J Turk Spinal Surg 2025;36(3):125-129
DOI: 10.4274/jtss.galenos.2025.62207

A COMPARATIVE STUDY OF ORTHOPEDIC SURGEONS AND AI MODELS IN THE CLINICAL EVALUATION OF SPINAL SURGERY

Muhammed Taha Demir[1], Yiğit Kültür[2]

[1]OrthoCure Clinic, İstanbul, Türkiye
[2]Yeni Yüzyıl University Gaziosmanpaşa Hospital, Department of Ortopaedics and Traumatology, İstanbul, Türkiye

ABSTRACT

Objective: Spinal surgery (SS) is an area characterized by high intra-operative challenges and higher complication rates compared to several other surgical specialties. The purpose of this study is to evaluate the effectiveness of artificial intelligence (AI) instruments-Chat Generative Pre-trained Transformer (ChatGPT)-4o, DeepSeek-V3, and Gemini Pro-in patient assessment and the clinical decision-making process compared with specialists of orthopedic surgery on a series of case-based and knowledge-based questions relevant to SS.

Materials and Methods: By two experienced orthopedic surgeons, a set of 50 questions has been created, including 25 requiring clinical judgement through the use of a case presentation format and 25 to test theoretical understanding. The test was given to two groups: Group 1 included three AI software programs (ChatGPT-4.0, DeepSeek-V3, Gemini Pro) and Group 2 included ten experienced orthopedic surgeons. The answers given were scored independently by the two expert surgeons.

Results: Group 2 performed significantly better than Group 1 in the case-based questions. There was a significant difference between the groups in one section (p=0.025), while there was no significant difference for the knowledge-based questions section (p=1.000). On the assessment of total correct responses, Group 2's performance was significantly better (p=0.036).

Conclusion: AI technologies have proved their utility for knowledge-based tasks but are dramatically inferior to clinicians for areas requiring clinical judgement and case analysis. Even if AI algorithms can become auxiliary tools, they should not take the clinician's place as the decision-maker.

Keywords: Artificial intelligence, spinal surgery, large language model

INTRODUCTION

In today's modern age, the need for instant and accessible information has increased exponentially across all areas, including the healthcare. This need encompasses not only the patient but also the healthcare professionals who, even with extensive training and higher-level expertise, often require up-to-date information to support clinical decision-making.

One of the most complex and risky fields in the realm of medicine is represented by spinal surgery (SS), being an area to which such technological support would be beneficial given the complexity of its clinical problems and the high risk of complications.

SS is marked by its application in anatomically critical regions, long operation times, complex postoperative care, increased morbidity and mortality, and the risk of extensive rehabilitation if there are complications-factors serving to significantly increase medicolegal risk. Therefore, SS requires support at the logistical level. Under such circumstances, the use of artificial intelligence (AI)-based tools through healthcare settings has emerged as an attractive strategy for the improvement of decision support and the enhancement of patient safety.

Navigation systems, computer programs for pedicle screw insertion, advanced radiological evaluation devices, and neurological monitoring systems, are now being used intraoperatively during spinal surgical procedures, helping to reduce surgical risks[1]. It has also been proposed that AI systems may provide benefits in diagnostic processes, prognostic analyses, and treatment planning[2,3]. Beyond their present uses during surgery, AI also has the potential to improve preoperative risk evaluations as well as standard and complicated postoperative management.

AI is a set of technologies that mimic the cognitive processes of humans, such as thought processes, learning, and problem-solving. One subset of AI, large language models, is designed specifically to understand natural language and absorb information from varied sources like scientific papers, books,

Address for Correspondence: Yiğit Kültür, Yeni Yüzyıl University Gaziosmanpaşa Hospital, Department of Ortopaedics and Traumatology, İstanbul, Türkiye
E-mail: yigitkulturr@hotmail.com
ORCID ID: orcid.org/0000-0001-8201-6994
Received: 06.06.2025 **Accepted:** 26.06.2025 **Publication Date:** 08.07.2025
Cite this article as: Demir MT, Kültür Y. A comparative study of orthopedic surgeons and AI models in the clinical evaluation of spinal surgery. J Turk Spinal Surg. 2025;36(3):125-129

research journals, and online data. Chat Generative Pre-trained Transformer (ChatGPT) is a well-known application of large language model technology. Due to its multimodal architecture, ChatGPT-4o can perform case-based analysis in the medical field and demonstrates remarkable expertise in critical thinking, literature synthesis, and clinical evaluation. Its use is particularly relevant in SS, due to its strengths in analyzing clinical cases, appraising patient trends, making academic evaluations, and interpreting images[1]. However, the use of this application is by subscription[4].

DeepSeek is another widely used AI model that is open-source; however, it does not have the function of processing image input[5]. It is claimed that this model is superior to ChatGPT in analyzing long medical papers, patient histories, and clinical studies[6]. It is also suggested that DeepSeek better follows the advancements in medical literature more dynamically and flexible[6,7]. The latest version, DeepSeek-V3, also offers offline functionality, thus enhancing data confidentiality[6]. On the other hand Bhattacharya et al.[8] reported that ChatGPT is superior in aspects of literature synthesis, clinical reasoning, medical education, and patient communication, DeepSeek is stronger in areas of surgical education, skill acquisition, patient teaching, and preoperative planning. Therefore, these two models play complementary roles.

In December 2023, the release of Google's Gemini model arrived with claims of improved reasoning capabilities as well as increased ability to handle complex tasks; however, their use in clinical settings remained somewhat constrained[9]. Nevertheless, Gemini has been suggested to be used as an adjunct to clinical decision-making processes[10-12]. With the growing debate over the use of AI to replace humans, it is important to consider the efficiency with which the models can read academic literature, understand it, and derive accurate conclusions in the field of medicine. This study compares the performances of orthopedic surgeons with three AI models-ChatGPT-4o, DeepSeek-V3, and Gemini Pro-in their accuracy for clinical decision-making scenarios and their theoretical knowledge capacity. The main focus is to examine the efficacy of AI systems within the context of preoperative patient evaluation and identify their reliability and efficiency compared to their human clinician counterparts across clinical decision-making scenarios.

MATERIALS AND METHODS

As the differences between the AI models carried less significance in the purview of this study, with the foremost aim being to identify the disparity in performance between AI and humans, ChatGPT-4o, DeepSeek-V3, and Gemini Pro were grouped as Group 1. On the other hand, ten orthopedic and traumatology surgeons with a minimum of 10 years of clinical experience were grouped as Group 2. The current study doesn't need to have authorization from an ethics committee because it doesn't involve patient interventions, procedural interventions, or the obtaining of personal health information.

In order to design the study question, two senior orthopedic surgeons formulated 50 study-type questions exclusively based on SS. Of these, 25 were case-based questions requiring clinical judgement, and the other 25 were on knowledge-based questions requiring theoretical knowledge. The question content breaks up as follows: 4 on anatomy, 12 on trauma, 4 on tumors, 4 on infections, 8 on postoperative surgical complications, 3 on physical examination, 7 on deformities, 5 on degenerative spine disease, and 3 on congenital spinal diseases. Since the DeepSeek model is unable to process images inputs, visual material or radiologic images were excluded from the questions developed for this study. The multiple choice questions were e-mailed to ten orthopedic and traumatology surgeons, instructing them to spend exactly 1 minute per question and record their answers. The answers were reviewed by the same surgeons who had formulated the questions. Concurrently, the same set of questions was administered to the three AI models, and their outputs were documented for subsequent analysis (Table 1). Statistical significance between the two groups was calculated by the Mann-Whitney U test. The correct answers rendered by AI models in the case-based and knowledge-based question sets were proportionally compared with those of the surgeons' group.

Statistical Analysis

In the current study, the evaluation results of the AI models-ChatGPT, GEM, and DeepSeek-were compared with those of ten orthopedic surgeons. The three AI models were placed in a single group, and the ten surgeons were placed in another group. The number of correct answers was taken both in absolute terms and in percentages. To compare the two groups, the Mann-Whitney U test, a non-parametric statistical method, was used. The reason for the choice of this specific test was the small sample sizes and the predicted non-normal distribution of the data, as it is an appropriate and stringent method for the comparison of two independent groups. The count of correct answers for each participant was counted separately for the clinical case-based questions (the first 25 questions), the fact-dependent knowledge-based questions (the last 25 questions), and the total of 50 questions. Independent Mann-Whitney U tests were performed for each of the above three categories. A p-value of less than 0.05 is considered statistically significant. All the analysis steps were performed using the SciPy package in the Python programming language.

RESULTS

In the case-based questioning analysis, Group 2 performed best compared to all other groups, with an overall accuracy of 88.8%. In Group 1, DeepSeek-V3 was found to be the best-performing model with an accuracy of 44%, which is half the rate of the

Table 1. The number of correct responses generated by artificial intelligence systems for case-based and knowledge-based questions

	ChatGP-4.o (n)	DeepSeek V3 (n)	Gemini Pro (n)
Case-based questions (n=25)	10	11	7
Knowledge-based questions (n=25)	19	20	16
Total (n=50)	29	31	23

ChatGPT: Chat Generative Pre-trained Transformer

Table 2. The individual and overall average accuracy rates of artificial intelligence models (Group 1), and the average correct response rate of orthopedic surgeons (Group 2)

Group	Case-based question accuracy rate (%)	Knowledge-based question accuracy rate (%)	Overall accuracy rate (%)
ChatGPT-4o	40.0	76.0	58.0
Gemini Pro	28.0	64.0	46.0
DeepSeek-V3	44.0	80.0	62.0
Group 1	37.3	73.3	55.3
Group 2	88.8	72.0	80.4

ChatGPT: Chat Generative Pre-trained Transformer

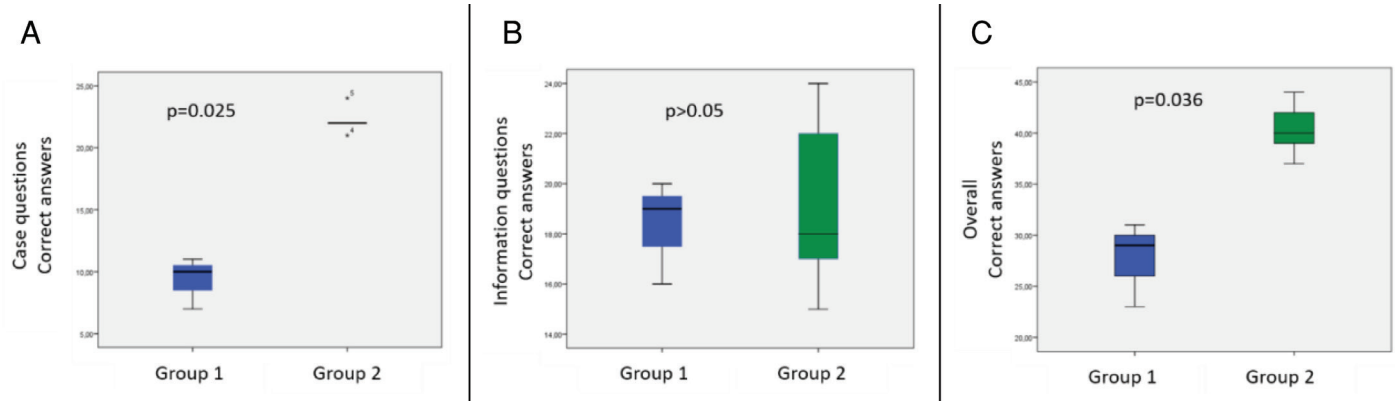

Figure 1. Comparison of correct answer numbers of (A) case-based questions, (B) knowledge-based questions and (C) overall between Group 1 (artificial intelligent) and Group 2 (orthopedic surgeons). Blue and green boxes represent Group 1 and Group 2, respectively

surgeons. ChatGPT-4o and Gemini Pro had accuracy rates of 40% and 28%, respectively (Table 2).

In the knowledge-based questions, DeepSeek-V3 had an accuracy of 80%, ChatGPT-4o demonstrated an accuracy of 76%, and Gemini Pro registered at 64%. On the other hand, Group 2 averaged 72%. As far as the overall performance is concerned, the AI models were again exceed by the Group 2 team who had an overall average score of 80.4%. Out of the AI models tested, the highest score was achieved by DeepSeek-V3 at 62.0%, followed by ChatGPT-4o with 58.0%, and then Gemini Pro with 46.0%. The overall success rate of 55.3% for Group 1 was calculated.

The Mann-Whitney U test was utilized to analyze statistically the test results from Group 1 and Group 2. The percentage of correct answers to case-based queries across Group 2 demostrated significantly higher performance compared to Group 1 (p=0.025). On the other hand, no statistical difference between the two groups was observed pertaining to knowledge-based questions (p=1.000). With respect to the total number of correct answers across the test, Group 2 revealed significantly improved performance compared to Group 1 (p=0.036) (Figure 1).

DISCUSSION

The growing use of AI models by healthcare professionals and patients has seen numerous clinical assessments on the potential applications and limitations of the technologies across the healthcare area, as seen through the numerous clinical studies[1-3,5,6,8,11,12]. The performances of ChatGPT-3.5 and ChatGPT-4o on the United States Medical Licensing Examination have been compared, indicating that the two

models passed the examinations, specifically clinical decision areas[13]. Another study, on the other hand, compared the diagnostic skill of ChatGPT to those of healthcare professionals and demonstrated that ChatGPT to have a limited understanding of examination questions[14]. Another study, on the use of wrist radiographs, tested the performances of ChatGPT-4o, Gemini 1.5, and DeepSeek-V3, with all failing to be identified as being useful clinical decision support systems[15]. A seperate study stated that diagnostic processes and systematic reviews would be aided using AI, postulating that tools such as ChatGPT and Gemini would become useful adjuncts to the clinical practice, but should not be entrusted to independently guide decision-making[12].

A comparison between ChatGPT, Gemini, and DeepSeek revealed stark differences in their performance in advanced situations requiring clinical judgement, highlighting the premise that such models should only exist as auxiliary tools and not the primary decision-maker[11]. In an orthopedic board test, AI models performed better on test items where analytical reasoning is not required[16].

This study entailed presenting 50 questions, which specific to SS, to ChatGPT-4o, Gemini Pro, DeepSeek-V3, and a group of ten experienced clinicians. The major objective was to compare the reasoning of AI models with that of human clinicians in situations requiring clinical judgement. The findings of this study stated that human clinicians perform superior than AI systems in the decision-making aspect when it comes to realistic case-based scenarios; however, AI models can perform as well as clinicians in situations entailing knowledge-based testing. These findings imply that while AI technologies can have some value in performing data-dependent tasks, they are largely insufficient to replace human expertise in clinical problem-solving and judgement on a case-by-case basis.

Although the AI system has shown proficiency in diagnostic and knowledge-related performance, it has yet to achieve the level of reliability needed for autonomous use in clinical decision-making. This study supports the current trend and emphasizes the importance of using AI technologies as supporting tools for healthcare professionals, not substitutes for them as main primary decision-makers. A wide range of clinical studies that compared various AI models have shown varied results[5,10,15]. In this analysis, overall accuracy percentages for the 50 integrated case-based and knowledge-based questions were 62% for DeepSeek-V3, 58% for ChatGPT-4o, and 46% for Gemini Pro. In a study focused on musculoskeletal radiology, ChatGPT proved to be more accurate compared to DeepSeek[17]. On the other hand, another study reported that DeepSeek provided more understandable replies compared to ChatGPT, credited to its high reasoning ability[18]. In this study, the results indicate that the DeepSeek models have higher overall accuracy, while ChatGPT-4o has similar performance for case-based and knowledge-based questions. However, the Gemini models performed generally worse.

The problems of verifiability and accountability of information created through the use of AI remain controversial topics. AI models utilize datasets limited to publicly available information up to a specified date. This constraint naturally raises the prospect of ignoring the newest literature and developments in the field of medicine. In one study analyzing different questions over time, it was noted that the accuracy of ChatGPT declined as the recency of the question improved[14]. These results suggest that the accuracy of the AI technologies may change in time and may not always match the current medical information. This finding shows that the accuracy of AI programs is time-dependent, indicating that they may not always have the most updated medical information. Because of this, our study sought to analyze the up-to-date validity of the AI programs by creating new test items and presenting them to the AI programs for preliminary testing.

An important limitation of the use of AI is the fact that the provided information often has no corroboration from credible scientific sources. Empirical research has shown that many of the references provided by ChatGPT-4o are scientifically unreliable, and DeepSeek-V3 has been shown to generate fake citations[19]. This fact makes the AI technology used in clinical decision support unreliable, thus posing great risks to patient safety[5,20]. Decisions from AI systems can lead to incorrect conclusions or late treatment, which may have great medical and legal consequences. Additionally, the lack of accountability of AI models represents a great shortcoming with regards to safety and responsibility in healthcare service provision[4,21]. For this reason, it is critical that AI systems are used only as auxiliary devices having human governance, the final decision authority resting entirely with the clinician[22,23].

Many studies on the application of AI to the field of SS havehighlighted the future potential of AI algorithms to become useful tools for preoperative planning and intraoperative assistance[22,23]. There is evidence showing ChatGPT is 68% successful at generating appropriate ideas relevant to spine surgery[23]. Additionally, it has been suggested that AI can represent an ideal asset for the development of educational resources, the simulation of complex clinical scenarios, the construction of personalized learning paths for medical students, and postoperative patient surveillance[6,22-25]. Given the relatively high complication rates of SS during intra-and postoperative periods compared to other surgical fields, this field requires strong technological support and an acceptance of new methodologies. The current study suggests further advancement of the AI technologies used in SS to position them among trustworthy auxiliary resources for healthcare professionals.

Study Limitations

A key limitation of the current study is the inability of DeepSeek to read images. Thus, radiology-and visually based assessments-that are critical to SS-cannot be examined. Additionally, the study only had 25 clinical cases, and this would limit the

generalization of the findings. The study used answers from ten expert orthopedic surgeons; inclusion of more clinicians would enhance both the reliability and the generalizability of the results. Further studies that involve more clinical involvement and large question sets are needed to obtain a more reliable evaluation.

CONCLUSION

The traditional view of medicine as an art emphasizes the role of numerous individual factors such as sociocultural context, cognitive capacity, medical history, and individual circumstances to bring the healing to fruition. Based on this model, it is technologically impossible for AI programs to fully understand the various human factors and generate context-relevant recommendations. Rather than viewing the technologies of AI as autonomous decision makers, it is more fitting to think of such applications as clinical practice-assisting instruments, tools for immediate access to relevant information, and reinforcement of decision support systems for diagnosis and therapy. These technologies should be envisioned as supportive tools to complement clinical decision-making and not to replace healthcare professionals; they are supportive factors strengthening clinical judgment. AI technologies have proved their utility for knowledge-based tasks but are dramatically inferior to clinicians for areas requiring clinical judgement and case analysis.

Ethics

Ethics Committee Approval- Informed Consent: The current study does not require ethics committee and informed consent because it does not involve patient interventions, procedural interventions, or the acquisition of personal health information.

Footnotes

Authorship Contributions

Surgical and Medical Practices: M.T.D., Y.K., Concept: M.T.D., Y.K., Design: M.T.D., Y.K., Data Collection or Processing: M.T.D., Analysis or Interpretation: M.T.D., Y.K., Literature Search: M.T.D., Y.K., Writing: M.T.D., Y.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Charles YP, Lamas V, Ntilikina Y. Artificial intelligence and treatment algorithms in spine surgery. Orthop Traumatol Surg Res. 2023;109:103456.
2. Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. J Dent Res. 2021;100:232-44.
3. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. Cancer Lett. 2020;481:55-62.
4. OpenAI. Terms of use. Last Accessed date: 30.06.2025. Available from: https://openai.com/policies/terms-of-use/
5. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. 2025;25:638.
6. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus. 2025;17:e79221.
7. Xu S, Hua W, Zhang Y. OpenP5: an open-source platform for developing, training, and evaluating LLM- based recommender systems. Conference Paper: SIGIR 2024: The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.
8. Bhattacharya K, Bhattacharya S, Bhattacharya N, Bhattacharya N. DeepSeek versus ChatGPT in surgical practice. Indian J Surg. 2025.
9. Gemini. Last Accessed date: 30.06.2025. Available from: https://gemini.google.com
10. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. Cureus. 2024;16:e57795.
11. Seth I, Marcaccini G, Lim K, Castrechini M, Cuomo R, Ng SK, et al. Management of Dupuytren's disease: a multi-centric comparative analysis between experienced hand surgeons versus artificial intelligence. Diagnostics (Basel). 2025;15:587.
12. Wong CR, Zhu A, Baltzer HL. The accuracy of artificial intelligence models in hand/wrist fracture and dislocation diagnosis: a systematic review and meta-analysis. JBJS Rev. 2024;12.
13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
14. Yigitbay A. Evaluation of ChatGPT's performance in the Turkish board of orthopaedic surgery examination. Med Bull Haseki. 2024;62:243-9.
15. Marcaccini G, Seth I, Xie Y, Susini P, Pozzi M, Cuomo R, Rozen WM. Breaking bones, Breaking Barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. J Clin Med. 2025;14:1983.
16. Karapınar SE, Dinçer R, Coşkun HS, Kaya Ö. Who is more successful in a spınal surgery examination? Chatgpt-3.5/4.0 or a resident doctor? J Turk Spinal Surg. 2025;36:88-91.
17. Uldin H, Saran S, Gandikota G, Iyengar KP, Vaishya R, Parmar Y, et al. A comparison of performance of DeepSeek-R1 model-generated responses to musculoskeletal radiology queries against ChatGPT-4 and ChatGPT-4o - a feasibility study. Clin Imaging. 2025;123:110506.
18. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025;198:105871.
19. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access. 2023;8:e23.00056.
20. Itchhaporia D. Artificial intelligence in cardiology. Trends Cardiovasc Med. 2022;32:34-41.
21. DeepSeek Privacy Policy. Last Accessed date: 30.06.2025. Available from: https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html
22. Nasirov R. The role of Claude 3.5 Sonet and ChatGPT-4 in posterior cervical fusion patient guidance. World Neurosurg. 2025;197:123889.
23. Herzog I, Mendiratta D, Para A, Berg A, Kaushal N, Vives M. Assessing the potential role of ChatGPT in spine surgery research. J Exp Orthop. 2024;11:e12057.
24. Kalanjiyam GP, Chandramohan T, Raman M, Kalyanasundaram H. Artificial intelligence: a new cutting-edge tool in spine surgery. Asian Spine J. 2024;18:458-71.
25. Uz C, Umay E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis. 2023;26:1343-9.