


RESEARCH

Open Access



Comparative evaluation of artificial intelligence models GPT-4 and GPT-3.5 in clinical decision-making in sports surgery and physiotherapy: a cross-sectional study

Sönmez Saglam^{1*} , Veysel Uludag², Zekeriya Okan Karaduman¹, Mehmet Arıcan¹, Mücahid Osman Yücel¹ and Raşit Emin Dalaslan¹

Abstract

Background The integration of artificial intelligence (AI) in healthcare has rapidly expanded, particularly in clinical decision-making. Large language models (LLMs) such as GPT-4 and GPT-3.5 have shown potential in various medical applications, including diagnostics and treatment planning. However, their efficacy in specialized fields like sports surgery and physiotherapy remains underexplored. This study aims to compare the performance of GPT-4 and GPT-3.5 in clinical decision-making within these domains using a structured assessment approach.

Methods This cross-sectional study included 56 professionals specializing in sports surgery and physiotherapy. Participants evaluated 10 standardized clinical scenarios generated by GPT-4 and GPT-3.5 using a 5-point Likert scale. The scenarios encompassed common musculoskeletal conditions, and assessments focused on diagnostic accuracy, treatment appropriateness, surgical technique detailing, and rehabilitation plan suitability. Data were collected anonymously via Google Forms. Statistical analysis included paired t-tests for direct model comparisons, one-way ANOVA to assess performance across multiple criteria, and Cronbach's alpha to evaluate inter-rater reliability.

Results GPT-4 significantly outperformed GPT-3.5 across all evaluated criteria. Paired t-test results ($t(55) = 10.45$, $p < 0.001$) demonstrated that GPT-4 provided more accurate diagnoses, superior treatment plans, and more detailed surgical recommendations. ANOVA results confirmed the higher suitability of GPT-4 in treatment planning ($F(1, 55) = 35.22$, $p < 0.001$) and rehabilitation protocols ($F(1, 55) = 32.10$, $p < 0.001$). Cronbach's alpha values indicated higher internal consistency for GPT-4 ($\alpha = 0.478$) compared to GPT-3.5 ($\alpha = 0.234$), reflecting more reliable performance.

Conclusions GPT-4 demonstrates superior performance compared to GPT-3.5 in clinical decision-making for sports surgery and physiotherapy. These findings suggest that advanced AI models can aid in diagnostic accuracy, treatment planning, and rehabilitation strategies. However, AI should function as a decision-support tool rather than a substitute for expert clinical judgment. Future studies should explore the integration of AI into real-world clinical workflows, validate findings using larger datasets, and compare additional AI models beyond the GPT series.

*Correspondence:
Sönmez Saglam
dr.sonmezsaglam@gmail.com

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



© The Author(s) 2025. **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/>.

Keywords Artificial intelligence, Large language models, Sports surgery, Physiotherapy, Clinical decision-making, Rehabilitation, Musculoskeletal disorders

Introduction

The utilization of artificial intelligence (AI) technologies in healthcare has gained significant momentum in recent years, particularly in clinical decision-making processes [1–4]. Advances in natural language processing (NLP) technologies have highlighted the potential applicability of large language models (LLMs) like the GPT series in addressing complex tasks such as medical decision-making and patient management. The existing literature demonstrates the promise of AI in improving diagnostic accuracy and optimizing treatment processes across various medical fields. For instance, Lopez et al. (2020) [5] explored AI's impact in cardiology, while Clark et al. (2024) [6] examined its accuracy in classifying dermatological conditions. However, there is a notable gap in the literature regarding the performance of AI-driven technologies in specialized fields such as sports surgery and physiotherapy, which demand highly specialized expertise.

Sports surgery and physiotherapy are domains characterized by intricate clinical decision-making processes that require expertise in both diagnosis and treatment planning. Accurate diagnosis and effective treatment strategies in these areas not only improve patient outcomes but are also critical for athletes to sustain their professional careers. Despite the growing body of research on AI applications in general medical practice, there remains a lack of comprehensive studies assessing its impact in these niche areas [7–11]. Although some recent studies have investigated AI's role in musculoskeletal disorders, the extent to which AI can contribute to specific clinical decision-making processes in sports surgery and physiotherapy has not been fully explored [12–15].

In recent years, AI and large language models (LLMs) have been increasingly integrated into decision support systems in orthopedics and physiotherapy. Several studies have examined the potential role of AI in improving diagnostic and treatment decision-making. Kunze et al. (2024) evaluated the ability of GPT-4 to diagnose and triage patients with knee pain, demonstrating that AI can assist in clinical decision-making by improving diagnostic consistency and reducing variability in physician assessments [12]. Similarly, Lintz et al. (2024) investigated the capacity of AI models to accurately classify patients requiring surgical intervention for foot and ankle surgery, finding that AI-supported triage could improve patient management efficiency [13]. In another study, Nwachukwu et al. (2025) analyzed the extent to which current LLMs align with evidence-based clinical guidelines in the

management of musculoskeletal diseases and highlighted discrepancies between AI-generated treatment recommendations and expert guidelines [14]. Furthermore, Truhn et al. (2023) assessed GPT-4's capability to generate orthopedic treatment recommendations based on MRI reports, demonstrating its potential role in radiology-based decision support systems [15]. Beyond orthopedics, AI applications in physiotherapy have also been examined. Villagrán et al. (2024) explored how LLMs can be used in physiotherapy education to provide automated feedback to students, indicating that AI has the potential to support both clinical decision-making and educational frameworks [16].

Despite the emerging evidence supporting AI's integration into musculoskeletal healthcare, there is a lack of comprehensive research evaluating its direct impact on clinical decision-making in sports surgery and physiotherapy. Current studies primarily focus on AI's diagnostic accuracy and triage capabilities, but its role in treatment planning, surgical technique recommendations, and rehabilitation program design remains largely unexamined. Given the complexity of sports-related injuries and the necessity for individualized rehabilitation approaches, understanding AI's ability to provide clinically relevant and evidence-based recommendations is essential.

This study represents the first attempt to systematically evaluate and compare the performance of GPT-4 and GPT-3.5 AI models in clinical decision-making within the fields of sports surgery and physiotherapy. By assessing these models in key domains—diagnostic accuracy, treatment suitability, surgical technique detailing, and rehabilitation plan validity—this research aims to address the existing knowledge gap. The findings of this study are expected to inform healthcare professionals about the strengths and limitations of AI-driven decision support systems, contributing to the broader integration of AI in specialized medical practice.

Methods

Study design

This study was designed as a cross-sectional, observational, and comparative research project to evaluate the performance of GPT-4 and GPT-3.5 AI models in clinical decision-making processes in sports surgery and physiotherapy. The study adhered to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines to ensure methodological rigor [17]. The performance of both models was analyzed based on four key clinical criteria: diagnostic accuracy,

treatment suitability, surgical technique detailing, and the validity of rehabilitation plans. The study protocol was approved by the Duzce University Non-Interventional Clinical Research Ethics Committee (No: 2024/215, Date: 21/10/2024) and conducted in accordance with the principles of the Declaration of Helsinki. Informed consent was obtained from all participants before data collection.

Participants

A total of 56 professionals specializing in sports surgery and physiotherapy participated in the study. Participants were recruited through professional networks and hospital affiliations to ensure a diverse representation of expertise. Participants were required to have a minimum of 5 years of clinical experience, be actively working in their field, and have basic knowledge of AI technologies. Those from unrelated medical specialties, retirees, and individuals with less than 5 years of experience were excluded. All participants were blinded to the AI model that generated each scenario to minimize bias in their evaluations.

Data collection and scenario development

Data collection was conducted via Google Forms to ensure accessibility and anonymity. Participants were presented with 10 standardized clinical scenarios, which are detailed in Table 1, and asked to evaluate the AI-generated responses using a 5-point Likert scale. The clinical scenarios were developed based on real-world patient data extracted from the hospital's information system, analyzing the last 10 years of patient records from orthopedic clinics. The most frequently encountered musculoskeletal conditions in sports surgery and physiotherapy were identified based on case prevalence and clinical significance. Two independent orthopedic surgeons and two independent physiotherapists reviewed the selected cases and finalized the scenarios to ensure clinical validity and diversity. Each scenario was independently processed through GPT-4 and GPT-3.5, which generated corresponding diagnoses, treatment plans, surgical recommendations, and rehabilitation protocols. AI-generated outputs were reviewed and validated by the independent panel before being presented to study participants. The details of the 10 clinical scenarios evaluated in this study are presented in Table 1.

Table 1 Clinical scenario examples used in the study

Scenario no	Scenario title	Patient profile summary	Diagnosis summary	Treatment summary
1	ACL Tear	23-year-old male, professional basketball player, severe knee pain, instability, MRI confirms ACL tear	ACL tear	Arthroscopic ACL reconstruction with hamstring/patellar tendon graft, followed by structured physiotherapy.
2	Rotator Cuff Tear	35-year-old female, professional swimmer, shoulder pain, weakness, MRI confirms partial supraspinatus tear	Supraspinatus tendon tear	Arthroscopic repair, immobilization for 4–6 weeks, followed by active rehabilitation exercises.
3	Meniscus Tear	28-year-old male, amateur football player, locking sensation in knee, MRI confirms medial meniscus tear	Medial meniscus tear	Arthroscopic meniscus repair or partial meniscectomy, gradual progression to full weight-bearing and functional exercises.
4	Shoulder Dislocation	27-year-old male, bodybuilder, recurrent shoulder dislocations, X-ray confirms anterior shoulder dislocation	Recurrent anterior shoulder dislocation	Latarjet procedure for stabilization, early passive mobilization postoperatively, and advanced strengthening exercises.
5	Achilles Tendon Rupture	30-year-old male, amateur runner, severe pain in Achilles region, MRI confirms full-thickness rupture	Achilles tendon rupture	Open surgical repair, immobilization in a plantarflexed position, gradual transition to functional and sports-specific rehabilitation.
6	Lateral Epicondylitis (Tennis Elbow)	32-year-old female, professional tennis player, persistent elbow pain unresponsive to conservative treatment	Lateral epicondylitis	Surgical debridement of Extensor Carpi Radialis Brevis tendon, progressive mobilization, and strengthening of wrist and elbow extensor muscles.
7	Plantar Fasciitis	40-year-old male, amateur runner, chronic heel pain worsened in the morning, ultrasound shows plantar fascia thickening	Plantar fasciitis	Partial fasciotomy, initial immobilization, followed by progressive functional rehabilitation and strength training.
8	Patellar Tendon Tear	29-year-old male, basketball player, inability to extend knee actively, MRI confirms full-thickness patellar tendon tear	Patellar tendon rupture	Open surgical repair, initial immobilization with knee brace, followed by quadriceps strengthening and functional recovery exercises.
9	Bankart Lesion (Labrum Tear)	24-year-old female, amateur volleyball player, shoulder instability, MRI confirms anteroinferior labrum tear	Anteroinferior labrum tear (Bankart)	Arthroscopic Bankart repair, gradual return to active motion and strengthening exercises, eventual progression to sports-specific activities.
10	Tibial Stress Fracture	26-year-old male, long-distance runner, localized shin pain, MRI confirms tibial stress fracture	Tibial stress fracture	Intramedullary nailing, early partial weight-bearing, progressive return to functional exercises and long-distance running.

Table 2 Demographic characteristics of the participants

Variable	Value
Clinician	31
Academic	25
Total Participants	56
Orthopedist (Clinician)	18
Orthopedist (Academic)	10
Total Orthopedists	28
Physiotherapist (Clinician)	13
Physiotherapist (Academic)	15
Total Physiotherapists	28
Experience (Mean) - Orthopedists	15.7
Experience (Std) - Orthopedists	6.9
Experience (Min) - Orthopedists	5
Experience (Max) - Orthopedists	25
Experience (Mean) - Physiotherapists	15.7
Experience (Std) - Physiotherapists	7.3
Experience (Min) - Physiotherapists	5
Experience (Max) - Physiotherapists	25

Statistical analysis

The study data were analyzed using various statistical methods to evaluate the performance of the GPT-4 and GPT-3.5 models in clinical decision-making within sports surgery and physiotherapy contexts. Descriptive statistics were used to summarize the demographic characteristics of the participants, including the distribution of clinicians and academics, as well as the average years of experience for orthopedists and physiotherapists.

A power analysis was conducted using G*Power 3.1 software to determine the minimum required sample size for detecting a significant difference between the two AI models. Assuming an effect size (Cohen's d) of 0.80, a significance level of $\alpha=0.05$, and a power of 0.80, the minimum required sample size was determined to be 32 participants. Since the study included 56 participants, the sample size was considered adequate for statistical comparisons [18]. A paired t-test was conducted to compare the overall performance scores of the GPT-4 and GPT-3.5 models, assessing whether a significant difference existed between them. Additionally, an independent t-test was used to compare the performance evaluations made by academics and clinicians for the GPT-4 model. A one-way analysis of variance (ANOVA) was employed to examine the model performance based on specific criteria such as diagnostic accuracy, treatment suitability, surgical technique detail, and rehabilitation plan appropriateness. In addition to p-values, eta squared (η^2) effect sizes were calculated to assess the magnitude of differences observed in the ANOVA analysis.

To measure internal consistency, Cronbach's alpha was calculated for both models. Furthermore, Cohen's d was computed to quantify the effect sizes of the differences observed between GPT-4 and GPT-3.5 and between

Table 3 T-test results for GPT-4 vs. GPT-3.5 and academics vs. clinicians (GPT-4)

Comparison	t-statistic	p-value
GPT-4 vs. GPT-3.5	10.45	< 0.001*
Academics vs. Clinicians (GPT-4)	-2.12	0.039*

* $p < 0.05$ indicates statistical significance

Table 4 One-way ANOVA results for the performance criteria

Criterion	F-statistic	p-value
Diagnosis Accuracy	28.45	< 0.001*
Treatment Suitability	35.22	< 0.001*
Surgical Technique Detail	25.67	< 0.001*
Rehabilitation Plan Appropriateness	32.10	< 0.001*

* $p < 0.05$ indicates statistical significance

academics and clinicians' evaluations of GPT-4. Statistical significance was set at $p < 0.05$ for all analyses.

Results

The performance of the GPT-4 and GPT-3.5 models was assessed based on several key criteria: diagnosis accuracy, treatment suitability, surgical technique detail, and rehabilitation plan appropriateness. A total of 56 participants, consisting of 31 clinicians and 25 academics, were involved in the evaluation. Among them, 28 were orthopedists (18 clinicians, 10 academics) and 28 were physiotherapists (13 clinicians, 15 academics). The average years of experience for orthopedists was 15.7 years ($SD=6.9$), and for physiotherapists, it was 15.7 years ($SD=7.3$). These demographic characteristics are summarized in Table 2.

A paired t-test revealed a statistically significant difference between the overall performance scores of GPT-4 and GPT-3.5, with GPT-4 outperforming GPT-3.5 across all scenarios ($t(55)=10.45$, $p<0.001$). Additionally, an independent t-test comparing the evaluations made by academics and clinicians for the GPT-4 model indicated that clinicians rated GPT-4 higher than academics did ($t(54) = -2.12$, $p=0.039$). These results are summarized in Table 3.

Further analysis using one-way ANOVA confirmed that GPT-4 demonstrated significantly better performance across all specific criteria evaluated. The most pronounced differences were observed in treatment suitability ($F(1, 55)=35.22$, $p<0.001$) and rehabilitation plan appropriateness ($F(1, 55)=32.10$, $p<0.001$). Significant differences were also found in diagnostic accuracy ($F(1, 55)=28.45$, $p<0.001$) and surgical technique detail ($F(1, 55)=25.67$, $p<0.001$), reinforcing the superiority of GPT-4. These results are detailed in Table 4.

Reliability analysis using Cronbach's alpha revealed moderate internal consistency in the evaluations of GPT-4 ($\alpha=0.478$), whereas GPT-3.5 showed lower consistency ($\alpha=0.234$). This suggests that participants were

more consistent in their ratings of GPT-4 across different criteria than in their ratings of GPT-3.5. To further evaluate the effect sizes of the observed differences, Cohen's d was calculated. The comparison of GPT-4 vs. GPT-3.5 yielded a large effect size ($d = 1.42$), whereas the difference between academic and clinical evaluations of GPT-4 showed a medium effect size ($d = 0.58$). These results are presented in Table 5.

The overall findings of this study demonstrate that GPT-4 significantly outperforms GPT-3.5 across all evaluated criteria, with the most notable differences observed in treatment suitability and rehabilitation plan appropriateness. These results indicate that GPT-4 may provide more clinically relevant and reliable decision-making support in the context of sports surgery and physiotherapy compared to GPT-3.5. However, further research is required to validate these findings in real-world clinical settings. The comparative performance of GPT-4 and GPT-3.5 across the four key criteria is illustrated in Fig. 1.

Figure 1 illustrates the mean performance scores of the GPT-4 and GPT-3.5 models across four key clinical decision-making criteria: diagnosis accuracy, treatment suitability, surgical technique detail, and rehabilitation plan appropriateness. The blue bars represent the performance of GPT-4, while the orange bars represent GPT-3.5. As shown, GPT-4 consistently outperforms GPT-3.5 across all evaluated criteria, with the most notable differences observed in treatment suitability and rehabilitation plan appropriateness. The performance scores are annotated above each bar to provide a clear comparison between the two models.

Table 5 Reliability and effect size analysis

Model/comparison	Cronbach's alpha	Cohen's d	Effect size interpretation
GPT-4	0.478	-	-
GPT-3.5	0.234	-	-
GPT-4 vs. GPT-3.5	-	1.42	Large
Academics vs. Clinicians (GPT-4)	-	0.58	Medium

Discussion

This study represents an important step in evaluating and comparing the performance of GPT-4 and GPT-3.5 in clinical decision-making within sports surgery and physiotherapy. While AI technologies have been extensively studied in various medical domains, their application in specialized areas such as sports surgery and physiotherapy has been under-explored [9, 19–21]. By addressing this gap, the study provides valuable insights into the capabilities and limitations of AI in these high-expertise medical fields. The findings demonstrate that GPT-4 significantly outperformed GPT-3.5 across multiple clinical criteria, including diagnostic accuracy, treatment suitability, surgical technique detailing, and rehabilitation plan appropriateness. These results align with previous studies highlighting AI's potential in enhancing clinical decision-making [22, 23], yet this is one of the first studies to directly assess AI performance in sports surgery and physiotherapy.

The results indicate that GPT-4 significantly outperformed GPT-3.5, as supported by statistical analyses. Paired t -test results ($t(55) = 10.45$, $p < 0.001$) confirmed that GPT-4 provided more accurate diagnoses, better treatment plans, and more detailed surgical techniques

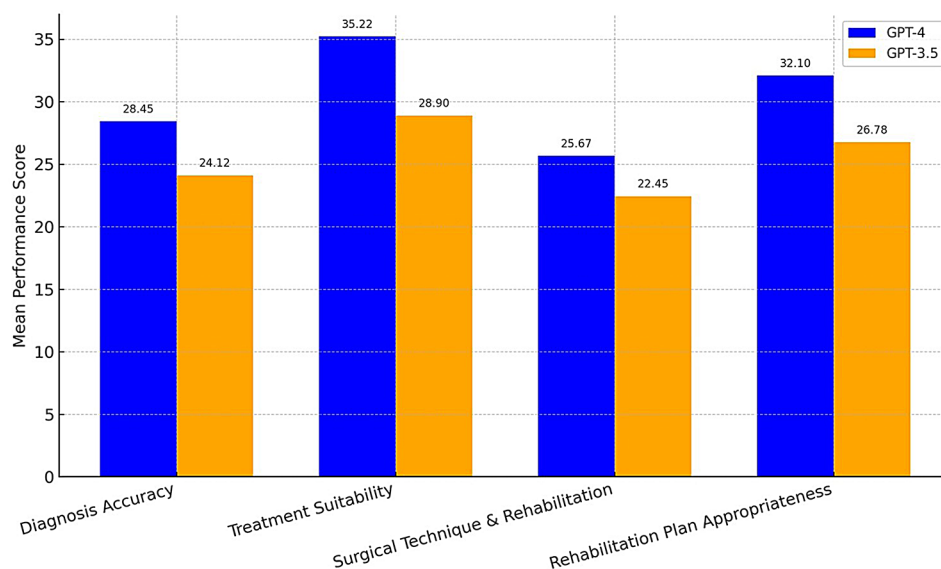


Fig. 1 Comparison of GPT-4 and GPT-3.5 models across performance criteria

compared to GPT-3.5. These results are consistent with prior studies, which have found that more advanced AI models, such as GPT-4, tend to exhibit superior performance in clinical decision support systems [24, 25]. This performance improvement can likely be attributed to GPT-4's more extensive training dataset and advanced natural language processing capabilities, which enable it to generate more accurate, reliable, and contextually appropriate medical recommendations. AI-driven decision support tools have been explored in various medical fields, with research indicating that LLMs can assist in triaging, diagnostic assessments, and treatment planning [12–14].

However, it is important to note that GPT-3.5 still demonstrated moderate performance, suggesting that less complex cases could still benefit from its clinical applications. Nonetheless, for more complex cases requiring precise decision-making, GPT-4 appears to be a more reliable and effective tool.

The findings of this study are in line with existing research, which has investigated the use of LLMs in medical decision-making. For instance, Kunze et al. (2024) demonstrated that GPT-4 performed well in knee pain triage, while Lintz et al. (2024) assessed its capability in surgical triage for foot and ankle conditions [12, 13]. Additionally, Nwachukwu et al. (2025) found that current LLMs do not fully align with evidence-based musculoskeletal treatment guidelines, raising concerns about the reliability of AI-generated recommendations [14]. The findings of this study further highlight that while GPT-4 can provide highly relevant clinical suggestions, it should still be used as a complementary tool rather than an independent decision-maker.

Similarly, Truhn et al. (2023) evaluated GPT-4's ability to generate orthopedic treatment recommendations from MRI reports, suggesting that LLMs can assist in image-based diagnostics [15]. This aligns with our findings, which indicate that GPT-4 demonstrated superior accuracy in treatment suitability and rehabilitation planning. Additionally, Villagrán et al. (2024) explored the role of AI in physiotherapy education, indicating that LLMs may enhance clinical reasoning and learning in medical training [16]. The integration of AI into physiotherapy and sports surgery education could further optimize the use of AI in clinical environments.

A key finding of this study was the higher internal consistency of GPT-4 compared to GPT-3.5, as demonstrated by Cronbach's alpha values (GPT-4: $\alpha=0.478$, GPT-3.5: $\alpha=0.234$). Although GPT-4 exhibited better consistency, the moderate reliability scores suggest some variability in AI-generated recommendations. This variation could be attributed to the nature of AI decision-making, where responses depend on training data and contextual interpretation. Additionally, participants may

have had different subjective evaluations of AI-generated clinical recommendations, further contributing to inter-rater variability. The results emphasize the need for continuous validation of AI-generated recommendations in clinical settings [26].

The integration of LLMs such as GPT-4 into clinical practice offers several potential benefits. AI-driven decision support tools have been shown to reduce cognitive load on healthcare professionals, improve diagnostic accuracy, and enhance treatment planning [27]. In sports surgery and physiotherapy, where rapid and evidence-based decision-making is crucial, AI models may serve as valuable clinical assistants. These models could be integrated into electronic health record systems to provide real-time decision support and automated treatment recommendations, reducing the burden on clinicians and improving efficiency.

Recent studies have highlighted the increasing role of AI and large language models (LLMs) in clinical decision-making and healthcare applications. Naqvi et al. (2024) emphasized that while AI-driven models hold promise for physiotherapy applications, their real-world integration requires careful validation, particularly in decision-support scenarios where clinical expertise remains irreplaceable [28]. Furthermore, Rossetini et al. (2023) reviewed the implications of AI in musculoskeletal rehabilitation, stressing both its potential benefits and limitations. They suggested that AI could enhance clinical workflows by supporting diagnostic and rehabilitation strategies but should not replace human clinical reasoning [29]. These findings align with our study, which demonstrated that GPT-4 significantly outperformed GPT-3.5 in diagnostic accuracy, treatment planning, and rehabilitation protocol design. However, as both studies suggest, AI should be viewed as an assistive tool rather than a substitute for expert clinical judgment. Future research should explore how AI can be effectively integrated into sports surgery and physiotherapy practice while maintaining clinician oversight and patient safety.

However, AI-generated outputs must always be interpreted by a trained clinician, as AI models lack contextual understanding, patient-specific considerations, and the ability to adapt to unforeseen clinical complexities. Future research should focus on improving AI transparency, ensuring clinical validation, and integrating AI into interdisciplinary healthcare workflows.

While this study provides novel insights, several limitations should be acknowledged. First, the study was limited to GPT-4 and GPT-3.5, meaning that the findings may not be generalizable to other AI models such as DeepSeek, Llama, or Gemini. Future research should compare a broader range of AI models to determine their relative effectiveness in clinical decision-making.

Additionally, the study relied on standardized clinical scenarios, which, while designed to reflect real-world cases, may not fully capture the variability and complexity of actual patient cases. Future research should incorporate real patient data and assess AI recommendations in a real-time clinical setting to validate these findings.

Another limitation is the potential bias in scenario creation, as two orthopedic surgeons and two physiotherapists reviewed and finalized the cases. While efforts were made to ensure unbiased scenario development, real-world clinical cases may introduce additional complexity. Future studies should explore AI performance in a more dynamic and diverse patient population.

Finally, AI models, including GPT-4, are susceptible to “hallucination” (the generation of inaccurate or misleading information). This issue remains a critical limitation of AI in healthcare, and future research should develop more robust safeguards to prevent AI-generated misinformation from influencing clinical decision-making [30].

Conclusions

This study provides strong evidence that GPT-4 outperforms GPT-3.5 in clinical decision-making within sports surgery and physiotherapy. The findings demonstrate that AI models, particularly GPT-4, can enhance diagnostic accuracy, improve treatment planning, and support healthcare professionals in making more informed decisions. However, AI should be used as a complementary tool rather than a replacement for clinical expertise.

Future research should focus on expanding the variety of AI models tested, utilizing larger and more diverse datasets, and further exploring AI's integration into real-world clinical environments. Additionally, the ethical, legal, and practical implications of AI in clinical decision-making should be further explored to ensure safe and effective AI adoption in healthcare.

Acknowledgements

Not available.

Author contributions

SS, VU and ZOK contributed to the analysis and interpretation of the data and to the writing and revision of the manuscript. SS and MA performed the surgical operations. ZOK, RED, MOY and VU contributed to data analysis, interpretation and writing. RED, MOY and MA contributed to data collection, analysis and methodology. ZOK, MA, SS and MOY contributed to experimental design, data collection and data revision.

Funding

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

Data availability

All data and materials can be requested from Dr. Sönmez Saglam when needed. (sonmezsaglam@duzce.edu.tr).

Declarations

Ethical approval

The study protocol was approved by the Duzce University Non-Interventional Clinical Research Ethics Committee (No: 2024/215, Date: 21/10/2024). The study was conducted in accordance with the principles of the Declaration of Helsinki.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Congress

Not available.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Orthopaedics and Traumatology, Faculty of Medicine, Duzce University, Duzce, Türkiye

²Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Duzce University, Duzce, Türkiye

Received: 27 December 2024 / Accepted: 7 April 2025

Published online: 14 April 2025

References

1. Pawelczyk J, Kraus M, Eckl L, Nehrer S, Aurich M, Izadpanah K, Siebenlist S, Rupp MC. Attitude of aspiring orthopaedic surgeons towards artificial intelligence: a multinational cross-sectional survey study. *Arch Orthop Trauma Surg.* 2024;144(8):3541–52.
2. Shelmerdine SC, Pauling C, Allan E, Langan D, Ashworth E, Yung KW, Barber J, Haque S, Rosewarne D, Woznitza N, et al. Artificial intelligence (AI) for paediatric fracture detection: a multireader multicase (MRMC) study protocol. 2024;14(12):e084448.
3. Ayorinde A, Mensah DO. Health care professionals' experience of using AI. *Syst Rev Narrative Synthesis.* 2024;26:e55766.
4. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan L-N. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. *BMC Med Inf Decis Mak.* 2024;24(1):1–8.
5. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, Kapa S, Lerman A, Luong C, Medina-Inojosa JR, et al. Artificial Intelligence in Cardiology: Present and Future. *Mayo Clinic Proc.* 2020;95(5):1015–1039.
6. Clark Lambert W, Grzybowski A. Dermatology and artificial intelligence. *Clin Dermatol.* 2024;42(3):207–9.
7. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. *J Experimental Orthop.* 2023;10(1):74.
8. Cheng K, Guo Q, He Y, Lu Y, Xie R, Li C, Wu H. Artificial intelligence in sports medicine: could GPT-4 make human Doctors obsolete?? *Ann Biomed Eng.* 2023;51(8):1658–62.
9. Tack C. Artificial intelligence and machine learning applications in musculoskeletal physiotherapy. *Musculoskelet Sci Pract.* 2019;39:164–9.
10. Sherazi A, Canes D. Comprehensive analysis of the performance of GPT-3.5 and GPT-4 on the American urological association self-assessment study program exams from 2012–2023. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada;* 2023.
11. Park Y-J, Pillai A, Deng J, Guo E, Gupta M, Paget M, Naugler C. Assessing the research landscape and clinical utility of large Language models: A scoping review. *BMC Med Inf Decis Mak.* 2024;24(1):72.
12. Kunze KN, Varady NH, Mazzucco M, Lu AZ, Chahla J, Martin RK, Ranawat AS, Pearle AD, Williams RJ. 3rd: The Large Language Model ChatGPT-4 Exhibits Excellent Triage Capabilities and Diagnostic Performance for Patients Presenting With Various Causes of Knee Pain. *Arthroscopy: the journal of arthroscopic & related surgery: Official publication of the Arthroscopy Association of North America and the International Arthroscopy Association.* 2024.

13. Lintz F, Acker A, Carvalho KA, Labidi M, Gonzi G, Munoz M-A, Joan Luo E, Bernasconi A, Easley ME. Cesar Netto cd: are large Language models efficient as triage tools for surgical management of foot and ankle patients?? *Foot Ankle Orthop.* 2024;9(4):247301142452473000142.
14. Nwachukwu BU, Varady NH, Allen AA, Dines JS, Altchek DW, Williams RJ III, Kunze KN. Currently available large Language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy: J Arthroscopic Relat Surg.* 2025;41(2):263–75. e266.
15. Truhn D, Weber CD, Braun BJ, Bressem K, Kathner JN, Kuhl C, Nebelung S. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci Rep.* 2023;13(1):20159.
16. Villagrán I, Hernández R, Schuit G, Neyem A, Fuentes-Cimma J, Miranda C, Hilliger I, Durán V, Escalona G, Varas J. Implementing artificial intelligence in physiotherapy education: A case study on the use of large Language models (LLM) to enhance feedback. *IEEE Trans Learn Technol.* 2024.
17. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, Initiative S. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg.* 2014;12(12):1495–9.
18. Demir S. A comparative analysis of GPT-3.5 and GPT-4.0 on a multiple-choice ophthalmology question bank: A study on artificial intelligence developments. *Romanian J Ophthalmol.* 2024;68(4):367–71.
19. Niel O, Bastard P. Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *Am J Kidney Diseases: Official J Natl Kidney Foundation.* 2019;74(6):803–10.
20. Hosny A, Parmar C, Quackenbush J. Artificial intelligence in radiology. 2018;18(8):500–10.
21. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, Allison T, Arnaout O, Abbosh C, Dunn IF, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. 2019;69(2):127–57.
22. Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, Sanchez-Barrueco A, Saga-Gutierrez C. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Archives oto-rhino-laryngology: Official J Eur Federation Oto-Rhino-Laryngological Soc (EUFOS): Affiliated German Soc Oto-Rhino-Laryngology - Head Neck Surg.* 2024;281(4):2081–6.
23. Lahat A, Sharif K. Assessing generative pretrained Transformers (GPT) in clinical Decision-Making: Comparative analysis of GPT-3.5 and GPT-4. 2024;26:e54571.
24. Yüce A, Yerli M, Misir A. Can Chat-GPT assist orthopedic surgeons in evaluating the quality of rotator cuff surgery patient information videos? *J Shoulder Elbow Surg.* 2025;34(1):141–6.
25. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus.* 2023;15(6):e40822.
26. Hirosawa T, Harada Y. Evaluating ChatGPT-4's Accuracy in Identifying Final Diagnoses Within Differential Diagnoses Compared With Those of Physicians: Experimental Study for Diagnostic Cases. 2024;8:e59267.
27. Santamato V, Tricase C, Faccilongo N, Iacoviello M, Marengo A. Exploring the impact of artificial intelligence on healthcare management: A combined systematic review and Machine-Learning approach. *Appl Sci.* 2024;14(22):10144.
28. Naqvi WM, Shaikh SZ, Mishra GV. Large Language models in physical therapy: time to adapt and adept. *Front Public Health.* 2024;12:1364660.
29. Rossetini G, Cook C, Palese A, Pillastrini P, Turolla A. Pros and cons of using artificial intelligence chatbots for musculoskeletal rehabilitation management. *J Orthop Sports Phys Ther.* 2023;53(12):728–34.
30. Khan MP, O'Sullivan ED. A comparison of the diagnostic ability of large Language models in challenging clinical cases. *Front Artif Intell.* 2024;7:1379297.

Publisher's note

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