

RESEARCH

Open Access



Evaluation of ChatGPT-5 responses to patient-centered questions on stromal vascular fraction for knee osteoarthritis: fair to good quality and content

Numan Mercan^{1*} , Ebubekir Eravsar² , Musa Ergin³ , Tevfik Çatal¹ , Sadettin Çiftçi² and Mustafa Akkaya^{4,5*}

Abstract

Background The global burden of symptomatic knee osteoarthritis (KOA) continues to grow, driving clinical interest in biological treatment options. Stromal vascular fraction (SVF), derived from adipose tissue, has gained attention as a potential therapy for KOA. As patients increasingly utilize large language models (LLMs) like ChatGPT for health information, this study aimed to evaluate the patient-information quality and readability of AI-generated responses to a curated set of patient-centered questions on SVF therapy for KOA.

Methods Thirty patient-centered questions were developed through literature review by two experts and then posed to ChatGPT-5 after asking it to answer from an orthopaedic specialist perspective. Responses from ChatGPT-5 were evaluated by four orthopaedic specialists using three quality instruments: DISCERN, NLAT-AI (assessing five domains: Accuracy, Safety, Appropriateness, Actionability, and Effectiveness), and the Mika et al. scoring system. Readability was assessed using five standard metrics.

Results Mean scores were as follows: DISCERN 44.61 ± 4.94 (Fair); for NLAT-AI domains, Accuracy 3.92 ± 0.50 (Good), Safety 3.23 ± 0.78 (Fair), Appropriateness 4.14 ± 0.32 (Good), Actionability 3.20 ± 0.67 (Fair), and Effectiveness 4.46 ± 0.35 (Excellent); NLAT-AI (Total/ sum of five domains) 18.95 ± 1.93 (Good); Mika et al. 2.45 ± 0.49 (Fair). Readability metrics indicated an 11th to 12th grade reading level.

Conclusion ChatGPT-5 provides fair-to-good quality responses to patient-centered questions about SVF in KOA. The answers are generally effective and clinically appropriate, with good accuracy, and often require only limited additional clarification. However, safety cautions and practical guidance are less consistently covered, and the reading level is relatively high. Further research is needed to clarify its role as an adjunct tool for patient education in this setting.

Keywords Artificial intelligence, ChatGPT, Knee osteoarthritis, Patient education, Regenerative medicine, Stromal vascular fraction

*Correspondence:

Numan Mercan
numanmercan@gmail.com
Mustafa Akkaya
mustafa@drakkaya.com

Full list of author information is available at the end of the article



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

The global burden and prevalence of symptomatic knee osteoarthritis (KOA) are projected to increase further by mid-century, placing significant pressure on both patients and healthcare systems [1, 2]. Current nonsurgical treatments, including patient education, weight management, exercise, and intra-articular injections, provide symptomatic relief but are insufficient to permanently alter disease progression, and their effects are often of limited duration [3, 4]. These limitations have led to clinical and research interest in biological therapies that can modulate inflammation, improve joint homeostasis, and promote structural healing [5].

Among orthobiologic options, adipose tissue-derived therapies, particularly uncultured stromal vascular fraction (SVF), stand out due to their tissue abundance and low morbidity [6, 7]. SVF is a heterogeneous cell mixture composed of adipose-derived mesenchymal stem cells, endothelial progenitor cells, pericytes, fibroblasts, and various immune cells [8]. These cells are thought to suppress synovial inflammation, support chondrocyte activity, and promote extracellular matrix remodeling through paracrine, immunomodulatory, and trophic mechanisms [9, 10]. Some studies have reported that intra-articular SVF injections significantly improve pain and function in KOA at 12–24 months follow-up [11].

Today, patients increasingly turn to digital resources, such as the internet and large language models (LLMs),

for health information [12, 13]. Digital resources are used for patient education in many areas, including musculoskeletal disorders, but their content quality and readability vary [14, 15]. ChatGPT, one of these digital resources, has been available since November 2022. It facilitates access to information by interacting with users through its text comprehension and production capabilities. It has rapidly become widespread across a wide range of applications, including healthcare [16, 17]. Its most recent model, ChatGPT-5, was released in 2025. As interest in regenerative and biologic approaches in orthopaedics continues to grow, patients increasingly seek information about options such as SVF before or alongside clinical care, and LLMs are being used more often as part of that search. For this reason, it is important to establish how usable and safe the information provided by these models is for patients, and to ensure that this information is also reviewed by clinicians.

Therefore, this study aimed to evaluate the patient-information quality and readability of AI-generated responses to a curated set of patient-centered questions on SVF therapy for KOA, when ChatGPT-5 was prompted to respond from an orthopaedic specialist perspective, and to describe the strengths and weak points of these responses. Thirty patient-centered questions were posed to ChatGPT-5, and its answers were evaluated using the DISCERN instrument for patient-information quality, the NLAT-AI framework for clinical suitability across accuracy, safety, appropriateness, actionability, and effectiveness, and the Mika et al. grading system to indicate whether the responses required additional explanation, together with standard readability indices. To our knowledge, this study is the first to perform a quality and readability analysis of the responses given by an LLM to questions regarding the use of SVF in KOA.

Materials and methods

Question generation and strategy for asking the artificial intelligence model

Ethical approval for this study was granted by the Local Ethics Committee of Selçuk University Faculty of Medicine (decision number: 2026/153). As the study involved no human participants, human data, or human tissue, informed consent was not applicable. The study was conducted in accordance with the principles of the Declaration of Helsinki.

A flowchart was constructed to illustrate the question selection process (Fig. 1). A search was conducted on PubMed for entries containing the terms 'stromal vascular fraction' and 'knee osteoarthritis' in the 'Title/Abstract' field. This initial search yielded 89 studies (as of August 20, 2025). These studies were then screened and refined to 67 that specifically addressed the use of SVF in KOA. From these 67 studies, a total of 1,088 questions

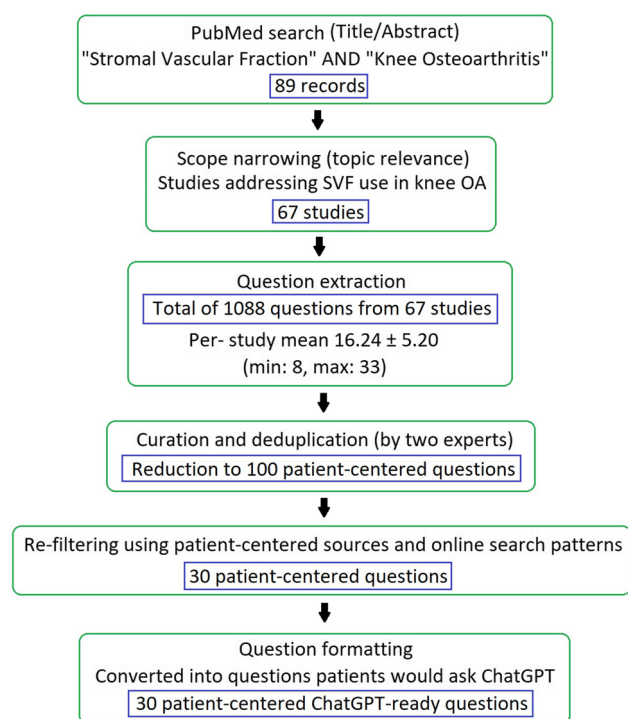


Fig. 1 Development of the 30 patient-centered questions used in this study

were extracted (N.M.). The mean number of questions per study was 16.24 ± 5.20 (range: 8 to 33). These 1,088 questions were consolidated into a refined list of 100 patient-centered questions through a curation process performed by two experts (M.A. (with a musculoskeletal system and regenerative medicine PhD) and N.M.). During this refinement, study-derived statements were translated into patient-style questions; duplicate or near-duplicate items were removed, research-style statements were excluded, and overlapping questions were merged into a single patient-friendly wording, yielding a refined list of 100 patient-centered questions. This list was then further refined to a core set of 30 questions by M.A. by filtering them against the themes and findings of key studies [18–20] and prioritizing common patient information needs and frequently searched topics. The final 30 patient-centered questions were subsequently phrased to simulate how a patient might pose them to an AI chatbot like ChatGPT-5 and were written in a short, direct FAQ style covering benefit, suitability, procedure, safety, and practical issues such as cost and approval status. The final wording was reviewed by M.A., and patients were not

involved in selecting or wording the final questions. The questions were categorized by content as follows: 2 questions pertained to the description and mechanism of the treatment, 10 focused on its efficacy, 6 addressed safety and eligibility concerns, 8 inquired about the procedure, and 4 questions related to evidence and funding (Table 1).

All questions were posed to ChatGPT-5 on the same day (August 22, 2025) to ensure that all responses were generated under the same model conditions and could be scored consistently. To reflect a patient seeking an orthopaedic specialist explanation and to keep the context consistent across all questions, we used the following prompt: “You are an experienced, patient-focused orthopaedic specialist with a PhD in regenerative medicine. You answer questions from patients about SVF (Stromal Vascular Fraction) therapy for knee osteoarthritis.” This wording was used to make the responder’s role explicit, not to search for an ideal prompt. Questions were asked from a newly created Gmail account after registering a new ChatGPT account. Each question was asked in an isolated session using the same system prompt,

Table 1 Questions posed to ChatGPT-5 and their categories

No	Question	Category
1	What is SVF therapy for knee osteoarthritis?	Description/Mechanism
2	Is SVF effective for knee osteoarthritis?	Efficacy
3	How does SVF work in knee osteoarthritis?	Description/Mechanism
4	Does SVF reduce pain in knee osteoarthritis?	Efficacy
5	Does SVF improve range of motion in knee osteoarthritis?	Efficacy
6	Does SVF reduce knee stiffness in knee osteoarthritis?	Efficacy
7	Does SVF repair cartilage in knee osteoarthritis?	Efficacy
8	Does SVF slow the progression of knee osteoarthritis?	Efficacy
9	Can SVF delay the need for a knee replacement?	Efficacy
10	At what disease stage is SVF most appropriate for knee osteoarthritis?	Safety/Eligibility
11	Who should not receive SVF for knee osteoarthritis?	Safety/Eligibility
12	Is SVF safe for older adults with knee osteoarthritis?	Safety/Eligibility
13	Is SVF safe for people with knee osteoarthritis who have other medical conditions?	Safety/Eligibility
14	Can SVF be given to both knees on the same day?	Procedure
15	How is SVF obtained for use in knee osteoarthritis?	Procedure
16	How is SVF administered for knee osteoarthritis?	Procedure
17	Is anesthesia required for an SVF knee injection?	Procedure
18	Should an SVF knee injection be performed under ultrasound guidance?	Procedure
19	Is any special preparation required before SVF for knee osteoarthritis?	Procedure
20	How long does an SVF knee injection take?	Procedure
21	How many SVF treatments are typically needed for knee osteoarthritis?	Procedure
22	When are benefits typically expected after SVF for knee osteoarthritis?	Efficacy
23	How long do the benefits of SVF last in knee osteoarthritis?	Efficacy
24	What are the most common side effects of SVF for knee osteoarthritis?	Safety/Eligibility
25	Are there any risks or complications with SVF for knee osteoarthritis?	Safety/Eligibility
26	Is SVF FDA-approved for knee osteoarthritis?	Evidence/Funding
27	How much does SVF cost for knee osteoarthritis?	Evidence/Funding
28	Is SVF for knee osteoarthritis covered by health insurance?	Evidence/Funding
29	What is the evidence level supporting SVF for knee osteoarthritis?	Evidence/Funding
30	How does SVF compare in effectiveness with other knee osteoarthritis injections?	Efficacy

preventing memory transfer between sessions. References were not requested as part of the questions or the opening wording. Questions and ChatGPT responses are provided in Supplementary Material 1.

Quality assessment

Three different quality scoring systems were used in the study, and it was planned to evaluate each response within a broader framework using these quality scoring systems with different contents (Fig. 2). The quality assessment of the responses was conducted by four observers (E.E., M.E., T.Ç., and S.Ç.), all orthopaedic specialists who use regenerative orthobiologics in their current clinical practice. All four observers rated all 30 responses using DISCERN, NLAT-AI, and the Mika et al. score.

The DISCERN scale was used to evaluate the quality of the patient information texts. DISCERN consists of 16 items, each scored on a scale of 1 to 5. Each text receives a total score ranging from 16 to 80, with higher scores indicating higher quality. The first 8 items of DISCERN assess reliability, the next 7 items evaluate treatment-specific details, and the final item rates the overall quality [21].

The Natural Language Assessment Tool for AI (NLAT-AI) scoring system was used to assess the suitability of the artificial intelligence responses for clinical use, based on a multidimensional expert evaluation framework [22]. This system covers five domains: Accuracy, Safety, Appropriateness, Actionability, and Effectiveness. In this study, the Accuracy score reflects the clinical correctness of the answer text as judged by the orthopaedic observers, not the checking of any references that may appear in the response. Each domain is scored on a 5-point Likert scale (5 = Strongly Agree, 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly Disagree).

The Mika et al. scoring system, which provides a grading model indicating whether ChatGPT responses require additional explanation, was also used [23]. This system was initially developed for ChatGPT responses

related to hip arthroplasty and has since been applied in other orthopaedic studies [24, 25]. Scores in this system range from 1 to 4 (1 = Excellent response not requiring clarification; 2 = Satisfactory requiring minimal clarification; 3 = Satisfactory requiring moderate clarification; 4 = Unsatisfactory requiring substantial clarification) (Table 2). Unlike the DISCERN and NLAT-AI scoring systems, a lower score in this system indicates better response quality.

Readability assessment

The readability of ChatGPT responses was evaluated with five algorithms commonly and establishedly used in health communication: Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Coleman-Liau Index (CLI) (Fig. 2). The use of multiple tools was preferred in order to reduce biases specific to a single formula. FRE produces a score between 0 and 100, and a higher score indicates easier readability. GFI and FKGL provide a grade-level (year/grade) output based on sentence length/word structure, and for example, 12 indicates that the text is approximately at the 12th-grade level. CLI similarly provides a grade level but is based on character counts. SMOG produces a grade-level estimate frequently used for health messages, based on the density of multisyllabic words. As a general rule, lower grade-level scores indicate easier readability. Readability analyses were performed by N.M. on the website readabilityformulas.com (accessed September 02, 2025).

Statistical analysis

The scores from four raters for DISCERN (Total), NLAT-AI (its five domains and Total), and Mika et al. scores were averaged at the response level. Descriptive statistics for these scores are presented as mean $\pm$ standard deviation (SD) and range (minimum–maximum). Inter-rater reliability was assessed using a two-way random-effects, absolute agreement, average-measures intraclass

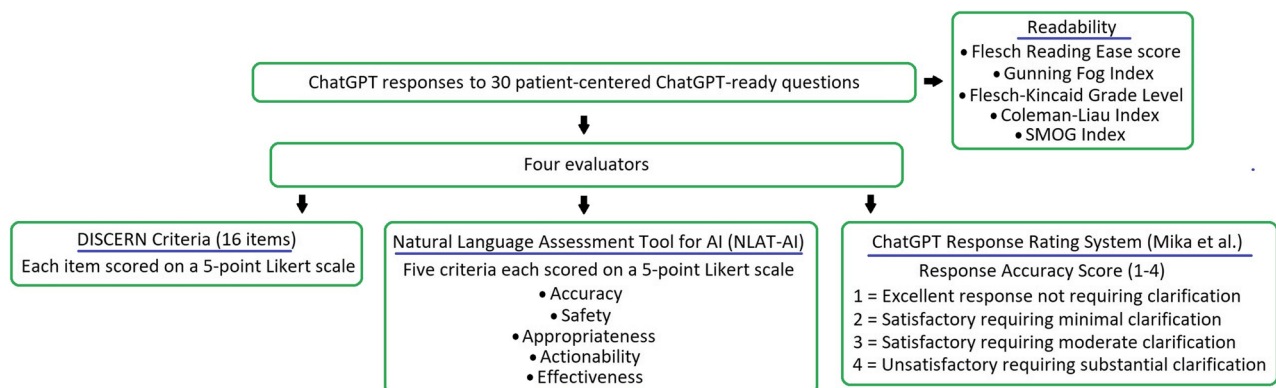


Fig. 2 Overview of the assessments performed for ChatGPT responses

Table 2 ChatGPT response rating system by Mika et al

Score	Description
1	Excellent response not requiring clarification
2	Satisfactory requiring minimal clarification
3	Satisfactory requiring moderate clarification
4	Unsatisfactory requiring substantial clarification

correlation coefficient, ICC(A,4), based on a two-way ANOVA model (Shrout & Fleiss / McGraw & Wong convention). For completeness, the single-measure ICC(A,1) was also calculated. The reliability was interpreted as follows: poor (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), and excellent (0.81–1.00) [26, 27]. The relationship between scales was examined using Spearman's rank correlation coefficient (ρ). To ensure consistent directionality across all scoring systems (where a high score indicates better quality), the Mika et al. scores were multiplied by -1 prior to analysis, as a lower raw score originally indicated better quality. The strength of the correlation ($|\rho|$) was classified with the following bands: negligible (< 0.20), weak (0.20–0.39), moderate (0.40–0.59), strong (0.60–0.79), and very strong (≥ 0.80). Group differences based on question categories (Efficacy, Procedure, Safety/Eligibility, Evidence/Funding, Description/Mechanism) for the scores (DISCERN, NLAT-AI (Total), Mika et al. score) and readability metrics were evaluated with the Kruskal–Wallis test. For significant results, pairwise comparisons were conducted using the Mann–Whitney U test with a Holm correction for multiple comparisons. In this section, data are reported as median (range) for ease of interpretation. Analyses were conducted in a Python 3.11 environment. All tests were two-tailed, and a p value of < 0.05 was considered statistically significant.

Results

The ChatGPT-5 responses generally exhibited a consistent patient information narrative structure. 90% (27/30) of responses began with an introductory paragraph that summarized the query content, and similarly, 90% (27/30) were structured with subheadings such as 'Here's how it works,' 'What to Expect,' 'Evidence & Considerations,' and 'Risks/Side Effects.' 70% (21/30) of responses contained a summary/conclusion section ending with phrases like 'In summary,' 'Bottom line,' or 'Overall.'

A common and comparable rating category framework was designed for the three scales. While established benchmarks from the literature were maintained for DISCERN, the score ranges for NLAT-AI (domains and total) and Mika et al. score were divided into five bands of equal width between the minimum and maximum scores, creating interpretation categories similar to those of DISCERN. This provided consistent and directly comparable thresholds (Excellent/Good/Fair/Poor/Very poor) across all scales (Table 3).

For the 30 responses rated by four observers, the mean total DISCERN score was 44.61 ± 4.94 (range: 34.25–51.25). The mean NLAT-AI domain scores were as follows: Accuracy 3.92 ± 0.50 (range: 3.00–5.00), Safety 3.23 ± 0.78 (range: 2.00–4.75), Appropriateness 4.14 ± 0.32 (range: 3.75–5.00), Actionability 3.20 ± 0.67 (range: 2.00–4.25), and Effectiveness 4.46 ± 0.35 (range: 4.00–5.00); the mean NLAT-AI (Total) score was 18.95 ± 1.93 (range: 15.75–23.25). In terms of the rating categories, the mean scores fell into the 'Fair' band for DISCERN; within NLAT-AI, Accuracy and Appropriateness were 'Good,' Effectiveness was 'Excellent,' whereas Safety and Actionability were 'Fair'; the NLAT-AI (Total) score was 'Good'; and the Mika et al. score was 'Fair' (Table 4).

Interobserver reliability, assessed by ICC(A,4), was found to be 0.821 for DISCERN, 0.801 for NLAT-AI (Total), and 0.746 for the Mika et al. score. On a domain basis, the values were: Safety 0.811, Accuracy 0.803, Appropriateness 0.684, Actionability 0.768, and Effectiveness 0.694. The accompanying single-measure ICC(A,1) values (for DISCERN 0.535, for NLAT-AI (Total) 0.502, for Mika et al. score 0.423) indicated moderate agreement at the individual rater level, supporting the use of mean scores for primary analyses (Table 4).

Regarding score distributions, DISCERN clustered most responses in the 'Fair' band. Within NLAT-AI, Appropriateness and especially Effectiveness were clustered in the upper bands (mostly Excellent), while Safety and Actionability were more broadly distributed across the middle bands. NLAT-AI (Total) was predominantly in the Good band, and the Mika et al. scores were concentrated in the 'Good/Fair' bands, being sparse at the extremes. This pattern summarizes the domain-specific quality variation and the comparability of the scales (Table 5).

Table 3 Rating categories and numerical thresholds for DISCERN, NLAT-AI, and Mika et al. scores

Rating Category	DISCERN	NLAT-AI Domains (1–5)	NLAT-AI Overall (5–25)	Mika (1–4)
Excellent	62.01–80.00	4.21–5.00	21.01–25.00	1.00–1.60
Good	50.01–62.00	3.41–4.20	17.01–21.00	1.61–2.20
Fair	38.01–50.00	2.61–3.40	13.01–17.00	2.21–2.80
Poor	27.01–38.00	1.81–2.60	9.01–13.00	2.81–3.40
Very poor	16.00–27.00	1.00–1.80	5.00–9.00	3.41–4.00

Table 4 Descriptive statistics for scales and inter-rater reliability

Scale	Mean $\pm$ SD	Range Min–Max	Rating Category	ICC(A,4)	Reliability Grade
DISCERN (Total)	44.61 $\pm$ 4.94	34.25–51.25	Fair	0.821	Excellent
NLAT-AI (Accuracy)	3.92 $\pm$ 0.50	3.00–5.00	Good	0.803	Good
NLAT-AI (Safety)	3.23 $\pm$ 0.78	2.00–4.75	Fair	0.811	Excellent
NLAT-AI (Appropriateness)	4.14 $\pm$ 0.32	3.75–5.00	Good	0.684	Good
NLAT-AI (Actionability)	3.20 $\pm$ 0.67	2.00–4.25	Fair	0.768	Good
NLAT-AI (Effectiveness)	4.46 $\pm$ 0.35	4.00–5.00	Excellent	0.694	Good
NLAT-AI (Total)	18.95 $\pm$ 1.93	15.75–23.25	Good	0.801	Good
Mika et al. score	2.45 $\pm$ 0.49	1.50–3.50	Fair	0.746	Good

Table 5 Distribution of rating categories according to scales ($n = 30$)

Scale	Excellent	Good	Fair	Poor	Very poor
DISCERN (Total)	0	6	20	4	0
NLAT-AI (Accuracy)	8	19	3	0	0
NLAT-AI (Safety)	4	9	8	9	0
NLAT-AI (Appropriateness)	10	20	0	0	0
NLAT-AI (Actionability)	3	10	9	8	0
NLAT-AI (Effectiveness)	22	8	0	0	0
NLAT-AI (Total)	5	18	7	0	0
Mika et al. score	1	9	13	6	1

Direction-aligned Spearman correlations (-Mika et al. score) indicated the level of agreement between the scales: The correlation between DISCERN and NLAT-AI (Total) was weak and statistically insignificant ($\rho = 0.301$, $p = 0.106$); the correlation between DISCERN and Mika et al. score was moderate and significant ($\rho = 0.466$, $p = 0.0095$); and the correlation between NLAT-AI (Total) and Mika et al. score was very strong and highly significant ($\rho = 0.874$, $p < 0.0001$). These patterns demonstrate that the Mika et al. score captures quality dimensions similar to both DISCERN and NLAT-AI (Total), while also suggesting that DISCERN and NLAT-AI (Total) themselves reflect partially different aspects of quality (Table 6).

The median word count of the responses was 265.5 (range: 148–477). Readability indices converged on a generally difficult-to-read profile: median FRE was 42.5 (range: 20–52), median FKGL was 10.9 (range: 9.25–14.25), median GFI was 12.6 (range: 10.60–15.30), median SMOG was 10.1 (range: 8.29–12.26), and median CLI was 13.8 (range: 11.19–17.89). Overall, this corresponds to approximately an 11th-grade reading level. The mean wordcount across responses is shown (Fig. 3),

whereas mean $\pm$ SD values for the readability metrics are presented (Table 7).

No significant differences were found in DISCERN, NLAT-AI (Total), or Mika et al. scores across the different question categories (Kruskal-Wallis, $P > 0.05$ for all) (Table 8). Similarly, no significant differences were observed in readability metrics across categories ($p > 0.05$). Only the word count showed a significant difference (Kruskal-Wallis $p = 0.002$; post-hoc analysis with Holm correction: Evidence/Funding > Procedure ($p = 0.036$) and > Efficacy ($p = 0.040$) (Table 9).

Discussion

As the role of LLMs in patient education expands, the systematic evaluation of the quality, reliability, and accessibility of the information they produce is becoming increasingly important. This study comprehensively evaluated the information quality, clinical relevance, and readability of ChatGPT-5-generated responses to 30 patient-centered questions regarding SVF in the management of KOA. Evaluation by four observers revealed that ChatGPT-5 responses were of fair-to-good quality across different metrics: a fair quality rating for the DISCERN score, a good quality rating for the NLAT-AI (Total) score, and a fair quality rating for the Mika et al. score. This aligns with previous reports on LLMs in orthopaedics [28, 29]. While this demonstrates the capacity of LLMs to provide structured information on complex medical topics, it also demonstrates that their use alone, without human expert supervision, remains a significant limitation.

The NLAT-AI subscale analysis provided important clues about the model's strengths and weaknesses. Responses showed strong performance in the Effectiveness and Appropriateness domains, but significant weaknesses in the Safety and Actionability domains (Table 5).

Table 6 Inter-scale Spearman rank correlations (direction-aligned)

Pair	Spearman ρ	Correlation strength	p value
DISCERN vs. NLAT-AI (Total)	0.301	Weak	0.1064
DISCERN vs. Mika et al. score (-)	0.466	Moderate	0.0095
NLAT-AI (Total) vs. Mika et al. score (-)	0.874	Very strong	< 0.0001

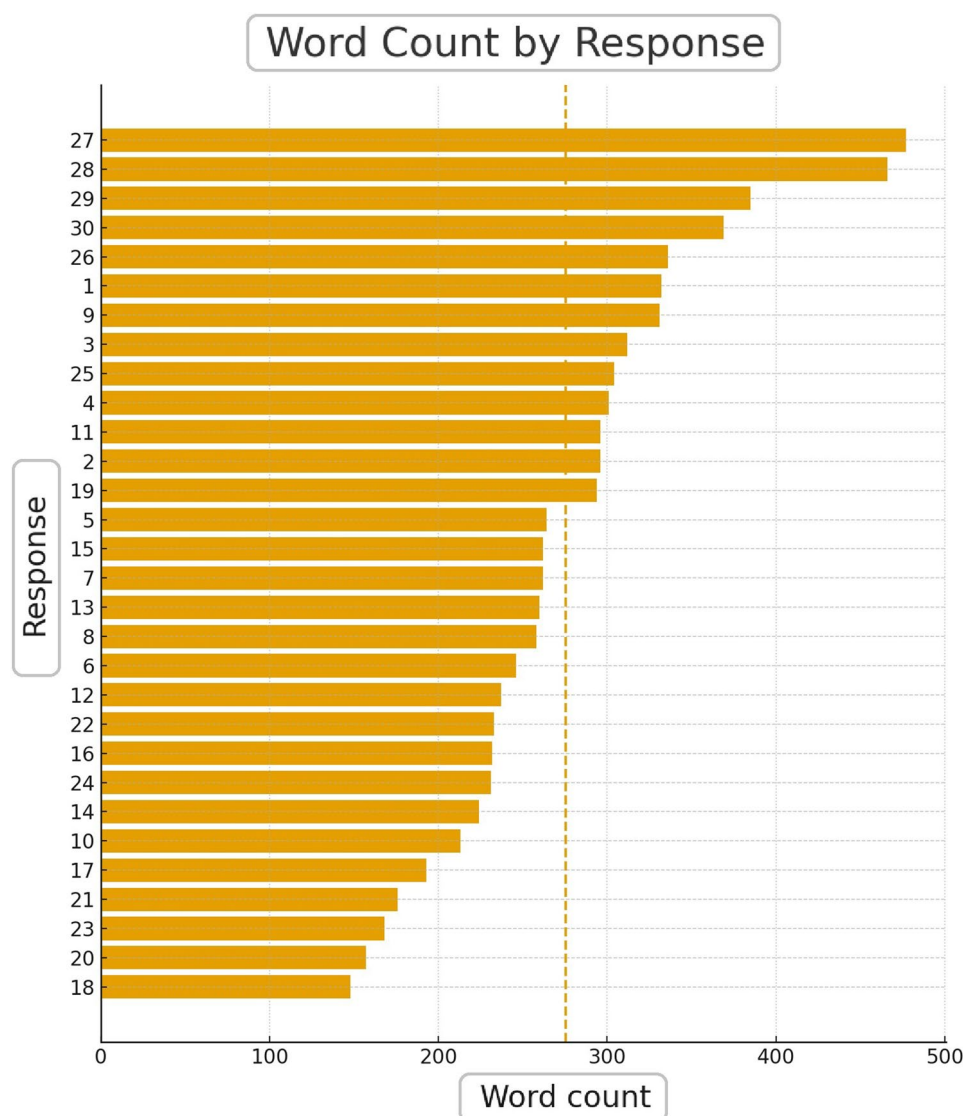


Fig. 3 Word count of ChatGPT responses across the 30 questions. The dashed vertical line indicates the mean word count (275.43) across all responses

Table 7 Descriptive statistics for readability metrics ($n=30$)

Metric	Mean $\pm$ SD	Range Min–Max
Word count	275.43 $\pm$ 80.03	148–477
Flesch Reading Ease	41.43 $\pm$ 6.67	20–52
Flesch-Kincaid Grade Level	10.91 $\pm$ 1.14	9.25–14.25
Gunning Fog Score	12.59 $\pm$ 1.14	10.60–15.30
SMOG Index	10.09 $\pm$ 0.98	8.29–12.26
Coleman Liau Index	13.82 $\pm$ 1.36	11.19–17.89

Clinically, this matters because an answer may sound convincing yet still fail to clearly state who the treatment may not be suitable for, what the key risks are, or when particular caution is needed. This finding is consistent with previous studies reporting that LLMs often fail to provide safety warnings, contraindications, and clear next-step recommendations, but rather excel at

conveying general information [30–32]. These weaknesses are particularly critical for experimental treatments with immature evidence bases, such as SVF [33]. The deficiencies in safety-related information and clear guidance on actionable steps highlight a significant limitation, as patients may not receive adequate warnings regarding potential contraindications or practical next steps, which could lead to harmful outcomes in real-world clinical settings. For SVF therapy in KOA, understanding the potential risks and assessing the treatment's suitability for the individual patient is at least as important as understanding how it may work. This also helps explain the weak DISCERN vs. NLAT-AI (Total) correlation. DISCERN does cover treatment benefits and risks, but it mainly reflects overall patient-information quality and whether the content supports informed decision-making in a balanced manner. NLAT-AI, by contrast,

Table 8 Comparison of scales by question category

Scale	Efficacy (n = 10)	Procedure (n = 8)	Safety/ Eligibility (n = 6)	Evidence/ Funding (n = 4)	Description/ Mechanism (n = 2)	p value*
DISCERN	47.88	40.12	42.38	45.12	45.00	0.069
(Total)	(44.25–51.25)	(34.25–50.75)	(38.00–51.25)	(40.75–49.00)	(42.25–47.75)	
NLAT-AI	17.12	18.38	19.62	21.00	17.75	0.128
(Total)	(15.75–21.00)	(17.00–23.25)	(18.00–22.00)	(18.50–21.50)	(16.00–19.50)	
Mika et al. score	2.50	2.75	2.25	2.25	2.50	0.503
	(1.75–3.00)	(1.75–3.50)	(1.50–2.75)	(1.75–2.50)	(2.00–3.00)	

Data presented as median (range)

*Kruskal–Wallis test

Table 9 Comparison of readability metrics by question category

Metric	Efficacy (n = 10)	Procedure (n = 8)	Safety/ Eligibility (n = 6)	Evidence/ Funding (n = 4)	Description/ Mechanism (n = 2)	p value*
Word count	263 (168–369)	208.5 (148–294)	248.5 (213–304)	425.5 (336–477)	322 (312–332)	0.002
Flesch-Reading Ease	42.5 (20–50)	43.5 (34–50)	43.5 (34–52)	38 (30–52)	41.5 (41–42)	0.823
Flesch-Kincaid Grade Level	10.69 (9.72–14.25)	10.57 (9.84–12.26)	10.80 (9.43–12.15)	11.75 (9.25–12.78)	10.59 (10.46–10.72)	0.943
Gunning Fog Index	11.80 (10.60–15.30)	12.55 (11.00–14.00)	13.05 (11.80–13.90)	13.05 (12.20–14.50)	12.65 (12.10–13.20)	0.282
SMOG Index	9.52 (8.29–12.14)	10.20 (9.44–11.00)	10.36 (8.79–11.27)	10.50 (10.18–12.26)	9.74 (9.29–10.19)	0.275
Coleman-Liau Index	14.32 (13.29–17.89)	12.84 (11.63–15.25)	13.01 (11.19–14.76)	14.22 (13.24–15.05)	14.09 (13.90–14.28)	0.236

Data presented as median (range)

*Kruskal–Wallis test. Post-hoc (Holm-corrected Mann–Whitney U): Evidence/Funding vs. Procedure, $p=0.036$; Evidence/Funding vs. Efficacy, $p=0.040$; other comparisons were not significant

evaluates clinical suitability more directly, including separate domain ratings for Accuracy, Safety, and Actionability. As a result, a response may score well on DISCERN while still scoring lower on NLAT-AI when warnings are not stated clearly enough or when practical guidance is limited. For patient-relevant quality, both perspectives are important. However, Safety and Actionability are especially critical when the content may influence patient decisions, because they reflect whether risk warnings and clear practical guidance are adequately provided.

Readability analyses showed that the responses were written at approximately a high-school level (Table 7), which is above the commonly recommended reading level for patient materials by the American Medical Association (AMA) and the National Institutes of Health (NIH). This is consistent with literature in orthopaedics showing that both traditional web-based content and LLM outputs have similar readability issues [34–36]. High readability levels can make it difficult for patients with low health literacy to understand and use this information. This can obviously be mitigated by simplifying the language and formatting the text in a patient-friendly way. The orthopaedic specialist prompt used in our queries may not accurately reflect how patients ask questions

in real life or what they typically need from an answer, and may have resulted in the model giving more formal responses, which could have contributed to the higher readability levels and the very high scores in the NLAT-AI Effectiveness domain.

The lack of statistically significant differences in DISCERN, NLAT-AI total, Mika et al. scores, and readability metrics across question categories suggests that ChatGPT-5 uses a consistent, category-independent response generation template. This may indicate that ChatGPT-5 adopts a predefined approach to providing general medical information rather than the content of the question. For example, a patient question inquiring about the effectiveness of a treatment and a question asking about the details of the procedure were answered with a similar structure (introduction, subheadings, summary) and at a similar level of language complexity. This consistency can be interpreted as an indication that ChatGPT-5 has internalized the standard medical text structures in the training data. However, this also means that ChatGPT-5 may be insufficient to produce more context-sensitive and deeply focused responses, particularly in more specific and patient-critical categories such as safety and procedure. The literature has reported that LLMs can provide

answers of similar complexity even to medical questions of varying difficulty levels [37–39]. These findings suggest that there is still room for improvement in generating customized content specific to the question type and tailoring readability to context for LLMs.

It is noteworthy that, a statistically significant difference was found only in word count across question categories. Specifically, responses to questions in the Evidence/Funding category were significantly longer than those in the Procedure and Efficacy categories. This suggests that ChatGPT-5 tends to produce more comprehensive and detailed responses to explain complex and multidimensional topics such as FDA approval, insurance coverage, cost, and level of scientific evidence. Conversely, it tends to produce more concise and structured responses in categories that contain relatively standard and descriptive information, such as Procedure or Efficacy. It should be kept in mind that the distribution of question numbers in different question categories may have affected the statistical power of comparisons between categories.

Our findings become more relevant when contextualized within the current state of clinical evidence and regulatory framework for SVF in KOA. While existing literature suggests that SVF can significantly improve pain and function in the short- to medium-term and that its safety profile is generally acceptable, there is considerable heterogeneity across studies in terms of cell preparation methods (enzymatic/mechanical), dose, protocol, and follow-up durations [11, 40, 41]. This methodological heterogeneity may explain why ChatGPT-5 performs relatively well in compiling general and promising information from the literature on efficacy and overall suitability, but falls short on context-sensitive, complex, and rapidly changing issues such as patient selection, contraindications, risk management, procedural details, and, particularly, regulatory status [42]. For example, the stringent regulation of adipose-derived SVF in the United States because it exceeds the minimal manipulation criterion [6] makes it imperative that the model provides clear, up-to-date, and jurisdiction-specific information on this topic.

Our study findings offer several practical implications for clinical use and future research. First, the systematic integration of safety and actionability appears important. Adding more structured safety and actionability elements to LLM responses, including risk-benefit balance, absolute and relative contraindications (red flags), management of potential side effects, and when to consult a physician, may help improve the Safety and Actionability scores of the NLAT-AI. This is an area that may be explored further, including through prompt design. Second, readability optimization is also important for effective patient information delivery. Presenting LLM outputs in a simpler, more patient-friendly format and aiming for a lower reading level may be worth

considering, particularly to better serve patients with varying information needs and literacy levels. The third important implication relates to transparency, source attribution, and uncertainty management. Consistent with the DISCERN logic, responses could include the date of creation, references to sources used, and clear statements regarding the strength/limitations of the evidence (e.g., “Randomized controlled trials (RCTs) are limited,” “Long-term outcomes remain uncertain”), which may help align patients’ expectations with a realistic level. For rapidly evolving therapies such as SVF and those with evolving regulatory frameworks, it may be worth exploring domain-specific and regularly updated knowledge bases that provide LLMs with access to current evidence-based data and current guidance from regulatory organisations like FDA.

Our findings demonstrate ChatGPT-5’s potential as a complementary tool for patient information, but also highlight areas for improvement (e.g., targeted improvements such as engineering to improve response quality and reliability, readability optimization, transparent referencing, and dynamic information update systems). It is anticipated that when the integration of SVF and similar innovative therapies into clinical practice is supported by high-quality evidence from larger-scale, long-term, blinded RCTs and standardized reporting frameworks, ChatGPT-5’s role in patient information and decision support in this dynamic field will be strengthened and become more reliable. These findings are particularly relevant for SVF in KOA, where patient selection, expectation management, and risk discussion are central to the consultation. If caution points and practical direction are incomplete, a reader may be more likely to form unrealistic expectations, misunderstand suitability, or pursue the intervention without appropriate individualized evaluation and follow-up planning. Given the weaker Safety and Actionability, this content should be used only as a starting point for discussion with a physician, not as a stand-alone guide for personal decisions.

This study has several limitations that must be acknowledged. The questionnaire, though rigorously derived from scientific literature, consisted of a limited set of 30 questions designed to simulate patient inquiries. Although the questions were phrased in a direct patient-question style, patients were not involved in selecting the final question set or reviewing the final wording, which may limit how well the set reflects real-life phrasing, patient priorities, and the more spontaneous or emotional nature of real-world patient questions. Future studies may further strengthen the question set by inviting patients to review the wording and to indicate which concerns they prioritize most. The question set used in this study was not a psychometrically validated instrument; therefore, the findings should be considered exploratory and

interpreted with caution when generalizing beyond this question pool. The evaluation of responses, despite excellent inter-rater reliability, was based on the subjective judgments of a panel of orthopaedic specialists. These observers are clinically involved in orthobiological treatments, which may create potential value judgment bias, and assessments by other groups, such as primary care physicians or patients, might yield different perspectives on quality and appropriateness. Because we asked for an orthopaedic specialist perspective, the answers may have been longer and more formal; therefore, readability and Effectiveness scores may be higher than what might be seen when patients ask questions without such guidance. The generalizability of our findings is also constrained, as responses were collected on a single date from a single AI model (ChatGPT-5), whose performance is subject to change with updates. Accordingly, our results should be understood as a snapshot of that specific date, and the same questions may yield different answers as the model changes over time. Future work may repeat the same question set at different times and with different ways of asking the same questions. Our assessment was also limited to English content, which may not reflect performance in other languages. Additionally, we deliberately abstained from verifying the scientific sources occasionally cited by ChatGPT-5. Whether citations appear can depend on how the question is asked, and even when correct sources are provided, they may not be practical for many patients to read and use. Therefore, this study does not comment on citation validity; our ratings reflect the clinical correctness of the answer text as evaluated by orthopaedic specialists. While this was considered ethically responsible to avoid legitimizing potentially hallucinated references, it limited our ability to assess the model's absolute performance in providing source transparency and evidence-based information. The strengths of this study include the use of multiple, validated assessment tools (DISCERN, NLAT-AI, Mika et al. scoring), a four-observer design, a standardized rating scheme, and high interobserver reliability.

Conclusion

In conclusion, when asked to answer from an orthopaedic specialist perspective, ChatGPT-5 provided fair-to-good quality responses to patient-centered questions about SVF in KOA, with strong performance in effectiveness, accuracy, and appropriateness. However, more limited coverage of cautions and practical guidance, together with a relatively high reading level, reduces its suitability as a stand-alone information source. In this setting, the responses are better considered as supportive material to complement clinical discussion rather than replace it, and further research is needed to clarify how such outputs can be used for patient education.

Abbreviations

AMA	American Medical Association
CLI	Coleman-Liau Index
FDA	U.S. Food and Drug Administration
FKGL	Flesch-Kincaid Grade Level
FRE	Flesch Reading Ease
GFI	Gunning Fog Index
ICC	Intraclass Correlation Coefficient
KOA	Knee Osteoarthritis
LLM	Large Language Model
NIH	National Institutes of Health
NLAT-AI	Natural Language Assessment Tool for Artificial Intelligence
SMOG	Simple Measure of Gobbledygook
SVF	Stromal Vascular Fraction

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12891-026-09731-y>.

Supplementary Material 1.

Authors' contributions

Conceptualization, N.M., E.E. and M.E.; methodology, N.M., T.Ç. and S.Ç.; software, N.M. and S.Ç.; validation, M.A.; formal analysis, M.E. and T.Ç.; investigation, N.M. and E.E.; resources, N.M. and S.Ç.; data curation, E.E., M.E. and T.Ç.; writing—original draft preparation, N.M., E.E., M.E. and T.Ç.; writing—review and editing, N.M., S.Ç. and M.A.; visualization, N.M. and T.Ç.; supervision, S.Ç. and M.A.; project administration, N.M., E.E. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the Local Ethics Committee of Selçuk University Faculty of Medicine (decision number: 2026/153). Because the study did not involve any human participants, human data, or human tissue, informed consent was not required. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Orthopaedics and Traumatology, Kahramanmaraş Necip Fazıl City Hospital, Kahramanmaraş 46050, Turkey

²Department of Orthopaedics and Traumatology, Selçuk University Faculty of Medicine, Konya 42130, Turkey

³Department of Orthopaedics and Traumatology, Cihanbeyli State Hospital, Konya 42850, Turkey

⁴Department of Orthopaedics and Traumatology, Faculty of Medicine, Yuksek Ihtisas University, Ankara 06530, Turkey

⁵Department of Orthopaedics and Traumatology, Private Koru Ankara Hospital, Ankara 06510, Turkey

Received: 25 October 2025 / Accepted: 12 March 2026

Published online: 14 March 2026

References

- Global regional, national burden of osteoarthritis. 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Rheumatol.* 2023;5(9):e508–22.
- Safiri S, Kolahi AA, Smith E, Hill C, Bettampadi D, Mansournia MA, et al. Global, regional and national burden of osteoarthritis 1990–2017: a systematic analysis of the Global Burden of Disease Study 2017. *Ann Rheum Dis.* 2020;79(6):819–28.
- Bannuru RR, Osani MC, Vaysbrot EE, Arden NK, Bennell K, Bierma-Zeinstra SMA, et al. OARSi guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. *Osteoarthr Cartil.* 2019;27(11):1578–89.
- Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee. *Arthritis Care Res.* 2020;72(2):149–62.
- Tong L, Yu H, Huang X, Shen J, Xiao G, Chen L, et al. Current understanding of osteoarthritis pathogenesis and relevant new approaches. *Bone Res.* 2022;10(1):60.
- Bourin P, Bunnell BA, Casteilla L, Dominici M, Katz AJ, March KL, et al. Stromal cells from the adipose tissue-derived stromal vascular fraction and culture expanded adipose tissue-derived stromal/stem cells: a joint statement of the International Federation for Adipose Therapeutics and Science (IFATS) and the International Society for Cellular Therapy (ISCT). *Cytotherapy.* 2013;15(6):641–8.
- Nguyen A, Guo J, Banyard DA, Fadavi D, Toronto JD, Wirth GA, et al. Stromal vascular fraction: A regenerative reality? Part 1: Current concepts and review of the literature. *J Plast Reconstr aesthetic surgery: JPRAS.* 2016;69(2):170–9.
- Ramakrishnan VM, Boyd NL. The Adipose Stromal Vascular Fraction as a Complex Cellular Source for Tissue Engineering Applications. *Tissue Eng Part B Reviews.* 2018;24(4):289–99.
- Kuroda K, Kabata T, Hayashi K, Maeda T, Kajino Y, Iwai S, et al. The paracrine effect of adipose-derived stem cells inhibits osteoarthritis progression. *BMC Musculoskelet Disord.* 2015;16:236.
- Desando G, Cavallo C, Sartoni F, Martini L, Parrilli A, Veronesi F, et al. Intra-articular delivery of adipose derived stromal cells attenuates osteoarthritis progression in an experimental rabbit model. *Arthritis Res therapy.* 2013;15(1):R22.
- Hong Z, Chen J, Zhang S, Zhao C, Bi M, Chen X, et al. Intra-articular injection of autologous adipose-derived stromal vascular fractions for knee osteoarthritis: a double-blind randomized self-controlled trial. *Int Orthop.* 2019;43(5):1123–34.
- The Lancet Digital H. Large language models: a new chapter in digital health. *Lancet Digit health.* 2024;6(1):e1.
- Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. *Health care Sci.* 2023;2(4):255–63.
- Roberts H, Zhang D, Dyer GS. The Readability of AAOS Patient Education Materials: Evaluating the Progress Since 2008. *J bone joint Surg Am volume.* 2016;98(17):e70.
- Karimi AH, Shah AK, Hecht CJ 2nd, Burkhart RJ, Acuña AJ, Kamath AF. Readability of Online Patient Education Materials for Total Joint Arthroplasty: A Systematic Review. *J Arthroplast.* 2023;38(7):1392–9.
- Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11(6):887.
- Mesko B. The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals. *J Med Internet Res.* 2023;25:e48392.
- Sullivan B, Platt B, Joiner J, Jacobs C, Conley C, Landy DC, et al. An Investigation of Google Searches for Knee Osteoarthritis and Stem Cell Therapy: What are Patients Searching Online? *HSS journal: the musculoskeletal journal of. Hosp Special Surg.* 2022;18(4):485–9.
- Oo WM, Linklater J, Siddiq MAB, Fu K, Hunter DJ. Comparison of ultrasound guidance with landmark guidance for symptomatic benefits in knee, hip and hand osteoarthritis: Systematic review and meta-analysis of randomised controlled trials. *Australasian J ultrasound Med.* 2024;27(2):97–105.
- Lenhard NK, Williams EE, Lape EC, MacFarlane LA, Losina E, Katz JN. Patient Perspectives Surrounding Intraarticular Injections for Knee Osteoarthritis: A Qualitative Study. *Arthritis Care Res.* 2022;74(3):410–9.
- Charnock D, Shepherd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Commun Health.* 1999;53(2):105–11.
- Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, et al. Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. *J Med Internet Res.* 2024;26:e55939.
- Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. *J bone joint Surg Am volume.* 2023;105(19):1519–26.
- Kalem M, Balaban K, Kocaoğlu H, Kindan P, Şahin E. Evaluation of ChatGPT responses to common patient questions on ankle fusion. *Foot Ankle Surg.* 2025;31(8):708–31.
- Tekin SB, Ince K, Tekin BG, Servet E, Bozgeyik B. Evaluation of Google and ChatGPT responses to common patient questions about scoliosis. *Spine Deform.* 2026;14(1):3–18.
- Celik H, Polat O, Özcan C, Camur S, Kilinc BE, Uzun M. Assessment of the quality and reliability of the information on rotator cuff repair on YouTube. *Orthopaedics & Traumatology, Surgery & Research: OTSR.* 2020;106(1):31–4.
- Uzun M, Cingoz T, Duran ME, Varol A, Celik H. The videos on YouTube® related to hallux valgus surgery have insufficient information. *Foot ankle surgery: official J Eur Soc Foot Ankle Surg.* 2022;28(4):414–7.
- Kasapovic A, Ali T, Babasz M, Bojko J, Gathen M, Kaczmarczyk R, et al. Does the Information Quality of ChatGPT Meet the Requirements of Orthopedics and Trauma Surgery? *Cureus.* 2024;16(5):e60318.
- Yüce A, Erkurt N, Yerli M, Misir A. The Potential of ChatGPT for High-Quality Information in Patient Education for Sports Surgery. *Cureus.* 2024;16(4):e58874.
- Sparks CA, Fasulo SM, Windsor JT, Bankauskas V, Contrada EV, Kraeutler MJ, et al. ChatGPT Is moderately accurate in providing a general overview of Orthopaedic conditions. *JB JS Open Access.* 2024;9(2):e23.00129.
- Jurayj A, Nerys-Figueroa J, Espinal E, Gaudiani MA, Baes T, Mahylis J, et al. Evaluating if ChatGPT can answer common patient questions compared with orthoinfo regarding rotator cuff tears. *J Am Acad Orthop Surg Glob Res Rev.* 2025;9(3):e24.00289.
- Casciato D, Mateen S, Cooperman S, Pesavento D, Brandao RA. Evaluation of Online AI-Generated Foot and Ankle Surgery Information. *J foot ankle surgery: official publication Am Coll Foot Ankle Surg.* 2024;63(6):680–3.
- Im GI. Current Status of Stromal Vascular Fraction from Adipose Tissue in the Clinical Application for Osteoarthritis Treatment. *Tissue Eng regenerative Med.* 2025;22(6):747–54.
- Sabharwal S, Badarudeen S, Unes Kunju S. Readability of online patient education materials from the AAOS web site. *Clin Orthop Relat Res.* 2008;466(5):1245–50.
- Yoo M, Jang CW. Presentation suitability and readability of ChatGPT's medical responses to patient questions about on knee osteoarthritis. *Health Inf J.* 2025;31(1):14604582251315587.
- Çiftci S, Şahin E, Aktaş SA, Safali S, Durgut F, Aydın BK. How understandable are the patient education materials about flat foot on the Internet for parents? *Medicine.* 2023;102(6):e32791.
- Marder RS, Abdelmalek G, Richards SM, Nadeau NJ, Garcia DJ, Attia PJ et al. ChatGPT-3.5 and –4.0 do not reliably create readable patient education materials for common orthopaedic upper- and lower-extremity conditions. *Arthrosc Sports Med Rehabil.* 2025;7(1):101027.
- Roster K, Kann RB, Farabi B, Gronbeck C, Brownstone N, Lipner SR. Readability and Health Literacy Scores for ChatGPT-Generated Dermatology Public Education Materials: Cross-Sectional Analysis of Sunscreen and Melanoma Questions. *JMIR dermatology.* 2024;7:e50163.
- Luo Z, Kam SC, Kim JY, Hu W, Lin C, Park HJ, et al. Does the quality and readability of information related to Varicocele obtained from ChatGPT 4.0 remain consistent across different models of inquiry? *World J Mens Health.* 2026;44(1):161–70.
- Zhang S, Xu H, He B, Fan M, Xiao M, Zhang J, et al. Mid-term prognosis of the stromal vascular fraction for knee osteoarthritis: a minimum 5-year follow-up study. *Stem Cell Res Ther.* 2022;13(1):105.
- Goncharov EN, Koval OA, Bezuglov EN, Encarnacion Ramirez MDJ, Engelgard M, et al. Stromal vascular fraction therapy for knee Osteoarthritis: a systematic review. *Medicina (Kaunas).* 2023;59(12):2090.
- Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, et al. Examining the Role of Large Language Models in Orthopedics: Systematic Review. *J Med Internet Res.* 2024;26:e59607.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.