and, in the cases of differing KL classification grades, were routinely found to underestimate the degree of OA present. Radiographic evaluation is 1 of multiple factors surgeons should consider while providing care for knee OA.

Acknowledgments

The authors would like to thank Elaina Vitale, a medical librarian, for her assistance with conducting a review of current literature in the preparation of this manuscript.

Conflicts of interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: D. S. Jevsevar has stock or stock options in Risalto Healthcare and serves on the AAOS Device, Biologics, and Technology Committee; AAOS Registry Oversight Committee; and the AAHKS EBPC.

Figure 2. Standing anteroposterior, posteroanterior Rosenberg, and sunrise view radiographs of a patient who underwent right total knee arthroplasty. The blinded raters and the operative surgeon disagreed on the Kellgren-Lawrence (KL) classification of osteoarthritis present. Both blinded raters assigned a KL grade 4, while the operative surgeon assigned KL grade 3.

For full disclosure statements refer to https://doi.org/10.1016/j.artd.2022.03.022.

References

[1] Lavernia CJ, Guzman JF, Gachupin-Garcia A. Cost effectiveness and quality of life in knee arthroplasty. Clin Orthop Relat Res 1997;(345):134.
[2] Miyasaka KC, Ranawat CS, Mullaji A. 10- to 20-year followup of total knee arthroplasty for valgus deformities. Clin Orthop Relat Res 1997;(345):29.
[3] Hawker G, Wright J, Coyte P, et al. Health-related quality of life after knee replacement. J Bone Joint Surg Am 1998;80(2):163.
[4] Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380(9859):2163.
[5] Courtney PM, Frisch NB, Bohl DD, Della Valle CJ. Improving value in total hip and knee arthroplasty: the role of high volume hospitals. J Arthroplasty 2018;33(1):1.
[6] Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am 2007;89(4):780.
[7] Riddle DL, Jiranek WA, Hayes CW. Use of a validated algorithm to judge the appropriateness of total knee arthroplasty in the United States: a multicenter longitudinal cohort study. Arthritis Rheumatol 2014;66(8):2134.
[8] Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16(4):494.
[9] Schiphof D, Boers M, Bierma-Zeinstra SM. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann Rheum Dis 2008;67(7):1034.
[10] Appropriate use criteria for the management of osteoarthritis of the knee. Rosemont, IL: AAOS; 2016.
[11] Lawson EH, Gibbons MM, Ingraham AM, Shekelle PG, Ko CY. Appropriateness criteria to assess variations in surgical procedure use in the United States. Arch Surg 2011;146(12):1433.
[12] Fitch K, Bernstein SJ, Aguilar MD, et al. The RAND/UCLA appropriateness method user's manual. Santa Monica, CA: RAND Corporation; 2001. https://www.rand.org/pubs/monograph_reports/MR1269.html.
[13] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420.
[14] Team RC. In: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
[15] Revelle W. In: Psych: procedures for psychological, psychometric, and personality research. Evanston, IL: Northwestern University; 2020.
[16] Bishop YM, Fienberg SE, Holland PW. Discrete multivariate analysis theory and practice. Cambridge: MIT Press; 2007. https://doi.org/10.1007/978-0-387-72806-3 [accessed 19.08.20].
[17] Gamer M, Fellows I, Singh P. Irr; various coefficients of interrater reliability and agreement. In: Coefficients of interrater reliability and agreement for quantitative. R-project; 2012. https://CRAN.R-project.org/package=irr [accessed 19.08.20].
[18] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155.
[19] Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res 2016;474(8):1886.
[20] Dowsey MM, Nikpour M, Dieppe P, Choong PF. Associations between pre-operative radiographic changes and outcomes after total knee joint replacement for osteoarthritis. Osteoarthritis Cartilage 2012;20(10):1095.
[21] Keurentjes JC, Fiocco M, So-Osman C, et al. Patients with severe radiographic osteoarthritis have a better prognosis in physical functioning after hip and knee replacement: a cohort-study. PLoS One 2013;8(4):e59500.
[22] Riis A, Rathleff MS, Jensen MB, Simonsen O. Low grading of the severity of knee osteoarthritis pre-operatively is associated with a lower functional level after total knee replacement: a prospective cohort study with 12 months' follow-up. Bone Joint J 2014;96-B(11):1498.
[23] Dowsey MM, Spelman T, Choong PF. Development of a prognostic nomogram for predicting the probability of nonresponse to total knee arthroplasty 1 year after surgery. J Arthroplasty 2016;31(8):1654.