Original Paper

Performance of a Large Language Model for Predicting Clinical and Radiographic Outcomes in Knee Osteoarthritis: Single-Center Prospective Observational Study

Zhiqiang Li, BSc; Yuhan Zhang, BSc; Mingwei Li, BSc, MD

Department of Rheumatology and Immunology, Fuxing Hospital, Capital Medical University, Beijing, China

Corresponding Author:
Mingwei Li, BSc, MD
Department of Rheumatology and Immunology
Fuxing Hospital
Capital Medical University
No. 20A, Fuxingmenwai Street, Xicheng District
Beijing, 100038
China
Phone: 86 18747772761
Email: dr_limingwei@126.com

Abstract

Background: Large language models (LLMs) can extract information from unstructured clinical text, but it remains unclear whether their ability to assess current status translates into reliable predictions of longitudinal outcomes.

Objective: This study aimed to evaluate DeepSeek's ability to assess baseline clinical and radiographic status in knee osteoarthritis (KOA), its performance in predicting 6-month outcomes, and its interrun reproducibility.

Methods: This study used a single-center, prospective observational cohort design to conduct a blinded, retrospective evaluation of the LLM. A total of 128 patients with primary KOA were enrolled, of whom 111 (86.7%) completed the 6-month follow-up and were included in the primary analysis. DeepSeek-V4-Flash and a rheumatologist independently assessed KOA diagnoses, baseline Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scores, and Kellgren-Lawrence (KL) grading based on the same standardized interview records and imaging reports and predicted 6-month outcomes. The LLM was run independently 3 times. The actual outcomes determined by another senior physician served as the reference, and the results were compared with the physician's predictions and the persistence model.

Results: The LLM showed high baseline diagnostic agreement with the physician (105/111, 94.6%) and good agreement on the WOMAC and KL, but 6-month predictions (WOMAC mean absolute error=5.92-6.58; KL exact agreement=74/111, 66.7%-76/111, 68.5%) were worse than those of the physician comparator (mean absolute error=3.86; 96/111, 86.5%) and not superior to those of the persistence model (mean absolute error=5.43). Despite high interrun intraclass correlation coefficients (0.982 and 0.965), exact agreement between runs was poor (35/111, 31.5% and 7/111, 6.3%).

Conclusions: The evaluated DeepSeek configuration may support structured baseline KOA assessment. However, its 6-month predictive accuracy and patient-level stability remain limited. Prognostic applications require evidence of added value over a persistence model.

(JMIR Med Inform 2026;14:e107092) doi: 10.2196/107092

KEYWORDS

knee osteoarthritis; large language model; DeepSeek; Western Ontario and McMaster Universities Osteoarthritis Index; WOMAC; Kellgren-Lawrence classification; outcome prediction

XSL•FO
RenderX

Introduction

Background

Knee osteoarthritis (KOA) is a chronic musculoskeletal disease with a highly heterogeneous clinical course; even among patients with similar baseline symptoms or radiographic severity, subsequent changes in pain, function, and structure may follow different trajectories [1,2]. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and the Kellgren-Lawrence (KL) classification are commonly used measures in clinical assessments and longitudinal studies. Accurately estimating future changes in these measures aids in risk stratification, follow-up planning, and individualized management. In response to this need, various disease-specific models developed in recent years have demonstrated the potential to predict pain and structural progression in KOA [3-6].

Large language models (LLMs) can extract clinical information from unstructured data—such as symptom descriptions, medical histories, treatment histories, and imaging reports—and perform tasks such as medical question answering, clinical reasoning, and information summarization [7,8]. In the field of osteoarthritis, studies have shown that LLMs can identify hip osteoarthritis or KOA from questionnaires, generate diagnostic and treatment recommendations generally consistent with those of specialists using anonymous medical records, and convert free-text knee magnetic resonance imaging reports into structured scores [9-11]. Studies indicate that LLMs can extract information relevant to a patient's current clinical condition from clinical narratives and convert it into clinically meaningful, structured judgments; however, it remains unclear whether they can be further used to predict individual outcomes.

Model-Based Longitudinal Outcome Prediction

Although models such as DeepSeek have demonstrated strong clinical reasoning capabilities in diagnosis and treatment decision-making for multidisciplinary cases [12], longitudinal outcome prediction differs in nature from tasks involving the identification, induction, or classification of existing information. Therefore, the evaluation of LLMs' longitudinal predictive capabilities must be based on outcomes observed in real-world follow-up data. Previous research on medical LLMs has emphasized the need to expand the scope of evaluation beyond simple accuracy to include real-world clinical data, workflow adaptability, and multidimensional performance metrics [13,14]. Furthermore, generative models may produce different outputs for the same input [15]. It is still necessary to validate whether LLMs can reliably predict longitudinal outcomes for KOA based on clinical narratives while maintaining stable outputs in a real-world follow-up cohort.

This study aimed to evaluate DeepSeek's ability to identify KOA and assess baseline WOMAC scores and report-derived KL grades from standardized clinical interview records and descriptive radiology reports. We also aimed to compare its predictions of 6-month WOMAC scores and report-derived KL grades with those of one rheumatologist and a persistence model and assess the reproducibility of its outputs across 3 independent runs.

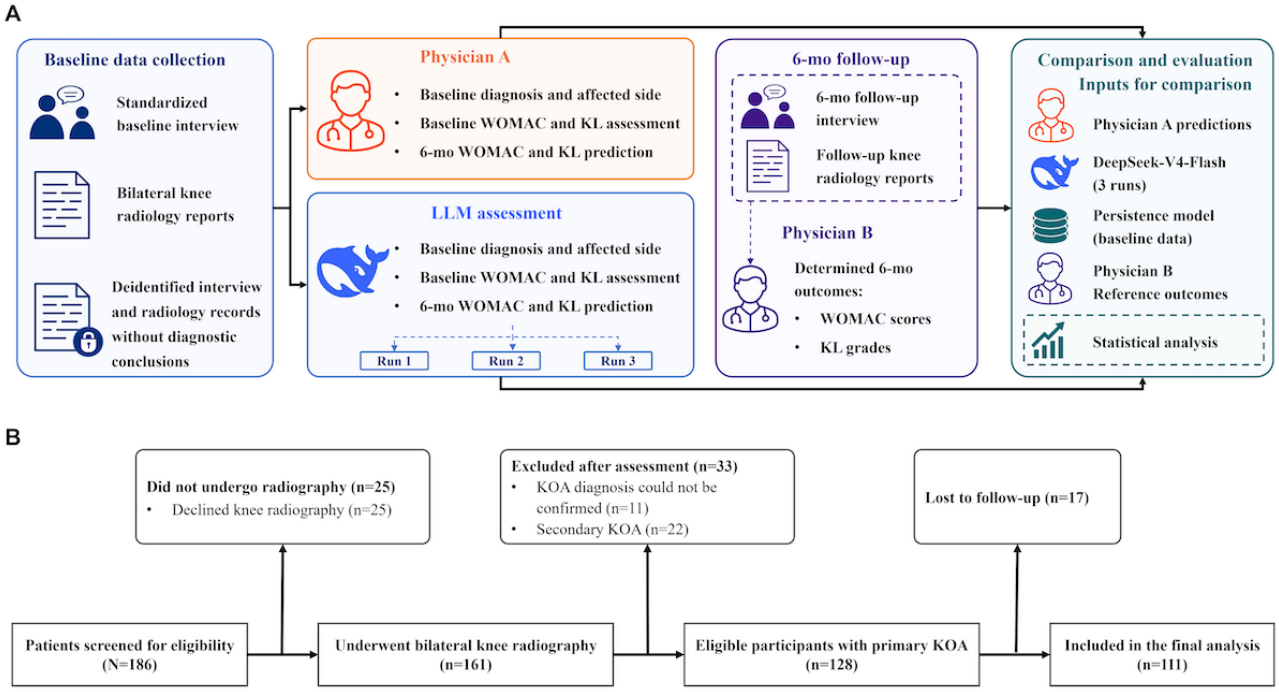
Methods

Study Design and Participants

This was a single-center prospective observational cohort study. A rheumatologist predicted the 6-month outcomes at baseline. The LLM evaluation was conducted retrospectively after completion of follow-up. Patients were consecutively enrolled from January to November 2025 in the Rheumatology and Immunology Department and the Orthopaedics Department at Fuxing Hospital, which is affiliated with Capital Medical University. Patients were eligible for enrollment if they had a definitive diagnosis of primary KOA according to the Guidelines for the Diagnosis and Treatment of Osteoarthritis (2018 edition) [16], completed a standardized baseline interview, and underwent anteroposterior and lateral radiographs of both knees. Patients who could not be definitively diagnosed, those with secondary KOA, or those with insufficient baseline data to complete the physician and LLM assessments were excluded. The primary analysis was limited to patients who completed follow-up and had complete outcome data. The overall study design and evaluation process are shown in Figure 1A. Reporting of the LLM evaluation was guided by the TRIPOD-LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model version) checklist [17].

XSL•FO
RenderX

Figure 1. Study design and participant flow. (A) Baseline data collection, parallel assessment, and prediction by rheumatologist A and the large language model (LLM), 6-month reference outcomes determined by rheumatologist B, and comparison with the persistence model. (B) Participant screening and final inclusion. KL: Kellgren–Lawrence; KOA: knee osteoarthritis; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.



Clinical Data Collection and WOMAC Assessment

Researchers trained according to a standardized protocol conducted baseline assessments using a standardized interview guide and recorded patients' responses in natural language. The interview guide and recording template are provided in Multimedia Appendix 1. Prior treatment was defined as receiving at least one KOA-related intervention within the 3 months prior to baseline, including medication, intra-articular injections, physical or rehabilitation therapy, and the use of assistive devices. Regular exercise was defined as exercising 4 or more times per week for 30 or more minutes per session; frequent stair climbing was defined as climbing 4 or more times per day, with each session involving 30 or more steps or 2 flights of stairs; and weight control was defined as actively undertaking dietary or exercise interventions to manage body weight.

The 96-point WOMAC scale was used to assess knee pain, stiffness, and physical function [18,19]. At the time of submission to the physician and the LLM for evaluation, all interview data were retained as natural language text without numerical coding. Responses containing explicit severity ratings were scored according to the WOMAC scale's categories from 0 to 4 points. When descriptions were ambiguous, the physician and the LLM scored them independently based on semantic analysis. The same procedure was followed for the 6-month follow-up, which was conducted via outpatient visits, hospital admissions, or telephone calls, with the follow-up window set at 6 months ±15 days after baseline.

Radiographic Assessment

Baseline and 6-month follow-up examinations followed a standardized protocol, with anteroposterior and lateral radiographs of both knees acquired using the same digital radiography system (Discovery XR656; GE Healthcare). The imaging data were independently reviewed by 2 radiologists with 5 and 8 years of experience, who each prepared descriptive reports. In cases of disagreement, a third radiologist with 15 years of experience reviewed the images and made the final determination. The radiologists were blinded to the time of examination, physician predictions, LLM outputs, and clinical outcomes; they described only radiographic findings and did not directly assess the KL grade. The KL grades at baseline and 6 months were determined by the respective clinical evaluators based on interview data and radiology reports.

The baseline assessment, based on comprehensive physician interview records and imaging reports, was used to confirm the diagnosis of KOA. For patients with bilateral KOA, the target knee was selected according to predefined rules: priority was given to the side with a higher KL grade; if the grades were the same, the side with more severe symptoms was selected; if they were still the same, the right knee was selected. Once the target knee was determined, it remained unchanged, and all baseline and follow-up outcomes were assessed based on that side.

Physician Assessment and Prediction

A board-certified rheumatologist with 10 years of clinical experience (physician A) independently reviewed standardized baseline interview records and imaging reports with diagnostic conclusions redacted to establish the diagnosis of KOA; assess baseline WOMAC subscores, total scores, and KL grading; and predict corresponding 6-month outcomes. All assessments and predictions were based solely on baseline data; physician A remained blinded to the actual 6-month outcomes and LLM outputs. The 6-month outcomes were independently assessed by another rheumatologist (physician B) with 25 years of clinical experience based on follow-up interview records and imaging

XSL•FO
RenderX

reports from the same period, including the WOMAC subscores, total score, and KL grading. Physician B was blinded to physician A's assessments and predictions, as well as the LLM outputs.

LLM Assessment and Prediction

On June 19 and 20, 2026, DeepSeek was accessed via the official web interface in Instant Mode (DeepSeek-V4-Flash) [20]. Deep Thinking and Smart Search were left enabled by default. No search activity was visible. The exact back-end build and sampling parameters were unverified. The model was not fine-tuned. Both the LLM and physician A used exactly the same deidentified baseline data, including standardized interview records and imaging reports with diagnostic conclusions removed; neither had access to the original x-ray images. A standardized prompt was used, instructing the model to determine the most likely diagnosis; assess the baseline WOMAC subscales, total score, and KL classification; and predict the corresponding 6-month outcome. The affected side was extracted from the diagnostic responses provided by both physician A and the LLM; cases where the diagnosis was correct but the side was not specified were classified as "side inconsistency." The prompt specified the use of the WOMAC scale from 0 to 96 points and required the output to be a single, unambiguous numerical value or grade. The standardized interview template and complete 4-step prompt sequence are provided in Multimedia Appendix 1.

To evaluate interrun reproducibility, 3 independent runs were conducted for 111 patients. For each run, a new session was created using the same input and prompts, with consistent case order and no prior run information provided. Reruns were required only for out-of-range WOMAC scores or an incorrect scale version; none were needed. All 333 LLM outputs were structured, extracted, and entered by an independent evaluator and then cross-checked against the original responses one by one by a second evaluator.

Radiographic progression was defined as an increase of 1 or more grades in report-derived KL at 6 months relative to the baseline grade assessed by physician A. The direction of change in the WOMAC score was classified as a decrease (<0), no change (0), or an increase (>0) based on the difference between the 6-month total score and the baseline total score. In the persistence model, the baseline WOMAC score and KL classification assessed by physician A were used as the corresponding predictors for the 6-month values.

Statistical Analysis

Continuous variables are expressed as means and SDs or medians and IQRs, and categorical variables are expressed as frequencies and percentages. Paired comparisons between baseline and 6-month data were performed using the Wilcoxon signed-rank test. Correlation between continuous outputs was assessed using the Pearson correlation coefficient. Agreement was evaluated using the intraclass correlation coefficient (ICC(2,1); 2-way random effects, absolute agreement, and single measurement) and Bland-Altman plots. Ordered categorical outputs were evaluated using exact agreement rates and the linear-weighted κ coefficient. Predictive performance was characterized using the mean absolute error (MAE), root mean squared error (RMSE), and signed error, and calibration was assessed using the calibration slope and calibration curves. The 95% CI for the MAE was estimated using 5000 bias-corrected and accelerated bootstrap repetitions.

The 3 LLM run results and the prediction errors of different methods were compared using the Friedman test; when overall differences were significant, the Wilcoxon signed-rank test was performed using the Holm correction method, and the performance of imaging progression identification was evaluated using sensitivity, specificity, positive predictive value, negative predictive value, and balanced accuracy. Overall and pairwise comparisons between methods were performed using the Cochran *Q* test and the exact McNemar test, respectively. A linear mixed-effects model with patient-specific random intercepts was used to analyze factors associated with the absolute prediction error rates of the LLM. All tests were 2 sided; $P<.05$ was considered statistically significant. Statistical analysis and graphing were performed using SPSS Statistics (version 27.0; IBM Corp) and Prism (version 11; GraphPad Software), respectively.

Missing data were assessed using the Little missing completely at random (MCAR) test and comparisons of baseline characteristics by follow-up status. The primary analysis included 111 completers; inverse probability weighting (IPW) was used for sensitivity analysis. Multiple imputation was considered but not performed.

Ethical Considerations

This study was approved by the ethics committee of Fuxing Hospital, Capital Medical University (approval 2025FXHEC-KSP088). Written informed consent was obtained from all participants. Deidentified baseline interview records and radiology reports were used for the physician and LLM evaluations. Participants received no compensation for taking part in the study.

Results

Characteristics of the Study Participants and 6-Month Outcomes

This study screened a total of 186 patients; 25 refused x-ray examination, 11 could not be definitively diagnosed with KOA, and 22 had secondary KOA. A total of 128 patients met the inclusion criteria, of whom 17 were lost to follow-up; ultimately, 111 patients were included in the analysis (Figure 1B). The mean age of the analyzed patients was 56.8 (SD 11.1) years; 60.4% (67/111) were female, and 63.1% (70/111) had bilateral knee involvement (Table 1). The baseline WOMAC total score was 19 (IQR 13-25) points, and at 6 months, it was 21 (IQR 13.5-27) points; the difference was not statistically significant ($P=.16$; Table 2). At both baseline and 6 months, KL grade 2 was the most common, accounting for 69.4% (77/111) and 66.7% (74/111), respectively. During follow-up, 13.5% (15/111) of the patients experienced radiographic progression, 85.6% (95/111) maintained stable grading, and 0.9% (1/111) experienced a decline in grading.

XSL•FO
RenderX

Table 1. Baseline characteristics of the study participants (N=111).

Characteristics	Values
Demographic and clinical characteristics	
Age (y), mean (SD)	56.8 (11.1)
Female sex, n (%)	67 (60.4)
BMI (kg/m^2), mean (SD)	25.69 (2.28)
Affected knee, n (%)	
Right	28 (25.2)
Left	13 (11.7)
Bilateral	70 (63.1)
Occupation, n (%)	
Worker	37 (33.3)
Farmer	30 (27)
Public institution employee	23 (20.7)
Retired	17 (15.3)
Athlete	4 (3.6)
Treatment characteristics, n (%)	
KOA[a]-related treatment	71 (64)
Oral NSAIDs[b]	39 (35.1)
Topical NSAIDs	40 (36)
Other analgesics	14 (12.6)
Intra-articular injection	19 (17.1)
Rehabilitation treatment	53 (47.7)
Behavioral characteristics, n (%)	
Assistive device use	22 (19.8)
Regular exercise	77 (69.4)
Frequent stair use	62 (55.9)
Weight control	51 (45.9)

[a]KOA: knee osteoarthritis.

[b]NSAID: nonsteroidal anti-inflammatory drug.

XSL•FO
RenderX

Table 2. Observed clinical and radiographic outcomes at baseline and the 6-month follow-up (N=111).

Outcome	Baseline	6-mo follow-up	*P* value
WOMAC[a]outcomes, median (IQR)			
Total score (range 0-96)	19 (13-25)	21 (13.5 to 27)	.16
Pain score (range 0-20)	4 (3-6)	5 (3 to 6)	.46
Stiffness score (range 0-8)	2 (1-2)	2 (1 to 2)	.22
Physical function score (range 0-68)	13 (9-18)	15 (9 to 19)	.12
Change	—[b]	2 (−4 to 5)	—
WOMAC change in direction, n (%)			
Decreased	—	45 (40.5)	—
Unchanged	—	3 (2.7)	—
Increased	—	63 (56.8)	—
KL[c] grade, n (%)			<.001
Grade 0	1 (0.9)	1 (0.9)	>.99
Grade 1	11 (9.9)	6 (5.4)	.06
Grade 2	77 (69.4)	74 (66.7)	.61
Grade 3	22 (19.8)	29 (26.1)	.07
Grade 4	0 (0)	1 (0.9)	>.99
KL grade change, n (%)			
Decreased	—	1 (0.9)	—
Unchanged	—	95 (85.6)	—
Increased	—	15 (13.5)	—
Radiographic progression, n (%)	—	15 (13.5)	—

[a]WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

[b]Not applicable. Changes and progression were assessed relative to baseline.

[c]KL: Kellgren-Lawrence.

Both 6-month outcomes were missing for 13.3% (17/128) of the participants, who could not be contacted or did not return. Participants lost to follow-up had lower baseline WOMAC scores and less regular exercise (*P*=.009 and *P*=.002, respectively; Table S1 in Multimedia Appendix 2). The Little test yielded χ^2_3=7.1 (*P*=.07). The main WOMAC performance ordering was unchanged in the IPW sensitivity analysis (Table S2 in Multimedia Appendix 2).

Cross-Sectional Diagnosis and Baseline Assessment Findings

Across the 3 LLM runs, diagnostic outputs were identical. KOA diagnostic agreement with physician A was 94.6% (105/111), and agreement for both diagnosis and affected side was 88.3% (98/111; Table 3). The baseline WOMAC total scores from the 3 LLM runs were highly correlated with the physician assessments (Pearson *r*=0.959-0.976; Figure 2A-C) and exhibited high consistency (ICC=0.956-0.973), with an MAE of 1.42 to 1.94 points and a mean signed difference of +0.81 to +1.04 points (Table 3). Scatterplots and Bland-Altman analyses showed that the 3 runs exhibited similar consistency characteristics (Figure 2D-F). The exact agreement rate between the 3 LLM runs and physician A's baseline KL grading ranged from 80.2% (89/111) to 83.8% (93/111), and none of the grading discrepancies exceeded 1 grade. The linear-weighted κ coefficients ranged from 0.680 to 0.727, and the MAE in grading ranged from 0.162 to 0.198 (Table 3).

XSL•FO
RenderX

Table 3. Cross-sectional diagnostic and baseline assessment performance of the 3 large language model (LLM) runs (N=111)[a].

Domain and metric	LLM run 1	LLM run 2	LLM run 3
Diagnosis, n (%)			
KOA[b] diagnosis concordance	105 (94.6)	105 (94.6)	105 (94.6)
Diagnosis and affected side agreement	98 (88.3)	98 (88.3)	98 (88.3)
Baseline WOMAC[c] total score (range 0-96)			
LLM-assessed mean (SD)	21.08 (10.38)	21.25 (10.26)	21.03 (10.13)
Pearson *r*	0.976	0.962	0.959
ICC[d](2,1; absolute agreement)	0.973	0.957	0.956
MAE[e] (points)	1.42	1.94	1.91
Signed difference (points), mean (SD)	+0.86 (2.25)	+1.04 (2.85)	+0.81 (2.94)
Baseline KL[f] grade			
Exact agreement, n (%)	89 (80.2)	93 (83.8)	92 (82.9)
Difference of ≤1 grade, n (%)	111 (100)	111 (100)	111 (100)
Linear-weighted κ	0.680	0.727	0.710
Grade MAE	0.198	0.162	0.171

[a]Knee osteoarthritis diagnosis concordance denotes baseline agreement between the LLM and physician A in patients with established knee osteoarthritis.

[b]KOA: knee osteoarthritis.

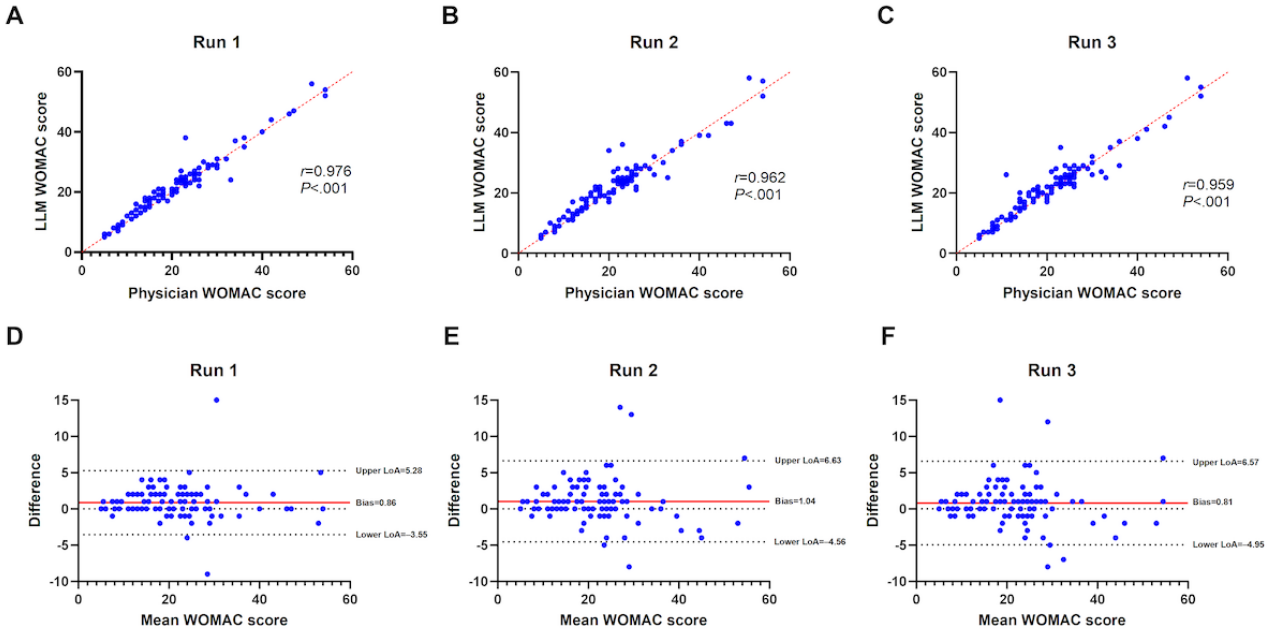
[c]WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index.

[d]ICC: intraclass correlation coefficient.

[e]MAE: mean absolute error.

[f]KL: Kellgren-Lawrence.

Figure 2. Agreement between large language model (LLM)– and physician-assessed baseline Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) total scores. (A-C) Correlations between the physician-assessed scores and those generated by LLM runs 1-3, respectively. (D-F) Corresponding Bland-Altman plots.



Predictive Performance of LLMs for 6-Month WOMAC Outcomes

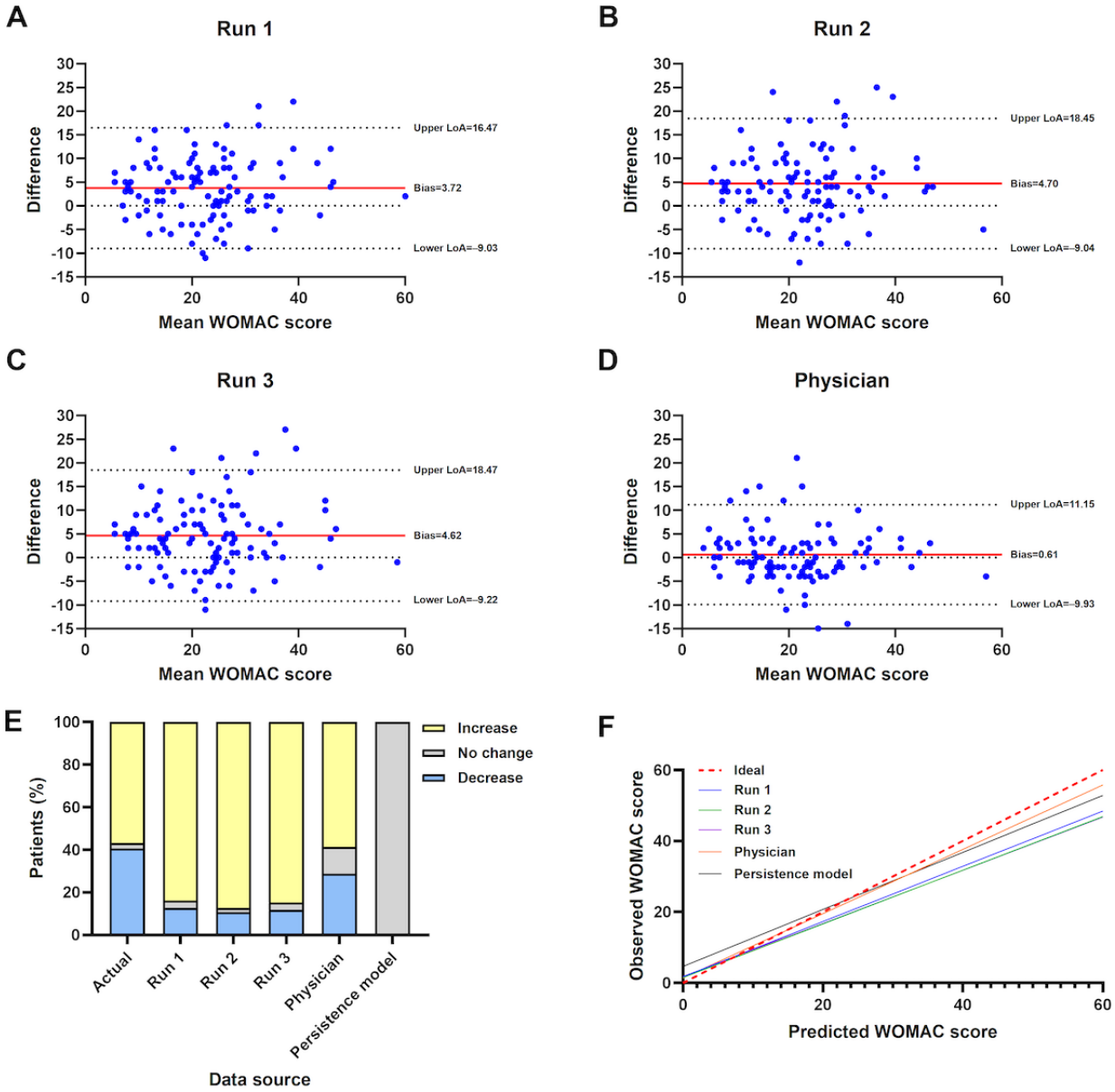
The mean actual WOMAC total score at 6 months was 20.89 points. The mean predicted scores from the 3 LLM runs ranged from 24.61 to 25.59 points, with an MAE of 5.92 to 6.58 points, an RMSE of 7.47 to 8.42 points, and a mean signed error of +3.72 to +4.70 points (Table 4). The LLM predictions showed a strong correlation with the observed values (*r*=0.779-0.814), and Bland-Altman analysis revealed a positive prediction bias

XSL•FO
RenderX

in all 3 runs (Figure 3A-D). During actual follow-up, the WOMAC total score fluctuated in both directions, whereas the LLM clearly tended to predict higher scores (Figure 3E). The calibration slope of the LLM ranged from 0.748 to 0.778 (Figure 3F).

The MAE of the physician's predictions was 3.86 points, which was lower than the 5.43 points for the persistence model (P=.001) and the 5.92 to 6.58 points for the LLM (P<.001; Table S3 in Multimedia Appendix 2). The mean signed error for the physician was +0.61 points, which was lower than the +3.72 to +4.70 points for the LLM (Table 4). Bland-Altman analysis showed minimal systematic bias between the physician's predictions and observed values (Figure 3D). The calibration slope of physician predictions was 0.906, which was closer to the ideal value of 1 than the results from the 3 LLM runs (Figure 3F).

Figure 3. Performance and calibration for predicting the 6-month Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) total score. (A-D) Bland-Altman plots comparing the predictions from large language model (LLM) runs 1-3 and the physician with the observed 6-month WOMAC total score, respectively. (E) Distribution of the observed and predicted directions of WOMAC change among the three LLM runs, the physician, and the persistence model. (F) Calibration curves for the three LLM runs, the physician prediction, and the persistence model. LoA: limits of agreement.



XSL•FO
RenderX

Table 4. Overall performance for predicting the actual 6-month Western Ontario and McMaster Universities Osteoarthritis Index total score.

Methods	Predicted mean (SD)	MAE[a] (BCa[b] 95% CI)	RMSE[c]	Signed error, mean (SD)	Pearson *r*
LLM[d] run 1	24.61 (10.87)	5.92 (5.08-6.77)	7.47	+3.72 (6.50)	0.814
LLM run 2	25.59 (10.70)	6.58 (5.66-7.55)	8.42	+4.70 (7.01)	0.779
LLM run 3	25.51 (10.83)	6.41 (5.43-7.42)	8.41	+4.62 (7.06)	0.779
Physician	21.50 (9.86)	3.86 (3.22-4.53)	5.39	+0.61 (5.38)	0.860
Persistence model	20.22 (10.31)	5.43 (4.77-6.14)	6.61	−0.68 (6.60)	0.797

[a]MAE: mean absolute error.

[b]BCa: bias-corrected and accelerated.

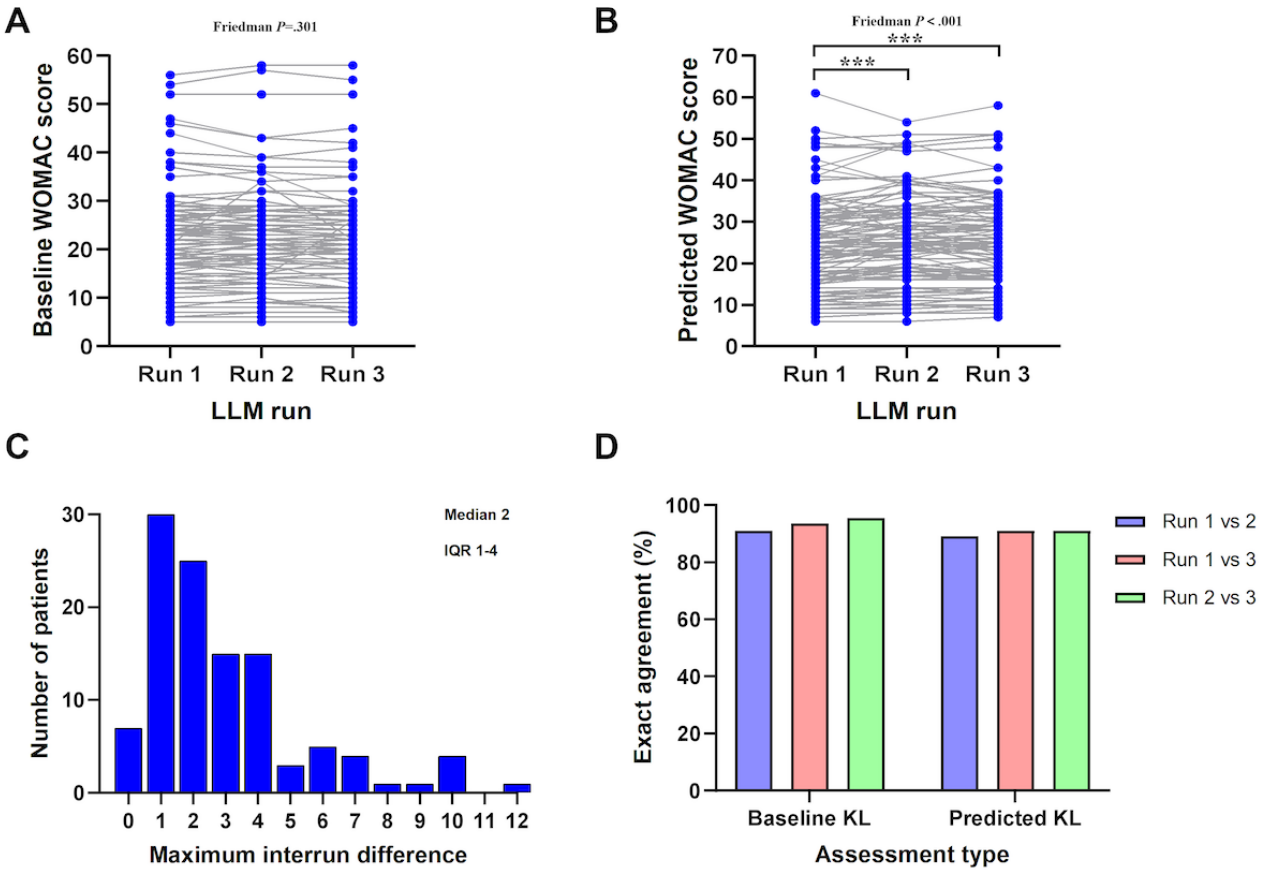
[c]RMSE: root mean squared error.

[d]LLM: large language model.

Reproducibility Across Multiple Runs of LLMs

The LLM achieved ICCs of 0.982 and 0.965 for the baseline WOMAC total score assessment and the 6-month WOMAC total score prediction, respectively; however, the proportions of cases demonstrating complete agreement across the 3 runs were only 31.5% (35/111) and 6.3% (7/111; Table S4 in Multimedia Appendix 2). There was no significant difference in the distribution of baseline WOMAC total scores across the 3 LLM runs (*P*=.30; Figure 4A), whereas the overall difference in the 6-month predicted scores was significant (*P*<.001; Figure 4B), with the predicted scores from the first run being lower than those from the second and third runs (Table S4 in Multimedia Appendix 2). The median maximum difference in the 6-month WOMAC predicted scores across the 3 LLM runs was 2 (IQR 1-4; Figure 4C) points. The 3 LLM runs were completely consistent in KOA diagnosis and determination of the affected side. The rates of complete agreement for the baseline and 6-month predicted KL grades were 90.1% (100/111) and 85.6% (95/111), respectively, and the rates of exact agreement between any 2 runs were also high (Figure 4D).

Figure 4. Reproducibility across 3 independent large language model (LLM) runs. (A) Baseline Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) total scores across the 3 runs. (B) Predicted 6-month WOMAC total scores across the 3 runs. (C) Distribution of the maximum interrun difference in the predicted 6-month WOMAC score. (D) Pairwise exact agreement for baseline and predicted 6-month Kellgren-Lawrence (KL) grades. ***P<.001.

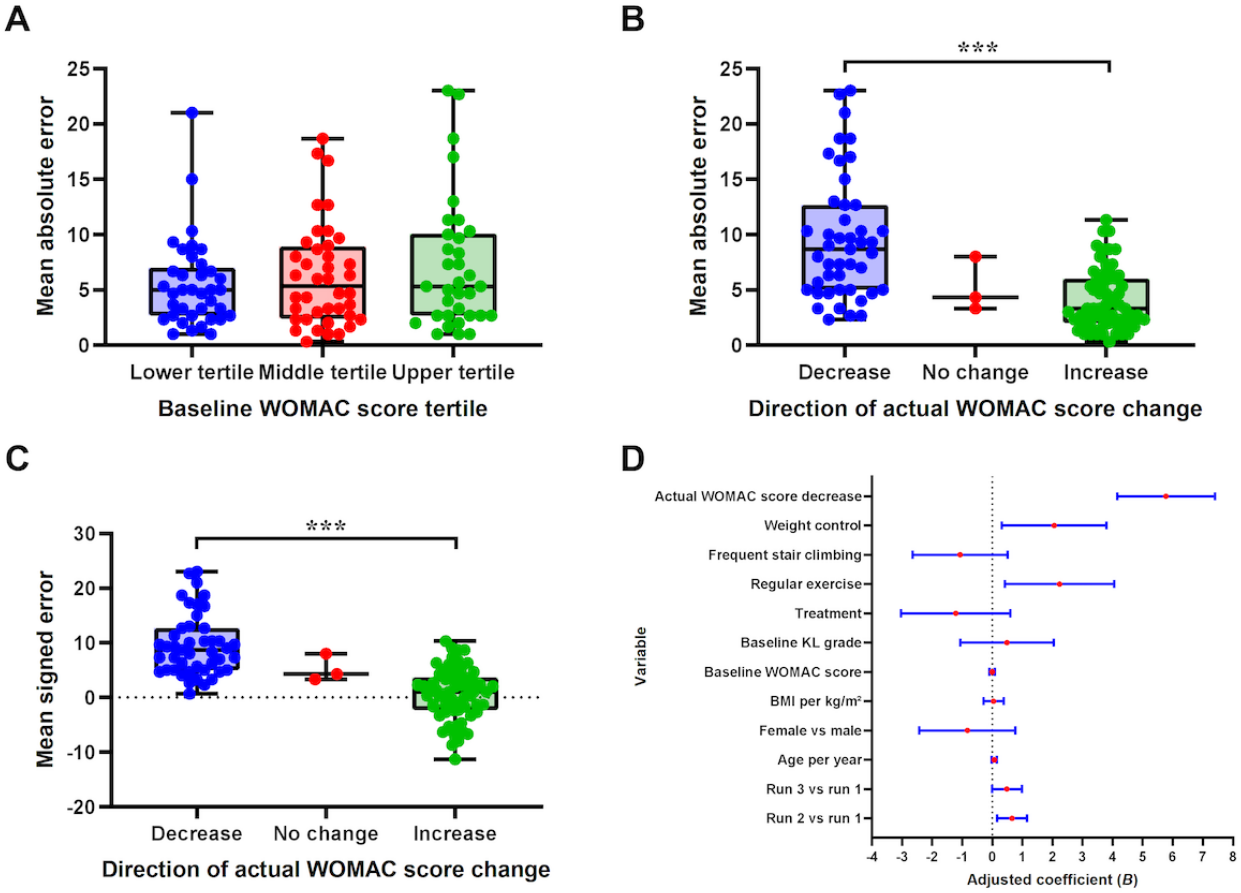


XSL•FO
RenderX

Model Error Patterns in LLMs and Related Factors

Between-group differences in patient-level MAE were not statistically significant among the low, medium, and high WOMAC baseline tertiles (*P*=.64; Figure 5A). Both the MAE and signed error differed significantly across groups based on the direction of change in actual WOMAC total scores (*P*<.001 in all cases; Table S5 in Multimedia Appendix 2). The MAE and signed error were higher in the group with decreasing scores than in the group with increasing scores (Figure 5B and Figure 5C). An exploratory mixed-effects model showed that regular exercise, active weight control, and a decrease in the actual WOMAC total score were all associated with larger absolute prediction errors, with the regression coefficient for a decrease in the actual score being the largest (β=5.774, 95% CI 4.151-7.396; *P*<.001; Figure 5D).

Figure 5. Patterns and associated factors of large language model (LLM) prediction error. (A) Mean absolute error according to baseline Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) tertiles. (B) Mean absolute error according to the actual direction of WOMAC change. (C) Mean signed error according to the actual direction of WOMAC change. (D) Adjusted coefficients and 95% CIs from the linear mixed-effects model. KL, Kellgren-Lawrence. ***P<.001.



LLM Predictions for 6-Month KL Grading

Across the 3 LLM runs, exact agreement with the observed 6-month KL grade ranged from 66.7% (74/111) to 68.5% (76/111) compared with 86.5% (96/111) for the physician comparator and 85.6% (95/111) for the persistence model (Table 5). Across the 3 LLM runs, 98.2% (109/111) of the predictions differed by no more than 1 grade from the actual grading; misclassifications primarily involved underestimating KL grade 3 as grade 2, while misclassifications by the physician were relatively rare (Figure 6A-D). Both the physician and the persistence model had lower MAEs than the 3 LLM runs (Table S6 in Multimedia Appendix 2). Only 13.5% (15/111) of the patients showed progression in report-derived KL grade. Across methods, sensitivity ranged from 0% (0/15) to 20% (3/15), with no statistically significant between-method difference (*P*=.26). For LLM run 2, sensitivity was 20% (3/15; 95% CI 4.3%-48.1%). Specificity ranged from 80.2% (77/96) to 83.3% (80/96) for the LLM, compared with 99% (95/96) for the physician and 100% (96/96) for the persistence model (overall *P*<.001). Other classification metrics are shown in Figure 6E to 6H.

XSL•FO
RenderX

Table 5. Performance for predicting the observed 6-month Kellgren-Lawrence grade (N=111).[a]

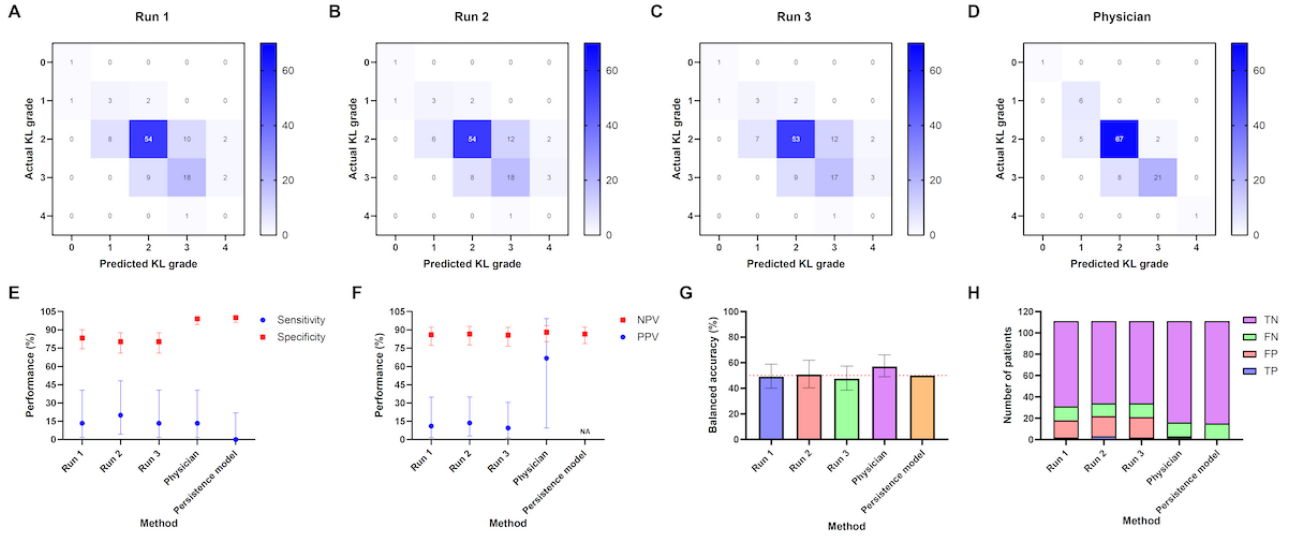
Method	Exact agreement, n (%; 95% CI)	Difference of ≤1 grade, n (%)	Linear-weighted κ	Grade MAE[b]
LLM[c] run 1	76 (68.5; 59.0-77.0)	109 (98.2)	0.482	0.333
LLM run 2	76 (68.5; 59.0-77.0)	109 (98.2)	0.486	0.333
LLM run 3	74 (66.7; 57.1-75.3)	109 (98.2)	0.462	0.351
Physician	96 (86.5; 78.7-92.2)	111 (100)	0.760	0.135
Persistence model	95 (85.6; 77.6-91.5)	111 (100)	0.736	0.144

[a]Exact binomial 95% CIs are reported for exact agreement.

[b]MAE: mean absolute error.

[c]LLM: large language model.

Figure 6. Prediction of the 6-month Kellgren-Lawrence (KL) grade and radiographic progression. (A-D) Confusion matrices for large language model (LLM) runs 1-3 and the physician prediction, respectively. (E) Sensitivity and specificity for identifying radiographic progression. (F) Positive and negative predictive values. (G) Balanced accuracy. (H) Numbers of true-positive, false-positive, false-negative, and true-negative classifications. Error bars in E and F represent exact binomial 95% CIs; those in G represent percentile 95% CIs from 10,000 bootstrap resamples stratified by progression status. The persistence model's balanced accuracy is 50% by construction. NPV: negative predictive value; PPV: positive predictive value.



Discussion

Principal Findings

This study evaluated DeepSeek's baseline assessment, 6-month outcome prediction, and operational stability in a prospective KOA cohort. The LLM showed high diagnostic agreement with a physician and performed well in WOMAC scoring and KL grading; however, its cross-sectional performance did not translate into reliable longitudinal predictions. This diagnostic agreement was assessed only in patients with established KOA and does not establish general diagnostic accuracy. The 6-month WOMAC prediction error and positive bias were higher than those of the physician comparator, and the model did not outperform the persistence model; its performance in KL grading and prediction of radiographic progression was also limited. Although interrun consistency was relatively high, absolute outputs still exhibited variability at the patient level.

Comparison With Prior Work

The evaluated model's strong baseline performance may reflect the structured information provided by standardized interviews and descriptive radiology reports. Previous studies have shown that LLMs can identify hip osteoarthritis or KOA using structured questionnaires, generate diagnoses consistent with those of specialist physicians using anonymized medical records, and extract imaging features and generate structured scores from free-text knee magnetic resonance imaging reports [9-11]. However, high correlations and ICCs alone do not establish clinically acceptable scoring accuracy. Absolute errors and Bland-Altman limits of agreement should also inform interpretation.

The evaluated LLM did not demonstrate added prognostic value over the persistence model, which assumes unchanged baseline status. Strong correlations with observed outcomes did not ensure accurate individual predictions, as reflected by the positive bias and poor calibration. Although DeepSeek demonstrated strong clinical decision-making capabilities in clinical case benchmarks [12], longitudinal prediction requires estimating state changes resulting from the combined effects of treatment, behavior, and disease fluctuations—dynamic information that is difficult to fully capture from a single baseline record. Existing research on medical LLMs has primarily focused on knowledge-based question-answering and

XSL•FO
RenderX

diagnostic tasks, with limited evaluation of real-world longitudinal outcomes and calibration [13,14].

The evaluated DeepSeek configuration tended to predict symptom worsening, particularly when patients actually improved, suggesting limited capture of bidirectional short-term symptom changes. Pain and function in KOA can follow different trajectories and are influenced by exercise, weight management, and other interventions [21,22]. Regular exercise and weight control are associated with larger prediction errors, suggesting that an LLM based solely on a single baseline data point may struggle to capture subsequent symptom trajectories related to behavior changes.

Previous studies have also shown that LLMs' performance in fine-grained KL grading remains limited [23,24]. In this study, the model was not provided with raw x-ray images but rather with descriptive imaging reports from which diagnostic conclusions had been removed. Baseline grading therefore involved extracting reported radiographic features, while longitudinal prediction targeted changes in report-derived KL grades. However, these findings do not establish performance in direct image-based progression prediction. Given the small number of progression events and class imbalance, overall grade agreement should be interpreted alongside progression sensitivity, specificity, and balanced accuracy.

Although ICCs were high across the 3 independent runs, exact agreement was low for both baseline WOMAC assessment and 6-month WOMAC prediction, with substantial interrun variability in some patients. The predicted values from the first run were lower than those from the subsequent 2 runs, suggesting that output fluctuations may include systematic run-level biases. Previous studies have also found that responses to the same questions are not entirely consistent across different time points or repeated runs [15,25]. High ICCs can coexist with run-dependent shifts in absolute scores even when patients remain similarly ranked. Such shifts could alter the estimated magnitude or direction of WOMAC change, limiting confidence in single-run predictions for individual patients.

The evaluated model may support clinical information extraction and preliminary baseline scoring under clinician supervision. Individualized prognosis requires anticipating future clinical changes, and its reliability cannot be inferred from strong baseline assessment performance. Previous studies have also shown that LLMs can produce KOA self-management materials and provide recommendations that are generally consistent with those of rehabilitation specialists [26,27]. In the future, unstructured information extracted by LLMs could be combined with raw imaging data, longitudinal symptom data, treatment adjustments, and behavioral trajectories to construct dynamic, multimodal, disease-specific models supervised by physicians.

Limitations

This study has several limitations. This single-center study had only 15 progression events, limiting precision; progression-related findings should therefore be considered exploratory. Furthermore, the 6-month follow-up period and the KL grading system may not be sufficient to capture subtle structural changes. Different evaluators assigned baseline and follow-up report-derived KL grades without prior calibration or formal agreement assessment, potentially affecting progression classification and related prediction metrics. This study evaluated one DeepSeek configuration and one physician predictor, limiting generalizability to other models, physicians, populations, and clinical settings. Attrition may have introduced selection bias. The nonsignificant MCAR test did not establish that data were MCAR, and the IPW sensitivity analysis depended on assumptions about missingness and model specification. The relevant exploratory findings require validation in larger samples and external cohorts.

Conclusions

The evaluated DeepSeek configuration may support structured baseline KOA assessment, but this does not establish its suitability for individualized prognosis. Prognostic applications should demonstrate added value beyond a persistence model, adequate calibration, and stable patient-level predictions in independent cohorts. Further evaluation across clinical settings, model versions, and physician comparators is needed to define the clinical role of the evaluated DeepSeek configuration.

Acknowledgments

The authors thank all study participants for their valuable contribution and the clinical and radiology staff of Fuxing Hospital, Capital Medical University, for their assistance with patient recruitment, follow-up assessments, and radiographic examinations.

Funding

This work was supported by the Capital's Funds for Health Improvement and Research (grant 2022-3-7024).

Data Availability

The data supporting the findings of this study are not publicly available to protect participant privacy. Deidentified data and the code for the supplementary analyses are available from the corresponding author on reasonable request.

Generative AI Use

No generative AI tools were used to draft or edit the manuscript, create figures, or develop analysis code. DeepSeek was used only as the model evaluated in this study as described in the Methods section.

XSL•FO
RenderX

Authors' Contributions

Conceptualization: ML
Data curation: ZL, YZ
Formal analysis: ZL
Investigation: ZL, YZ
Methodology: ML
Supervision: ML
Validation: YZ
Visualization: ZL
Writing—original draft: ZL
Writing—review and editing: YZ, ML

Conflicts of Interest

None declared.

Multimedia Appendix 1

Standardized interview record template and complete large language model prompt sequence.
[DOCX File , 22 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Supplementary analyses and results.
[DOCX File , 25 KB-Multimedia Appendix 2]

References

1. Sharma L. Osteoarthritis of the knee. N Engl J Med. Jan 07, 2021;384(1):51-59. [doi: 10.1056/NEJMcp1903768] [Medline: 33406330]
2. Mahmoudian A, Lohmander LS, Mobasheri A, Englund M, Luyten FP. Early-stage symptomatic osteoarthritis of the knee - time for action. Nat Rev Rheumatol. Oct 2021;17(10):621-632. [doi: 10.1038/s41584-021-00673-4] [Medline: 34465902]
3. Guan B, Liu F, Mizaian AH, Demehri S, Samsonov A, Guermazi A, et al. Deep learning approach to predict pain progression in knee osteoarthritis. Skeletal Radiol. Feb 2022;51(2):363-373. [FREE Full text] [doi: 10.1007/s00256-021-03773-0] [Medline: 33835240]
4. Joseph GB, McCulloch CE, Nevitt MC, Link TM, Sohn JH. Machine learning to predict incident radiographic knee osteoarthritis over 8 years using combined MR imaging features, demographics, and clinical factors: data from the Osteoarthritis Initiative. Osteoarthritis Cartilage. Feb 2022;30(2):270-279. [FREE Full text] [doi: 10.1016/j.joca.2021.11.007] [Medline: 34800631]
5. Schiratti JB, Dubois R, Herent P, Cahané D, Dachary J, Clozel T, et al. A deep learning method for predicting knee osteoarthritis radiographic progression from MRI. Arthritis Res Ther. Oct 18, 2021;23(1):262. [FREE Full text] [doi: 10.1186/s13075-021-02634-4] [Medline: 34663440]
6. Li X, Li C, Zhang P. Predictive models of radiographic progression and pain progression in patients with knee osteoarthritis: data from the FNIH OA Biomarkers Consortium project. Arthritis Res Ther. May 30, 2024;26(1):112. [FREE Full text] [doi: 10.1186/s13075-024-03346-1] [Medline: 38816759]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Pagano S, Strumolo L, Michalk K, Schiegl J, Pulido LC, Reinhard J, et al. Evaluating ChatGPT, Gemini and other large language models (LLMs) in orthopaedic diagnostics: a prospective clinical study. Comput Struct Biotechnol J. Dec 26, 2024;28:9-15. [FREE Full text] [doi: 10.1016/j.csbj.2024.12.013] [Medline: 39850460]
10. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol. Nov 28, 2023;24(1):61. [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
11. Xie Y, Hu Z, Tao H, Hu Y, Liang H, Lu X, et al. Large language models for efficient whole-organ MRI score-based reports and categorization in knee osteoarthritis. Insights Imaging. May 14, 2025;16(1):100. [doi: 10.1186/s13244-025-01976-w] [Medline: 40366500]
12. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. Aug 2025;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]

XSL•FO
RenderX

13. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
14. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. Sep 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
15. Zhu L, Mou W, Hong C, Yang T, Lai Y, Qi C, et al. The evaluation of generative AI should include repetition to assess stability. JMIR Mhealth Uhealth. May 06, 2024;12:e57978. [FREE Full text] [doi: 10.2196/57978] [Medline: 38688841]
16. Joint Surgery Group, Chinese Orthopaedic Association, Chinese Medical Association. 骨关节炎诊疗指南（2018年版）[Guideline for the Diagnosis and Treatment of Osteoarthritis (2018 Edition)] [Article in Chinese]. Chin J Orthop. 2018;38(12):705-715. [FREE Full text] [doi: 10.3760/cma.j.issn.0253-2352.2018.12.001]
17. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
18. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol. Dec 1988;15(12):1833-1840. [Medline: 3068365]
19. Xie F, Li SC, Goeree R, Tarride JE, O'Reilly D, Lo NN, et al. Validation of Chinese Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) in patients scheduled for total knee replacement. Qual Life Res. May 2008;17(4):595-601. [doi: 10.1007/s11136-008-9340-7] [Medline: 18415706]
20. DeepSeek-V4 preview: entering the era of affordable million-token context. DeepSeek. Apr 24, 2026. URL: https://deepseek.com/en/news/v4-preview/ [accessed 2026-10-28]
21. Bastick AN, Wesseling J, Damen J, Verkleij SP, Emans PJ, Bindels PJ, et al. Defining knee pain trajectories in early symptomatic knee osteoarthritis in primary care: 5-year results from a nationwide prospective cohort study (CHECK). Br J Gen Pract. Jan 2016;66(642):e32-e39. [FREE Full text] [doi: 10.3399/bjgp15X688129] [Medline: 26639946]
22. Lee AC, Harvey WF, Han X, Price LL, Driban JB, Bannuru RR, et al. Pain and functional trajectories in symptomatic knee osteoarthritis over up to 12 weeks of exercise exposure. Osteoarthritis Cartilage. Apr 2018;26(4):501-512. [FREE Full text] [doi: 10.1016/j.joca.2018.01.016] [Medline: 29391277]
23. Zhu J, Jiang Y, Chen D, Lu Y, Huang Y, Lin Y, et al. High identification and positive-negative discrimination but limited detailed grading accuracy of ChatGPT-4o in knee osteoarthritis radiographs. Knee Surg Sports Traumatol Arthrosc. May 2025;33(5):1911-1919. [doi: 10.1002/ksa.12639] [Medline: 40053915]
24. Temel MH, Erden Y, Bağcıer F. Evaluating artificial intelligence performance in medical image analysis: sensitivity, specificity, accuracy, and precision of ChatGPT-4o on Kellgren-Lawrence grading of knee X-ray radiographs. Knee. Aug 2025;55:79-84. [doi: 10.1016/j.knee.2025.04.008] [Medline: 40273525]
25. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res. Sep 15, 2023;25:e47621. [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]
26. Du K, Li A, Zuo QH, Zhang CY, Guo R, Chen P, et al. Comparing artificial intelligence-generated and clinician-created personalized self-management guidance for patients with knee osteoarthritis: blinded observational study. J Med Internet Res. May 07, 2025;27:e67830. [FREE Full text] [doi: 10.2196/67830] [Medline: 40332991]
27. Gürses Ö, Özüdoğru A, Tuncay F, Kararti C. The role of artificial intelligence large language models in personalized rehabilitation programs for knee osteoarthritis: an observational study. J Med Syst. Jun 03, 2025;49(1):73. [doi: 10.1007/s10916-025-02207-x] [Medline: 40459660]

Abbreviations

ICC: intraclass correlation coefficient
IPW: inverse probability weighting
KL: Kellgren-Lawrence
KOA: knee osteoarthritis
LLM: large language model
MAE: mean absolute error
MCAR: missing completely at random
RMSE: root mean squared error
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model version
WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

XSL•FO
RenderX

Edited by I Steenstra; submitted 15.Jul.2026; peer-reviewed by I Kalappurackal Mansoor, S Kivity; comments to author 13.Oct.2026; revised version received 22.Oct.2026; accepted 22.Oct.2026; published 12.Nov.2026

Please cite as:
Li Z, Zhang Y, Li M
Performance of a Large Language Model for Predicting Clinical and Radiographic Outcomes in Knee Osteoarthritis: Single-Center Prospective Observational Study
JMIR Med Inform 2026;14:e107092
URL: https://medinform.jmir.org/2026/1/e107092
doi: 10.2196/107092
PMID:

©Zhiqiang Li, Yuhan Zhang, Mingwei Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX