REVIEW ARTICLE

Journal of Pre-Clinical and Clinical Research
www.jpccr.eu

Are large language models capable of generating safe and guideline-concordant rehabilitation recommendations in orthopaedics and sports medicine? A review

Bartosz Gołembiewski[1,2,A-D]✉, Grzegorz Mulski[1,2,A-D], Martyna Manicka[1,2,A-D], Michał Baranowicz[1,2,B-D], Weronika Mazurkiewicz[1,2,B-D], Zuzanna Borecka[1,2,B-D], Agnieszka Sobczak[1,2,B-D], Alicja Szymczak[1,2,B-D], Anna Jaworowicz[1,2,B-D], Joanna Piasecka[1,2,B-D], Łukasz Łapaj[1,E-F]

[1] *Department of General Orthopedics, Musculoskeletal Oncology and Trauma Surgery, University of Medical Sciences, Poznań, Poland*
[2] *Student Scientific Society of Orthopaedics and Musculoskeletal Traumatology, University of Medical Sciences, Poznań, Poland*
A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E – Critical revision of the article, F – Final approval of the article

Gołembiewski B, Mulski G, Manicka M, Baranowicz M, Mazurkiewicz W, Borecka Z, Sobczak A, Szymczak A, Jaworowicz A, Piasecka J, Łapaj Ł. Are large language models capable of generating safe and guideline-concordant rehabilitation recommendations in orthopaedics and sports medicine? A review. J Pre-Clin Clin Res. doi:10.26444/jpccr/222263

Abstract

Introduction and Objective. Large language models (LLMs) are being increasingly used in medicine, including orthopedics and sports medicine, where they may support the development of rehabilitation recommendations and making decisions on return to physical activity. However, concerns remain regarding safety and adherence to clinical guidelines. The aim of the review is to summarize the available evidence on the quality of LLM-generated recommendations, their concordance with current guidelines, and their potential clinical implications.

Review Methods. A narrative review of the literature was conducted using the MEDLINE (PubMed) and Scopus databases, covering the period from November 2022 (after the widespread introduction of models such as ChatGPT) to 1 March 2026. English-language studies evaluating LLM-generated rehabilitation recommendations, their concordance with clinical guidelines, and safety aspects were included.

Brief description of the state of knowledge. LLMs can generate coherent and often useful rehabilitation recommendations, although their quality is variable. Many studies report only partial concordance with clinical guidelines, with key elements – such as exercise parameters or progression criteria – frequently omitted. Another important limitation is the sensitivity of responses to prompt formulation. While the recommendations are generally reasonable, they often require specialist verification before their use in clinical practice.

Summary. LLMs may serve as a valuable tool to support rehabilitation, particularly in patient education and treatment planning; however, they should not be considered a standalone source of clinical recommendations. Their use requires specialist oversight and further validation in clinical studies.

Key words

large language models, Artificial Intelligence, orthopaedics, rehabilitation, sports medicine

INTRODUCTION

In recent years, artificial intelligence has gained rapid traction in clinical medicine. Since the public release of generative models in late 2022, large language models have attracted growing attention for their potential role in diagnostics, patient education, and clinical decision support [1]. Their applicability in planning therapeutic treatment is also being analyzed with increasing frequency. In orthopedics and sports medicine, the abilty of large language models (LLMs) to create rehabilitation recommendations and advice on the possibility of returning to sports activities, is a subject of particular interest.

Rehabilitation is integral to the management of musculoskeletal injuries and disorders. Such cases represent one of the leading causes of reduced physical activity, with considerable clinical and economic burdens [2, 3]. In conditions such as anterior cruciate ligament (ACL) rupture, rotator cuff injury, or Achilles tendon rupture, the structure and execution of postoperative rehabilitation influences not only recovery time, but also the risk of reinjury re-curring injury and long-term functional outcomes [4, 5]. Contemporary standards of care are grounded in evidence-based medicine (EBM) and clinical practice guidelines (CPGs), which outline criteria for exercise progression, functional assessment, and safe return-to-sport (RTS) decision-making [4–6], often incorporating both objective and patient-reported outcome measures [7].

✉ Address for correspondence: Bartosz Gołembiewski, Department of General Orthopaedics, Musculoskeletal Oncology and Trauma surgery, University of Medical Sciences, Poznań, Poland.
E-mail: bartekgolembiewski@gmail.com

Received: 31.03.2026; accepted: 20.05.2026; first published: 27.05.2026

In this context, an important question is whether LLMs can generate rehabilitation recommendations that adhere to current standards and whether their use can be considered safe for clinical application. Previous studies suggest that models such as ChatGPT-4 can produce comprehensive rehabilitation programmes that incorporate therapeutic goals and elements of the International Classification of Functioning, Disability and Health (ICF); however, these outputs are not free from errors and over-simplifications [8]. In studies assessing knee osteoarthritis rehabilitation planning, GPT-4o and Gemini achieved 70–74% concordance with expert consensus – the primary limitations involved inadequate exercise parameter specification and poorly defined progression criteria [9]. Other investigations have demonstrated substantial variability in AI-generated recommendations depending on prompt formulation, raising concerns regarding their reproducibility, reliability, and stability in clinical practice [10].

Although systematic reviews on the use of LLMs in orthopedic surgery have already been published [11], studies specifically examining the generation of rehabilitation recommendations, their adherence to clinical guidelines, and their potential safety implications in sports medicine, remain limited. Given the rapid development of this technology and the increasing number of publications referring to reporting standards for AI-based research (e.g., CONSORT-AI, SPIRIT-AI) [12], the available evidence requires careful review and critical evaluation.

OBJECTIVE

The aim of this narrative review is to discuss current research on rehabilitation recommendations and exercise programmes created by large language models (LLMs) in orthopaedics and sports medicine. Particular attention is paid to assessing the extent to which the recommendations they propose are consistent with current clinical guidelines, as well as identifying potential risks arising from their use in clinical practice, especially in the context of decisions about a patient's return to sports activities.

MATERIALS AND METHOD

This narrative review of the literature discusses the use of large language models (LLMs) in generating rehabilitation recommendations and exercise programs in orthopaedics and sports medicine. A literature search was conducted in the MEDLINE (PubMed) and Scopus databases. The time frame covered the period from November 2022 (from the moment generative language models were made publicly available), to the date of the final literature search (1 March 2026). Only publications written in English were included.

The search strategy was based on a combination of key words related to LLM technology, rehabilitation, orthopaedics and sports medicine, including: 'large language model', "'LM', 'generative AI', 'ChatGPT', 'GPT-4', 'GPT-4o', 'Gemini', 'rehabilitation', 'physical therapy', 'exercise prescription', 'exercise programme', 'return to sport', 'sports medicine', and 'musculoskeletal'. Individual terms were combined using the boolean operators AND/OR.

The review included studies analyzing rehabilitation or postoperative management recommendations generated by LLM in a clinical context, as well as publications assessing their compliance with guidelines or safety aspects. The final set of literature included narrative reviews, systematic reviews and meta analyses addressing the use of large language models in healthcare, as well as original observational studies and exploratory analyses evaluating rehabilitation or exercise recommendations generated by LLM. In addition, selected foundational publications relevant to rehabilitation, return-to-sport criteria and musculoskeletal clinical guidelines were included when they provided important contextual background. The analysis excluded single case reports, conference abstracts, editorials, commentaries, letters to the editor, and non peer-reviewed publications. Studies focusing exclusively on other applications of artificial intelligence in medicine, such as diagnostic imaging or administrative tasks, without assessment of rehabilitation or exercise recommendations generated by LLM were also excluded.

The selection process involved an initial screening of titles and abstracts followed by full-text assessment of potentially relevant articles. Reference lists of included publications were also reviewed to identify additional relevant studies. A total of 8 studies met the inclusion criteria and were included in the final analysis.

As this study is a narrative review, no formal protocol registration, standardized risk-of-bias assessment, or independent duplicate screening procedures were applied. Due to the heterogeneity of study designs, evaluated models, and outcome measures, a meta-analysis was not performed. The results are presented in a descriptive form, focusing on their clinical significance and compliance with current guidelines.

RESULTS

The summarized characteristics of studies analyzing the use of large language models (LLMs) in generating rehabilitation recommendations in orthopaedics, rehabilitation and physical therapy are presented in Table 1. The available literature on the use of LLMs to generate rehabilitation recommendations in orthopaedics and physical therapy remains limited because they focus primarily on 3 topics: assessing the quality of recommendations generated by LLM, their compliance with clinical guidelines, and their potential use in clinical practice.

Gürses et al. analyzed rehabilitation programmes for patients with knee osteoarthritis generated by language models and found that LLM-generated recommendations showed substantial agreement with programmes developed by physical therapists, although they frequently lacked detailed specification of exercise parameters, such as intensity, number of repetitions, or load progression criteria [9]. Similarly, Mykhalko et al., evaluating individualized rehabilitation plans generated by ChatGPT-4o, concluded that although most proposed therapeutic strategies were considered potentially useful, specialist verification is required before their implementation in clinical practice [14].

Another important area of research involves evaluating whether recommendations generated by language models comply with current clinical guidelines. Safran et al., analyzing ChatGPT responses in musculoskeletal rehabilitation,

Table 1. Summary of of studies analyzing the use of large language models (LLMs) in generating rehabilitation recommendations in orthopedics, rehabilitation and physical therapy

Author, year	Clinical context	Anylized LLM(s)	Study design	Notable results
Gürses et al. [9], 2025	Rehabilitation in patients with knee osteoarthritis	ChatGPT-4o, Gemini	Observational study; comparison of LLM-generated and physiotherapist-designed rehabilitation programs	Moderate agreement between LLM-generated programs and expert consensus; most common shortcomings involved insufficiently detailed exercise parameters
Sawamura et al. [13], 2024	Answering clinical questions in physical therapy	ChatGPT	Cross-sectional study of model responses to guideline-based clinical questions in physiotherapy	The model generated mostly correct responses; however, explicit references to clinical guidelines and sources were often lacking
Mykhalko et al. [14], 2025	Individualized rehabilitation plans for comorbid patients (including musculoskeletal disorders)	ChatGPT-4o	Expert evaluation of LLM-generated rehabilitation plans	Most LLM-generated plans were considered applicable after review and modification; specialist supervision was recommended
Safran et al. [15], 2025	Orthopaedic rehabilitation	ChatGPT-4o, Gemini 2.5 Pro, DeepSeek-V3	Comparative analysis of model performance in clinical reasoning and treatment planning	Response quality differed between models
Safran et al. [16], 2025	Musculoskeletal rehabilitation	ChatGPT-4	Cross-sectional study evaluating guideline concordance of LLM responses	Partial concordance with clinical guidelines; many responses failed to meet all guideline criteria and lacked sufficient detail. The authors highlighted the role of LLMs as a supplementary decision-making tool.
Kim J. [17], 2025	Musculoskeletal physiotherapy	ChatGPT, DeepSeek	Comparative analysis of model response quality	ChatGPT generated more detailed responses; both models showed limitations in clinical justification and require further validation.
Gianola et al. [18], 2024	Rehabilitation and conservative treatment of lumbosacral radicular pain	ChatGPT-3.5	Cross-sectional study; comparison of model responses with clinical guidelines	Some responses were consistent with clinical guidelines; however, many lacked full guideline concordance and internal consistency
Hao et al. [19], 2025	Musculoskeletal physiotherapy	ChatGPT-4	Assessment of the model as a clinical decision-support tool	LLMs may support clinical decision-making but should not replace specialist judgment

reported that while some recommendations were consistent with current guidelines, the models frequently omitted key elements of therapeutic management or failed to take into account the full range of guideline criteria [16]. Similar findings were reported by Gianola et al., who compared ChatGPT-generated responses with clinical guidelines for the management of lumbosacral radicular syndrome. While the model was able to outline general therapeutic principles, its recommendations were not always fully aligned with current guidelines and depended on the prompt formulation [18].

Several studies have evaluated the quality of responses generated by LLMs. Sawamura et al., in their analysis of clinical questions in the field of physical therapy, showed that ChatGPT was able to generate substantively correct answers consistent with general physical therapy recommendations; the responses also frequently lacked precise references to scientific literature or current clinical guidelines [13]. Similar findings have been reported in studies comparing different language models. Safran et al. reported in their analysis, notable differences between models in both response quality and the justification of therapeutic recommendations [15]. Kim et al., in turn, noted that although LLMs are capable of generating coherent and logically structured responses, their recommendations are not always sufficiently detailed to serve as an independent basis for therapeutic decision-making, highlighting the need for validation of stand-alone models before their use in clinical practice [17]. In another study, Hao et al. analyzed the potential role of LLMs as a tool supporting clinical decision-making in musculoskeletal physiotherapy. They emphasized that language models may provide useful support in the analysis of clinical problems; however, their recommendations should be considered supportive-only, rather than a substitute for specialist evaluation [19], which is consistent with findings from previously described studies.

DISCUSSION

Guidelines for concordance and safety of LLM-Generated rehabilitation recommendations. Although LLMs are capable of generating coherent and often clinically appropriate recommendations, a key question concerns the extent to which these outputs align with current clinical guidelines [20]. In musculoskeletal rehabilitation, the formulation of recommendations requires consideration of several key elements, such as appropriate exercise parameters, load progression, return-to-activity criteria, and individualized rehabilitation planning [4,6]. Analyses of musculoskeletal rehabilitation have shown that some recommendations generated by LLMs are consistent with current clinical guidelines; however, the models frequently omit important elements of these guidelines and present recommendations in an overly general or incomplete manner [16,18]. Similar observations have been reported in studies evaluating language-model responses to clinical questions in physiotherapy in which many answers were generally consistent with therapeutic recommendations, but lacked precise references to scientific literature or specific guideline recommendations [13].

It is also noteworthy, that standardized rehabilitation protocols may sometimes require modification due to factors such as intraoperative changes in surgical technique or other surgery-related complications. For instance, although full weight-bearing is typically permitted after procedures such as arthroplasty, it may be limited due to intraoperative issues with implant fixation [21]. Moreover, comorbidities such as Parkinson's disease may influence postoperative rehabilitation, as affected patients are more prone to falls and gait disturbances [22, 23]. These situations are particularly challenging, as LLMs may not generate reliable recommendations in non-standard clinical scenarios.

An important aspect when evaluating the usefulness of LLMs in rehabilitation is the safety of the recommendations they generate. Several studies have emphasized that although language models can produce logical and coherent therapeutic strategies, their recommendations often require specialist verification, particularly regarding exercise dosing, load progression, and the consideration of individual patient characteristics [9, 14]. However these issues are not limited to rehabilitation in orthopaedics or physiotherapy. Studies examining the use of LLMs in other areas of medicine have shown that although they can generate linguistically and logically coherent responses, significant inconsistencies with clinical knowledge and guidelines have also been reported [14, 24, 25]. For this reason, most authors emphasize that LLM-generated recommendations should be primarily considered as a decision-support tools rather than autonomous sources of therapeutic guidance [16, 19, 26, 27].

Another issue when evaluating the usefulness of language models in rehabilitation planning is the stability of their responses and their reproducibility across different clinical scenarios. Studies examining the use of LLMs in medicine have repeatedly noted that the content of LLM-generated responses may vary depending on the formulation of the query, the amount of information provided in the prompt, or the specific model used [10, 18, 24–26]. This phenomenon is particularly relevant in rehabilitation, where treatment plans should rely on clearly defined clinical criteria and individual patient characteristics, and remain reproducible among patients with similar profiles to allow for standardized management. Analyses of clinical questions in physiotherapy and musculoskeletal rehabilitation have shown that although language models can generate logical and coherent therapeutic recommendations, their responses may differ in the level of detail and the scope of the proposed interventions [15, 17]. Similar observations have been reported in studies from other areas of medicine, suggesting that recommendations generated by LLMs may be sensitive to the context of the query but are not always fully reproducible when the same clinical problem is presented [28–30]. In practice, this suggests that recommendations generated by language models may be influenced by prompt formulation.

Clinical implications for orthopaedics and sports medicine. As discussed earlier, the potential applications of LLMs in orthopaedics and sports medicine primarily relate to their role as decision-support tools in clinical decision-making. In recent years, it has been increasingly suggested that such models may support the analysis of medical information, the generation of preliminary therapeutic recommendations, and the rapid retrieval and synthesis of current scientific knowledge [24, 25, 31]. In orthopaedic surgery, artificial intelligence–based technologies are already being used in areas such as medical imaging analysis, prediction of treatment outcomes, and surgical planning, highlighting the growing role of digital tools in clinical practice [32]. Here, LLMs may constitute an additional tool supporting clinicians in everyday clinical practice.

In orthopaedic rehabilitation, language models may be particularly useful in the planning of the therapy and patient education. Current literature suggests that LLMs may support the preparation of educational materials and help explain therapeutic recommendations in a way that is more understandable to patients, potentially improving adherence to rehabilitation recommendations [33]. It has also been emphasized that LLM-based tools may facilitate communication between clinicians and patients, for example, by simplifying complex medical information and adapting educational content to the patient's level of understanding [30,34]. In the context of rehabilitation, this may include explaining the principles of performing therapeutic exercises, the course of the rehabilitation process, and the rationale and importance of a gradual return to physical activity.

At the same time, the potential benefits of LLMs in orthopaedic rehabilitation are closely tied to their responsible and safe use in clinical practice. As discussed earlier, numerous studies emphasize that although language models can generate logical and convincing responses, they are not always fully consistent with current medical knowledge or clinical guidelines [18, 35]. An additional concern remains the phenomenon of so-called model hallucinations, in which models generate convincing but inaccurate information, which in the clinical context may have important implications for patient safety [36].

Another important area of potential LLM applications is medical education and support for the training of healthcare professionals, including physicians and physiotherapists. Language models may be used to analyze clinical cases, generate educational scenarios, and explain complex concepts, which may support the educational process [25]. In recent years, reports have also emerged suggesting that LLMs may serve as tools supporting the development of clinical reasoning by simulating medical cases and enabling interactive discussion of potential diagnostic and therapeutic strategies [30]. In terms of specialist training, these tools may facilitate access to current medical knowledge and support continuing professional development, which plays an important role in orthopaedic and rehabilitation practice [37].

Ethical, regulatory and implementation considerations. As the use of LLMs in medicine expands, increasing attention is also being paid to the ethical, legal, and regulatory issues associated with their implementation in healthcare. One of the most controversial and widely discussed issues relates to the reliability of AI-generated content and the legal responsibility for decisions made on the basis of recommendations generated by AI-based systems. The literature emphasizes that although LLMs may serve as useful clinical decision-support tools, their use requires supervision by qualified specialists and a clearly defined allocation of responsibility for potential errors [38, 39].

Another important concern is the limited transparency of LLMs and the difficulty in assessing the basis on which their responses are generated. Unlike many traditional clinical decision-support tools, LLMs operate in a manner that does not always allow tracing the underlying basis of a given recommendation [39–41], which may significantly complicate the evaluation of the reliability of generated responses and the identification of potential errors. For this reason, increasing attention is being paid to the development of more transparent standards for evaluating artificial intelligence–based tools and clear guidelines for their use in clinical practice [38–40].

Equally important are issues related to the regulation of such tools and their integration into healthcare systems. International organizations, including the World Health Organization (WHO) and regulatory authorities overseeing

medical technologies, highlight the need to establish clear legal frameworks for the use of artificial intelligence in medicine, addressing issues such as data security, algorithm transparency, and responsibility for clinical decision-making [42, 43]. In the context of rehabilitation, the implementation of LLM-based tools should be preceded by appropriate clinical validation and careful evaluation of their safety in clinical practice.

Finally, given the limitations of LLMs, their rapidly increasing use, and concerns regarding patient safety, developers of such models should take particular care to ensure that algorithms are designed in a way that clearly communicates that the information provided in response to rehabilitation-related queries does not constitute medical advice, and is intended for educational purposes only.

CONCLUSIONS

In summary, large language models show promise as supportive tools in orthopaedic and sports medicine rehabilitation, particularly in generating structured recommendations and assisting with patient education. However, current evidence indicates that their outputs are often incomplete, inconsistently aligned with clinical guidelines, and sensitive to prompt formulation. As a result, LLM-generated recommendations should not be used as standalone sources for therapeutic decision-making, but rather as additional tools requiring supervision of orthopaedic surgeon or rehabilitation specialist.

Future research should extend beyond response-based evaluations and focus on prospective, real-world clinical studies assesing their impact on rehabilitation outcomes, patient adherence, and return-to-sport decision-making. Standardized approaches to prompt design and response evaluation, along with validation of models in clinical settings, will be essential to determine their practical utility. It also appears essential to develop language models dedicated to medical applications, trained on validated clinical data and aligned with current clinical guidelines, as this could improve the reliability of generated recommendations, enhance guideline adherence, reduce the risk of errors and potentially unsafe outputs, and, importantly, strengthen the protection of sensitive patient data required for effective clinical use.

Acknowledgments. The authors would like to sincerely thank Assoc. Prof. Łukasz Łapaj, MD, PhD, for his invaluable support in the preparation and revision of the final manuscript. The review was conducted within the framework of the Student Scientific Society of Orthopaedics and Musculoskeletal Traumatology at the University pf Medical Sciences in Poznań, Poland.

Conflicts of interest. The authors declare no conflicts of interest.

Funding. The study did not receive any external funding.

REFERENCES

1. 1. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT-J, Li Y-CJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci. 2025;32:45. https://doi.org/10.1186/s12929-025-01131-z
2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. London, England. 2020;396:1204–22. https://doi.org/10.1016/S0140-6736(20)30925-9
3. Pop T, Szczygielska D, Druzbicki M. Epidemiology and cost of conservative treatment of patients with degenerative joint disease of the hip and knee. Ortop Traumatol Rehabil. 2007;9:405–12.
4. Grindem H, Snyder-Mackler L, Moksnes H, Engebretsen L, Risberg MA. Simple decision rules can reduce reinjury risk by 84% after ACL reconstruction: the Delaware-Oslo ACL cohort study. Br J Sports Med. 2016;50:804–8. https://doi.org/10.1136/bjsports-2016-096031
5. Kyritsis P, Bahr R, Landreau P, Miladi R, Witvrouw E. Likelihood of ACL graft rupture: not meeting six clinical discharge criteria before return to sport is associated with a four times greater risk of rupture. Br J Sports Med. 2016;50:946–51. https://doi.org/10.1136/bjsports-2015-095908
6. Ardern CL, Glasgow P, Schneiders A, Witvrouw E, Clarsen B, Cools A, et al. 2016 Consensus statement on return to sport from the First World Congress in Sports Physical Therapy, Bern. Br J Sports Med. 2016;50:853–64. https://doi.org/10.1136/bjsports-2016-096278
7. Chrzan D, Kusz D, Bołtuć W, Bryła A, Kusz B. Subjective assessment of rehabilitation protocol by patients after ACL reconstruction – preliminary report. Ortop Traumatol Rehabil. 2013;15:215–25. https://doi.org/10.5604/15093492.1058412
8. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. J Rehabil Med. 2023;55:jrm13373. https://doi.org/10.2340/jrm.v55.13373
9. Gürses ÖA, Özüdoğru A, Tuncay F, Kararti C. The Role of Artificial Intelligence Large Language Models in Personalized Rehabilitation Programs for Knee Osteoarthritis: An Observational Study. J Med Syst. 2025;49:73. https://doi.org/10.1007/s10916-025-02207-x
10. Yang Z, Zhang X, Li H, Ye J. More details, less variability? A crossover design study on the impact of information granularity on ChatGPT's training program stability. Biol Sport. Termedia. 2025;43:379–92. https://doi.org/10.5114/biolsport.2026.154148
11. Mo K, Lin R, Dunn E, Girgis G, Fang W, Walsh J, et al. Systematic Review on Large Language Models in Orthopaedic Surgery. J Clin Med. Multidisciplinary Digital Publishing Institute. 2025;14:5876. https://doi.org/10.3390/jcm14165876
12. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. British Medical Journal Publishing Group. 2020;370:m3164. https://doi.org/10.1136/bmj.m3164
13. Sawamura S, Bito T, Ando T, Masuda K, Kameyama S, Ishida H. Evaluation of the accuracy of ChatGPT's responses to and references for clinical questions in physical therapy. J Phys Ther Sci. 2024;36:234–9. https://doi.org/10.1589/jpts.36.234
14. Mykhalko Y, Dyditska S, Balatska L, Filak F, Rubtsova Y. AI-driven rehabilitation: evaluation of ChatGPT-4o for generating personalized physical rehabilitation plans in comorbid patients. Wiad Lek. Warsaw, Poland. 1960; 2025;78:753–9. https://doi.org/10.36740/WLek/203850
15. Safran E, Yaşasın Y. AI vs AI: clinical reasoning performance of language models in orthopedic rehabilitation. J Health Sci Med. MediHealth Academy Yayıncılık. 2025;8:825–31. https://doi.org/10.32322/jhsm.1743257
16. Safran E, Yildirim S. A cross-sectional study on ChatGPT's alignment with clinical practice guidelines in musculoskeletal rehabilitation. BMC Musculoskelet Disord. 2025;26:411. https://doi.org/10.1186/s12891-025-08650-8
17. Kim J. Comparing ChatGPT and DeepSeek for Generating Clinically Relevant Responses related to Physical Therapy. J Musculoskelet Sci Technol. Academy of KEMA. 2025;9:9–18. https://doi.org/10.29273/jmst.2025.9.1.9
18. Gianola S, Bargeri S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT Compared to Clinical Practice Guidelines in Making Informed Decisions for Lumbosacral Radicular Pain: A Cross-sectional Study. J Orthop Sports Phys Ther. 2024;54:222–8. https://doi.org/10.2519/jospt.2024.12151
19. Hao J, Yao Z, Tang Y, Remis A, Wu K, Yu X. Artificial Intelligence in Physical Therapy: Evaluating ChatGPT's Role in Clinical Decision

Support for Musculoskeletal Care. Ann Biomed Eng. 2025;53:9–13. https://doi.org/10.1007/s10439-025-03676-4
20. Lin M-J, Hsieh L-C, Chen C-K. Evaluating ChatGPT's Concordance with Clinical Guidelines of Ménière's Disease in Chinese. Diagnostics. Multidisciplinary Digital Publishing Institute. 2025;15:2006. https://doi.org/10.3390/diagnostics15162006
21. Mitchell O, Ward P, Petrov K. Weight-Bearing Status After Peri-Prosthetic Proximal Femur Fracture Open Reduction and Internal Fixation (ORIF) or Revision Arthroplasty: A Clinical Audit. Cureus. 2025;17:e90805. https://doi.org/10.7759/cureus.90805
22. Allen NE, Schwarzel AK, Canning CG. Recurrent falls in Parkinson's disease: a systematic review. Park Dis. 2013;2013:906274. https://doi.org/10.1155/2013/906274
23. Bloem BR, Hausdorff JM, Visser JE, Giladi N. Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena. Mov Disord Off J Mov Disord Soc. 2004;19:871–84. https://doi.org/10.1002/mds.20115
24. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47:33. https://doi.org/10.1007/s10916-023-01925-4
25. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
26. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5:e107–8. https://doi.org/10.1016/S2589-7500(23)00021-3
27. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230–43. https://doi.org/10.1136/svn-2017-000101
28. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6:120. https://doi.org/10.1038/s41746-023-00873-0
29. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv. 2023: n. pag. https://doi.org/10.48550/arXiv.2303.13375
30. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Nature Publishing Group. 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2
31. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. Npj Digit Med. Nature Publishing Group. 2025;8:263. https://doi.org/10.1038/s41746-025-01684-1
32. Han F, Huang X, Wang X, Chen Y, Lu C, Li S, et al. Artificial Intelligence in Orthopedic Surgery: Current Applications, Challenges, and Future Directions. MedComm. 2025;6:e70260. https://doi.org/10.1002/mco2.70260
33. 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. JMIR Publications Inc., Toronto, Canada. 2024;26:e59607. https://doi.org/10.2196/59607
34. Blease C, Bernstein MH, Gaab J, Kaptchuk TJ, Kossowsky J, Mandl KD, et al. Computerization and the future of primary care: A survey of general practitioners in the UK. PloS One. 2018;13:e0207418. https://doi.org/10.1371/journal.pone.0207418
35. Naqvi WM, Shaikh SZ, Mishra GV. Large language models in physical therapy: time to adapt and adept. Front Public Health. Frontiers. 2024;12:1364660. https://doi.org/10.3389/fpubh.2024.1364660
36. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. Npj Digit Med. Nature Publishing Group. 2025;8:274. https://doi.org/10.1038/s41746-025-01670-7
37. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. Multidisciplinary Digital Publishing Institute. 2023;11:887. https://doi.org/10.3390/healthcare11060887 (access: 2025.03.12)
38. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med. 1982. 2020;260:113172. https://doi.org/10.1016/j.socscimed.2020.113172
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Nature Publishing Group. 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7
40. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. Nature Publishing Group. 2019;1:206–215. https://doi.org/10.1038/s42256-019-0048-x
41. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. https://doi.org/10.1186/s12916-019-1426-2
42. European Commission, Directorate-General for Communications Networks, Content and Technology, High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Publications Office. 2019. https://data.europa.eu/doi/10.2759/346720 (access: 2025.12.12).
43. World Health Organization 2021. Ethics and governance of artificial intelligence for health. https://www.who.int/publications/i/item/9789240029200. (access: 2025.12.12).