The accuracy and reproducibility of the KL grade evaluation using Chat GPT in the Asian population

Takanori Iriuchishima[1,2], Sang-Yang Lee[1], Satoshi Okamoto[1], Daika Sato[1]
[1]Department of Orthopedic Surgery, Keijinkai Shiroyama Hospital, Ota, Japan
[2]Department of Functional Morphology, Nihon University School of Medicine, Tokyo, Japan
Email of Presenting Author: sekaiwoseisu@yahoo.co.jp

Disclosures: Takanori Iriuchishima (N), Sang-Yang Lee (N), Satoshi Okamoto (N), Taika Sato (N)

INTRODUCTION: The recent advancement of artificial intelligence (AI) has been remarkable. AI is being increasingly applied not only in manufacturing, business, and information technology, but also in the field of medicine. In particular, AI has already been used clinically in diagnostic imaging, especially in radiology and pathology. In the field of orthopedic surgery, however, the application of AI for the diagnosis of bone and joint diseases or trauma remains limited. Nonetheless, AI based on deep learning may be well suited for orthopedic diagnostic imaging, as hard tissues such as bones and joints tend to be easier to analyze than soft tissue organs like the lungs or intestines. The Kellgren-Lawrence (KL) grading system is a widely used clinical classification for knee osteoarthritis (OA) based on anterior-posterior knee radiographs. It assesses the severity of knee OA by evaluating joint space narrowing, osteophyte (bony spur) formation, and joint deformity. Based on the KL grade, orthopedic surgeons typically select appropriate treatment strategies for knee OA. Although the KL grading system is well known among orthopedic surgeons, if AI could accurately assess KL grades, it could enable non-specialist physicians to diagnose knee OA and determine suitable treatment options. ChatGPT, a generative AI developed in 2022, is now widely used around the world. It can answer questions and provide information using internet resources, and it also has capabilities in image generation and evaluation. While there is potential for ChatGPT to be used in medical contexts, its clinical utility has not yet been thoroughly investigated. The purpose of this study was to evaluate the accuracy and reproducibility of KL grade assessment using ChatGPT. The hypothesis was that ChatGPT could diagnose KL grades with high accuracy and reproducibility compared to evaluations made by orthopedic surgeons.

METHODS: Eighty knees from 80 subjects (61 female and 19 male; mean age 65 ± 19 years) with knee pain were included in this study. Subjects diagnosed with rheumatoid arthritis, a history of trauma, or a history of knee surgery were excluded. At the initial clinic visit, anterior-posterior (A-P) knee radiographs were obtained for all participants. First, all radiographs were evaluated by a knee surgeon with over 20 years of experience. The radiographs were classified into Kellgren-Lawrence (KL) grades 0 to 4 by the surgeon. The same A-P radiographs were then uploaded to ChatGPT and the model was prompted with the question: “What is the Kellgren-Lawrence grade of this knee and the recommended treatment?” The most severe KL grade reported by ChatGPT was recorded as its final response. ChatGPT was asked the same question twice on separate days. KL grades were analyzed both as individual grades (0–4) and as categorical classifications: mild OA (KL 0–2) and severe OA (KL 3–4). The diagnoses by the knee surgeon and ChatGPT were compared.
For statistical analysis, Cohen’s kappa coefficient was used to assess agreement. Intra-rater reliability (ChatGPT vs. ChatGPT) and inter-rater reliability (ChatGPT vs. knee surgeon) were evaluated using intraclass correlation coefficients (ICCs) calculated with SPSS software (version 29.0, IBM).

RESULTS SECTION: The KL grade diagnoses (grades 0 to 4) made by the knee surgeon were distributed as follows: 8, 5, 10, 28, and 29 cases, respectively (Table 1). ChatGPT evaluated KL grade based on features such as joint space narrowing, osteophyte formation, bone sclerosis, and joint deformity. In the standard five-grade KL classification (0–4), the agreement rates between the knee surgeon and ChatGPT were 65% (first trial) and 50% (second trial) (Table 1). The kappa coefficient for ChatGPT's test–retest reliability (first vs. second trial) was 0.31, and the intraclass correlation coefficient (ICC) was 0.745 (95% CI: 0.628–0.828). The kappa coefficients for agreement between ChatGPT and the surgeon were 0.52 (first trial) and 0.28 (second trial), with corresponding ICCs of 0.811 (95% CI: 0.721–0.875) and 0.659 (95% CI: 0.499–0.773), respectively. When KL grades were grouped into two categories—grades 0–2 (mild OA) and grades 3–4 (severe OA)—the agreement between the surgeon and ChatGPT increased to 85% (first trial) and 80% (second trial). The kappa coefficient for agreement between ChatGPT's two trials was 0.54, with an ICC of 0.538 (95% CI: 0.363–0.677). The kappa coefficients for agreement between ChatGPT and the surgeon were 0.64 (first trial) and 0.46 (second trial), and the corresponding ICCs were 0.646 (95% CI: 0.498–0.758) and 0.458 (95% CI: 0.269–0.614), respectively (Table 2). ChatGPT suggested treatment recommendations for knee OA based on the KL grade. For KL grade 0, it recommended “no specific treatment is needed” or “if asymptomatic, no intervention is needed.” For mild OA (KL grades 1 or 2), ChatGPT advised that “treatment is usually non-surgical at first, focusing on symptom relief and functional improvement” or “conservative treatment is best at this stage.” For KL grade 3 knees, it suggested symptom-based surgical interventions such as arthroscopic debridement, high tibial osteotomy (HTO), or knee arthroplasty. When KL grade 4 was diagnosed, ChatGPT recommended HTO or partial/total knee arthroplasty (UKA or TKA).

DISCUSSION: ChatGPT was able to evaluate the KL grade of the knee simply by uploading a radiographic image. Its responses showed a significant correlation with the diagnoses made by the knee surgeon. However, the reproducibility of ChatGPT’s responses was relatively low. In many cases, the KL grade was underestimated by AI compared to the surgeon’s diagnosis. Although ChatGPT was able to evaluate individual features such as osteophyte formation, joint space narrowing, and knee deformity, it appeared to lack the ability to integrate these parameters in the same manner as experienced surgeons. As the capabilities of ChatGPT continue to improve, its diagnostic performance may approach that of human physicians in the near future.

SIGNIFICANCE/CLINICAL RELEVANCE: ChatGPT was able to evaluate the KL grade of the knee simply by uploading a radiographic image. Its responses showed a significant correlation with the diagnoses made by the knee surgeon. However, the reproducibility of ChatGPT’s responses was relatively low.

Table 1.

KL grade	0	1	2	3	4
Knee Surgeon’s diagnosis	8	5	10	28	29 (cases)
Chat GPT-1st trial	2	11	12	29	26
Chat GPT-2nd trial	6	6	3	25	40

Table 2.

	Kappa coefficient	ICC
Chat GPT intra-rater reliability		
KL grade 0-4	0.31	0.745 (0.628-0.828)
KL grade 0-2/3-4	0.54	0.538 (0.363-0.677)
Chat GPT-Surgeon inter-rater reliability (1st/2nd trial)		
KL grade 0-4	0.52/0.28	0.811/0.743
KL grade 0-2/3-4	0.64/0.46	0.646/0.458