ORIGINAL ARTICLE

Interobserver Reliability of the Kellgren-Lawrence Classification of Degenerative Knee Osteoarthritis among Resident Physicians from the University of the Philippines – Philippine General Hospital

Alfredo P. Pacheco, MD and Gregorio Marcelo S. Azores, MD

Arthroplasty and Lower Extremity Reconstruction Service, Department of Orthopedics,
College of Medicine and Philippine General Hospital, University of the Philippines Manila

ABSTRACT

Objectives. To determine the interobserver reliability of the Kellgren-Lawrence classification among selected residents from departments forming the University of the Philippines – Philippine General Hospital (UP-PGH) Osteoarthritis Multidisciplinary Clinic (OAMDC).

Methods. From each department, 3 resident physicians (n = 9) were randomly chosen and tasked to categorize 20 knee anteroposterior and lateral radiographs into KLC grades. Inter-observer reliability was assessed using Fleiss's kappa coefficient (κ).

Results. Results show that there was 31.90% (κ = 0.3190 ± 0.0228, p-value < 0.05) agreement beyond chance in KLC grading of the radiograph series among all participating residents. Sub-group analyses of interobserver reliability in terms of departmental affiliation noted a range of agreement beyond chance in KLC grading, from 10.52% (κ = 0.1052 ± 0.0779, p-value < 0.05) to 56.38% (κ = 0.5638 ± 0.0844, p-value < 0.05).

Conclusion. The findings reveal significant variability of agreement beyond chance in KLC grading, both within and among residents of UP-PGH OAMDC-participating clinical departments, which may reflect differences in training or competency and/or the documented limitations of the KLC system. Further investigation to improve diagnostic and severity assessment accuracy and uniformity in the institution is therefore warranted.

Key Words: Osteoarthritis, Arthritis, Knee, Kellgren-Lawrence

INTRODUCTION

Osteoarthritis (OA), a heterogeneous group of overlapping yet distinct joint disorders with similar biological, morphological, and clinical outcomes,[1] is the most common rheumatic affliction and the fourth leading cause of disability,[2] affecting an estimated 250 million people worldwide.[3] The 2003 Philippine National Nutrition and Health Survey (NNHeS) pegged the national prevalence of OA to 0.5% among those aged 20 years and 11% among those at and beyond 60 years of age, with females being affected approximately four-fold more than men.[4] With the global trends of population aging, sedentary lifestyle, and chronic lifestyle-associated diseases remarkably permeating to the Philippine setting, these figures arguably are now gross underestimation of a public health crisis.

Affecting one or several diarthrodial joints, often not discriminating between large or weight-bearing (e.g., knee

Poster presented in the Philippine Orthopedic Association Annual Convention on November 14, 2018, at the EDSA Shangri-La Hotel, Mandaluyong City.

Corresponding author: Gregorio Marcelo S. Azores, MD
Department of Orthopedics
College of Medicine and Philippine General Hospital
University of the Philippines Manila
Taft Avenue, Ermita, Manila 1000, Philippines
Email: gsazores@up.edu.ph

and hip joints) and small or non-weight-bearing joints (e.g., hand joints),[1] it characteristically manifests a polymorphic nature that remains difficult to rigorously define.[5] Likely in part due to the still-evolving understanding of its pathophysiology and factors related to epidemiologic risk, progression, protection, and prognosis,[1,5] a universally-accepted system for diagnosis and severity assessment for OA is yet to be seen. In turn, this could have adversely affected both the efficacy and range of management options for the disease, as all pharmacologic options are, at best, only symptomatic and not disease-modifying (i.e., able to reduce symptoms and at least slow the progression of the disease), and total knee arthroplasty, the only effective modality beyond symptomatic relief for OA, is both expensive and recommended only for severe cases.[1]

While plain radiography is central to the assessment of OA, various radiographic classification systems have appeared in the literature since the recognition of the disease as a distinct entity. Among this is the Kellgren-Lawrence Classification (KLC) system, also the first formal scheme for radiographic OA assessment.[6] This system facilitates the assessment of OA severity by assigning a grade from 0 to 4 to anteroposterior (AP) and lateral knee radiographs (Table 1). It has emerged as the most popular tool to objectively diagnose OA and was notably utilized in several landmark studies in the field.[7-9] Furthermore, it has been also integrated into several algorithms to guide clinical decision-making, specifically in terms of identifying patients who would benefit most from surgical intervention. However, the popularity of KLC does not excuse its potential for further improvement, which depends on timely validation and re-evaluation especially in the context of growing clinical evidence on the disease and patient-specific outcomes.[5]

Table 1. Kellgren-Lawrence Classification for Knee OA[5,6]

Grade	Description
0	No joint space narrowing or reactive changes
1	Doubtful joint space narrowing, possible osteophytic lipping
2	Definite osteophytes, possible joint space narrowing
3	Moderate osteophytes, definite joint space narrowing, some sclerosis, possible bone-end deformity
4	Large osteophytes, marked joint space narrowing, severe sclerosis, definite bone-end deformity

In the University of the Philippines – Philippine General Hospital (UP-PGH), the Osteoarthritis Multidisciplinary Clinic (OAMDC) was established by the Departments of Family and Community Medicine, Orthopedics, and Rehabilitation Medicine to facilitate comprehensive outpatient care for patients afflicted with OA. This clinic, manned primarily by resident physicians of the three participating clinical departments, utilizes the KLC system to guide diagnostic and severity assessment as well as management decisions. Thus, together with accuracy in the use of the classification system, uniformity, consistency, and reproducibility of interpretations within and among physicians in this clinic are of utmost importance. This study seeks to determine the interobserver reliability of the KLC system when utilized by resident physicians training under the component clinical departments of the UP-PGH OAMDC.

METHODOLOGY

From each participating clinical department of the UP-PGH OAMDC (Departments of Family and Community Medicine, Orthopedics, and Rehabilitation Medicine), three resident physicians were randomly selected via fishbowl draw and recruited, for a total of nine residents. Randomly-selected anteroposterior and lateral radiographs of 20 patients diagnosed with degenerative OA of the knees were used in the study. All radiographs were stored in digital form, with identifying markers hidden from the evaluators/subjects. The radiographs were then arranged beforehand into a series of images, and the sequence was preserved entirely when the series was shown individually to each reach resident. The residents were provided exactly one minute to classify each image into one of the five KLC grades before a particular image was removed from view and the next image in the series was shown.

Ratings from evaluators were recorded and logged electronically, which were then analyzed statistically using the Fleiss's kappa coefficient (κ)[10].

RESULTS

Statistical treatment of recorded KLC grading of the radiographic images by the resident physicians, using the Fleiss's kappa coefficient (κ), revealed an overall inter-observer reliability of 31.90% (κ = 0.3190 ± 0.0228 [standard error or SE]; 95% confidence interval [CI]: 0.2743, 0.3637; p-value = 0] agreement beyond chance among all selected evaluators in the study. Full agreement in given KLC grade among all subjects occurred only in one (5%) image (Image 5, unanimously given with a KLC grade of 4), while majority agreement (i.e., n ≥ 5 agreement on a particular KLC grade) was observed in all except one (95%) image (Image 7). Subgroup analyses of the inter-rater reliability of the KLC system in terms of department affiliation showed significantly variable agreement beyond chance. Resident physicians from Department A demonstrated 10.52% agreement beyond chance (κ = 0.1052 ± 0.0779; 95% CI: -0.0476, 0.2580; p-value = 0.1772) with unanimous agreement on 3 (15%) and majority agreement (i.e., n ≥ 2 of the subjects agreeing on the same KLC grade) on 14 (70%) images. Resident physicians from Department B demonstrated 56.38% agreement beyond chance (κ = 0.5638 ± 0.0844; 95% CI: 0.3984, 0.7292; p-value = 2.348×10^{-11}] with unanimous agreement on 11 (55%) and majority agreement on all (100%) images.

Residents from Department C accrued 24.75% agreement beyond chance (κ = 0.2475 ± 0.0797; 95% CI: 0.0913, 0.4037; p-value = 0.0019) with unanimous agreement on 5 (25%) and majority agreement on all (100%) images.

DISCUSSION

The Kellgren-Lawrence Classification (KLC) for knee osteoarthritis (OA) has been fully integrated into the management guidelines for the said disease in many parts of the world. This is impressive especially when the existence of at least 25 published classification schemes for knee OA is considered.[11] Aside from the historical and contemporary significance of KLC mentioned earlier, several healthcare insurance companies even require KLC documentation to approve billing coverage for total knee arthroplasty.[5] Most importantly in the context of the present study, this classification scheme is utilized as the central basis for the assessment and management of patients in the UP-PGH OAMDC. The results of the study show that, among resident physicians training under the clinical departments forming the Clinic, the agreement beyond chance in terms of classifying knee radiographs into KLC grades is well below half of the time (31.90 ± 2.28%). Analyzing the data by department affiliation of the resident physicians further revealed variable inter-observer agreement beyond chance (from 10.52 ± 7.79% in one department to 56.38 ± 0.844% in another). These discrepancies may arguably be explained by two general factors: (1) a possible effect of the differences in training, competency, and exposure to related cases across the three departments and (2) the inherent limitations of the KLC grading system. As there is no objective evidence, especially from the present study, to argue for or against the first factor, it is suggested that further efforts be directed to systematically explore this possibility. However, in light of the growing literature of criticism to the KLC system, the impact of the second factor on the results can be elaborated objectively.

It is also important to note that none of the published schemes for describing kappa (κ) values, notably including

Table 2. Kellgren-Lawrence Classification (KLC) grades were given by all resident physician subjects in the study (n = 9) to each radiograph. The two rightmost columns dichotomize the resident physicians into assigning KLC grades of less than KLC grade 2 (< 2) or at least KLC grade 2 (≥ 2)

Radiograph Number	Number of Resident Physicians						
	KLC Grade						
	0	1	2	3	4	< 2	≥ 2
1	0	0	0	6	3	0	9
2	0	4	5	0	0	4	5
3	0	2	7	0	0	2	7
4	0	7	2	0	0	7	2
5	0	0	0	0	9	0	9
6	0	0	6	2	1	0	9
7	0	0	3	3	3	0	9
8	0	5	4	0	0	5	4
9	0	6	3	0	0	6	3
10	0	0	3	5	1	0	9
11	0	0	8	1	0	0	9
12	0	0	6	3	0	0	9
13	0	5	3	1	0	5	4
14	0	0	0	5	4	0	9
15	0	0	5	4	0	0	9
16	0	7	1	1	0	7	2
17	0	0	8	1	0	0	9
18	0	0	3	6	0	0	9
19	0	0	4	4	1	0	9
20	0	0	7	2	0	0	9
κ ± SE	0.3190 ± 0.0228					0.4792 ± 0.0373	
95% CI	0.2743, 0.3637					0.4061, 0.5522	
p-value	0					0	

CI – confidence interval; κ – Fleiss's kappa coefficient; SE – standard error

Table 3. Kellgren-Lawrence Classification (KLC) grades given by resident physician subjects from Department A (n = 3) to each radiograph. The two rightmost columns dichotomize the resident physicians into assigning KLC grades of less than KLC grade 2 (< 2) or at least KLC grade 2 (≥ 2)

Radiograph Number	Number of Resident Physicians						
	KLC Grade						
	0	1	2	3	4	< 2	≥ 2
1	0	0	0	1	2	0	3
2	0	0	3	0	0	0	3
3	0	0	3	0	0	0	3
4	0	2	1	0	0	2	1
5	0	0	0	0	3	0	3
6	0	0	2	0	1	0	3
7	0	0	1	1	1	0	3
8	0	2	1	0	0	2	1
9	0	2	1	0	0	2	1
10	0	0	1	1	1	0	3
11	0	0	2	1	0	0	3
12	0	0	2	1	0	0	3
13	0	1	1	1	0	1	2
14	0	0	0	1	2	0	3
15	0	0	1	2	0	0	3
16	0	2	0	1	0	2	1
17	0	0	2	1	0	0	3
18	0	0	1	2	0	0	3
19	0	0	1	1	1	0	3
20	0	0	2	1	0	0	3
κ ± SE	0.1052 ± 0.0779					0.3464 ± 0.1291	
95% CI	-0.0476, 0.2580					0.0934, 0.5994	
p-value	0.1772					0.0073	

CI – confidence interval; κ – Fleiss's kappa coefficient; SE – standard error

Table 4. Kellgren-Lawrence Classification (KLC) grades given by resident physician subjects from Department B (n = 3) to each radiograph. The two rightmost columns dichotomize the resident physicians into assigning KLC grades of less than KLC grade 2 (< 2) or at least KLC grade 2 (≥ 2)

Radiograph Number	Number of Resident Physicians						
	KLC Grade						
	0	1	2	3	4	< 2	≥ 2
1	0	0	0	3	0	0	3
2	0	2	1	0	0	2	1
3	0	0	3	0	0	0	3
4	0	3	0	0	0	3	0
5	0	0	0	0	3	0	3
6	0	0	1	2	0	0	3
7	0	0	1	2	0	0	3
8	0	1	2	0	0	1	2
9	0	3	0	0	0	3	0
10	0	0	1	2	0	0	3
11	0	0	3	0	0	0	3
12	0	0	2	1	0	0	3
13	0	3	0	0	0	3	0
14	0	0	0	3	0	0	3
15	0	0	1	2	0	0	3
16	0	3	0	0	0	3	0
17	0	0	3	0	0	0	3
18	0	0	1	2	0	0	3
19	0	0	1	2	0	0	3
20	0	0	3	0	0	0	3
κ ± SE	0.5638 ± 0.0844					0.8222 ± 0.1291	
95% CI	0.3984, 0.7292					0.5692, 1.0753	
p-value	2.348 x 10^{-11}					1.904 x 10^{-10}	

CI – confidence interval; κ – Fleiss's kappa coefficient; SE – standard error

Table 5. Kellgren-Lawrence Classification (KLC) grades given by resident physician subjects from Department C (n = 3) to each radiograph. The two rightmost columns dichotomize the resident physicians into assigning KLC grades of less than KLC grade 2 (< 2) or at least KLC grade 2 (≥ 2)

Radiograph Number	Number of Resident Physicians						
	KLC Grade						
	0	1	2	3	4	< 2	≥ 2
1	0	0	0	2	1	0	3
2	0	2	1	0	0	2	1
3	0	2	1	0	0	2	1
4	0	2	1	0	0	2	1
5	0	0	0	0	3	0	3
6	0	0	3	0	0	0	3
7	0	0	1	0	2	0	3
8	0	2	1	0	0	2	1
9	0	1	2	0	0	1	2
10	0	0	1	2	0	0	3
11	0	0	3	0	0	0	3
12	0	0	2	1	0	0	3
13	0	1	2	0	0	1	2
14	0	0	0	1	2	0	3
15	0	0	3	0	0	0	3
16	0	2	1	0	0	2	1
17	0	0	3	0	0	0	3
18	0	0	1	2	0	0	3
19	0	0	2	1	0	0	3
20	0	0	2	1	0	0	3
κ ± SE	0.2475 ± 0.0797					0.2708 ± 0.1291	
95% CI	0.0913, 0.4037					0.0178, 0.5239	
p-value	0.0019					0.0359	

CI – confidence interval; κ – Fleiss's kappa coefficient; SE – standard error

those of Landis & Koch[12] and the Byrt[13] (which criticized the former), was used in interpreting the κ values obtained in the study. This is because (1) such schemes are (1) not validated especially in the clinical context being considered in the study, and (2) are artificially subject to prevalence, bias, number of categories and subjects,[13] thus reducing the value of descriptive labeling of agreements as "poor," "slight," "fair," "moderate," "good," "very good," "substantial," "excellent," and "almost perfect."

While KLC has been independently validated by several authors and groups,[5] these studies frequently reveal, implicitly or explicitly, the limitations of this classification system. For instance, published observer reliability values for KLC are highly variable across reports.[5] The original exposition of the classification scheme showed that among the joints examined for OA, the highest inter-observer and the second-highest intra-observer agreement correlation were attributed to knee AP radiographs,[6] but the clinical and statistical relevance of these findings were not extensively discussed and were rather "assumed." Another assumption regarding KLC is that it demonstrates a predictable trend of progression of OA from KLC grade 0 (no narrowing or reactive change) then KLC grade 1 (osteophyte formation) to KLC grade 4 (bone end alteration) – a supposition that is even yet to be confirmed even at present, even when this perspective is intuitively accepted for monitoring disease progression.[14] Further criticism is also directed toward the seemingly preferential dependence of the KLC system to osteophyte formation over joint space narrowing in terms of grade assignment, which often presents a diagnostic challenge in symptomatic OA patients with radiographs showing cartilage loss without osteophyte formation.[15]

Another contention on the validity of the KLC system is the rampancy of its versions, with subtle to obvious descriptive alterations, as they appear in the literature.[11,16] A total of five – the original version (as used in the present study) and four other alternatives – was previously documented by Schiphof et al.,[11] who then subsequently sought to assess the diagnostic impact of this plurality and determine the version with the highest association of patient-reported knee pain symptoms.[16] They found that while the difference in the number of cases classified as OA is little to small across all versions with KLC grade 2 – labeled as "definite/ mild osteoarthritis" versus KLC grade < 2 as "none/possible osteoarthritis" – as the diagnostic cut-off, the association of the original KLC version with self-reported pain symptom was the most relevant numerically and clinically.[16] When this dichotomy (KLC grade < 2 and KLC grade ≥ 2) is used in the findings of the present study, secondary analyses revealed significant improvement of agreement beyond chance both overall (increased by 16% to 47.92 ± 3.73%; p-value = 0) and among all department affiliations: 3-fold increase in Department A (to 34.64 ± 12.91%; p-value = 0.0073), 26-point increase in Department B (to 82.22 ± 12.91%; p-value = 1.904×10^{-10}) and 3-point increase in Department C (to 27.08 ± 12.91%; p-value = 0.0359). Thus, a formal adaptation of this dichotomous scheme especially in the context of first encounter-impression of possible OA cases is suggested.

The findings of the present study have significant implications for the system currently in place in UP-PGH (especially the OAMDC) to manage OA patients. Perhaps, further strengthening the recommendation from this paper to investigate and, in turn, to improve the clinical capacity of the institution according to the latest evidence is the outcome of the prospective longitudinal cohort study conducted by the United States-based Multicenter ACL Revision Study (MARS) Group, which (1) compared the interobserver reliability of six radiographic OA classification systems including KLC and (2) measured the degree of correlation between these classification systems and arthroscopic findings.[17] This study demonstrated that another classification scheme, the International Knee Documentation Committee (IKDC) system, generated the highest interobserver reliability and the reflected best the physical state of the joint as observed through arthroscopy. Furthermore, the Rosenberg view (45° posteroanterior flexion weight-bearing view) was shown to provide better inter-observer reliability than the routinely-used knee AP views.

CONCLUSION AND RECOMMENDATIONS

The results of the present study show significant variability of agreement beyond chance in terms of objective assessment of knee osteoarthritis (OA) using the Kellgren-Lawrence Classification (KLC) system both within and among residents of clinical departments forming the UP-PGH OAMDC. In light of these findings and the implications of the literature reviewed earlier, further investigation and other endeavors to improve the accuracy and uniformity in OA diagnosis and severity in the institution are warranted.

Statement of Authorship

Both authors participated in data collection and analysis, and approved the final version submitted.

Author Disclosure

Both authors declared no conflicts of interest.

Funding Source

This paper was self-funded by the authors.

REFERENCES

1. Martel-Pelletier J, Barr AJ, Cicuttini FM, Conaghan PG, Cooper C, Goldring MB et al. Osteoarthritis. Nat Rev Dis Primers. 2016;2: 16072. [DOI: 10.1038/nrdp.2016.72]
2. Fransen M, Bridgett L, March L, Hoy D, Penserga E, Brooks P. The epidemiology of osteoarthritis in Asia. Int J Rheum Dis. 2011;14(2):113-21. [DOI: 10.1111/j.1756-185X.2011.01608.x]
3. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systemic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2163-96. [DOI: 10.1016/ S0140-6736(12)61729-2]
4. Zamora G, Salido EO, Penserga EG. Clinical profile of Filipino patients with osteoarthritis seen at two arthritis clinics. Int J Rheum Dis. 2012;15(4):399-406. [DOI: 10.1111/j.1756-185X.2012.01758.x]
5. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res. 2016;474:1886-93. [DOI: 10.1007/s11999-016-4732-4]
6. Kellgren JH, Lawrence JS. Radiological assessment of osteoarthrosis. Ann Rheum Dis. 1957;16:494-502.
7. Felson DT, Naimark A, Anderson J, Kazis L, Castelli W, Meenan RF. The prevalence of knee osteoarthritis in the elderly: the Framingham osteoarthritis study. Arthritis Rheum. 1987;30:914-8.
8. Bagge E, Bjelle A, Valkenburg HA, Svanborg A. Prevalence of radiographic osteoarthritis in two elderly European populations. Rheumatol Int. 1992;12:33-8.
9. Scott WW, Lethbridge-Cejku M, Reichle R, Wigley FM, Tobin JD, Hochberg MC. Reliability of grading scales for individual radiographic features of osteoarthritis of the knee: the Baltimore longitudinal study of aging atlas of knee osteoarthritis. Invest Radiol. 1993;28:497-501.
10. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378-82.
11. Schiphof D, de Klerk BM, Koes BW, Bierma-Zeinstra S. Good reliability, questionable validity of 25 classification criteria of knee osteoarthritis: a systematic appraisal. J Clin Epidemiol. 2008;61: 1205-15. [DOI: 10.1016/j.jclinepi. 2008.04.003]
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.
13. Byrt T. How good is that agreement?. Epidemiology. 1996;7(5):561.
14. Braun HJ, Gold GE. Diagnosis of osteoarthritis: imaging. Bone. 2012;51(2):278-88. [DOI: 10.1016/j.bone.2011.11.019]
15. Spector TD, Cooper C. Radiographic assessment of osteoarthritis in population studies: whither Kellgren and Lawrence?. Osteoarthritis Cartilage. 1993;1(4):203-6.
16. Schiphof D, de Klerk BM, Kerkhof HJM, Hofman A, Koes BW, Boers M et al. Impact of different descriptions of the Kellgren and Lawrence classification criteria on the diagnosis of knee osteroarthritis. Ann Rheum Dis. 2011;70:1422-7. [DOI: 10.1136/ard.2010.147520]
17. The MARS (Multicenter ACL Revision Study) Group. Osteoarthritis classification scales: interobserver reliability and arthroscopic correlation. J Bone Joint Surg Am. 2014;96:1145-51. [DOI: 10.2106/ JBJS.M.00929]