Received: 05 Feb 2025 | **Accepted:** 18 Apr 2025 **Genel Tıp Derg. Volume** 35/**Issue** 3 (June), 487-492
DOI: 10.54005/geneltip.1634118

ORIGINAL ARTICLE

Artificial Intelligence Exercise Recommendations in Knee Osteoarthritis Rehabilitation: ChatGPT-4o and Gemini Advanced Example

Diz Osteoartriti Rehabilitasyonunda Yapay Zeka Egzersiz Önerileri: ChatGPT-4o ve Gemini Advanced Örneği

[1]Ömer Alperen Gürses, [1]Anıl Özüdoğru, [2]Figen Tuncay, [1]Caner Karartı

[1]School of Physical Therapy and Rehabilitation, Department of Physiotherapy and Rehabilitation, Kırşehir

[2]Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Kırşehir Ahi Evran University, 40100 Merkez, Kırşehir, Turkey

Correspondence

Ömer Alperen Gürses, PhD, School of Physiotherapy and Rehabilitation, Kırşehir Ahi Evran University, Kırşehir, Türkiye

E-Mail: omeralperengurses@gmail.com,

How to cite ?

Gürses ÖA, Özüdoğru A, Tuncay F, Karartı C. Artificial intelligence exercise recommendations in knee osteoarthritis rehabilitation: ChatGPT-4o and Gemini Advanced example. Genel Tıp Derg. 2025;35 (3): 487-492

Abstract

Aim: This study aimed to comparatively evaluate the propensity of the large language models ChatGPT-4o and Gemini Advanced to recommend personalized exercise based on patients' assessment data in knee osteoarthritis rehabilitation.
Methods: This observational study included 40 patients diagnosed with knee OA according to the American College of Rheumatology criteria. Demographic data, pain levels, range of motion, muscle strength, functional status, and balance were assessed using standardized clinical tests. ChatGPT-4o and Gemini Advanced generated three-phase rehabilitation programs based on these assessments. Exercise recommendations were analyzed across 12 parameters, and statistical comparisons were conducted using the Mann-Whitney U test and Spearman's correlation ($p<0.05$).
Results: ChatGPT-4o demonstrated statistically significant differences in 7 parameters: Phase 1 (quadriceps muscle strength, knee flexion angle, knee extension angle, and four-square step test; $p=0.017$, $p=0.012$, $p=0.033$, $p=0.043$), Phase 2 (quadriceps muscle strength and Lysholm scale; $p=0.032$, $p=0.040$), and Phase 3 (quadriceps muscle strength; $p=0.007$). In contrast, Gemini Advanced exhibited significant differences in only 2 parameters: Phase 1 (Lysholm scale score; $p=0.044$) and Phase 3 (quadriceps strengthening exercise; $p=0.047$). ChatGPT-4o appeared to integrate patient assessment data more effectively, but both models showed limitations in personalization.
Conclusions: While ChatGPT-4o and Gemini Advanced show potential for designing personalized knee OA rehabilitation programs, their recommendations remain constrained. Further improvements in dataset quality, real-time medical knowledge integration, and domain-specific training are needed to enhance their clinical utility.

Keywords: Artificial intelligence, ChatGPT, Gemini, Large language models, Physiotherapy, rehabilitation program, Knee osteoarthritis

ÖZ

Amaç: Bu çalışma, büyük dil modelleri ChatGPT-4o ve Gemini Advanced'in diz osteoartriti rehabilitasyonunda hastaların değerlendirme verilerine dayanarak kişiselleştirilmiş egzersiz önerme eğilimini karşılaştırmalı olarak değerlendirmeyi amaçlamıştır.
Gereç ve Yöntemler: Gözlemsel nitelikteki bu çalışmaya, Amerikan Romatoloji Koleji kriterlerine göre diz osteoartriti tanısı almış 40 hasta dahil edilmiştir. Demografik veriler, ağrı düzeyi, eklem hareket açıklığı, kas kuvveti ve fonksiyonel durum ve denge standart klinik testlerle değerlendirilmiştir. ChatGPT-4o ve Gemini Advanced, bu değerlendirmelere dayanarak üç fazdan oluşan rehabilitasyon programları oluşturmuştur. Egzersiz önerileri 12 parametre üzerinden analiz edilmiş, istatistiksel karşılaştırmalar Mann-Whitney U testi ve Spearman korelasyonu ile yapılmıştır ($p<0.05$).
Bulgular: Faz 1'de kuadriseps kas kuvveti, diz fleksiyon açısı, diz ekstansiyon açısı ve dört kare adım testi ($p=0.017$, $p=0.012$, $p=0.033$, $p=0.043$); Faz 2'de kuadriseps kas kuvveti ve Lysholm ölçeği ($p=0.032$, $p=0.040$); Faz 3'te ise kuadriseps kas kuvveti ($p=0.007$) olmak üzere ChatGPT-4o, 7 parametrede istatistiksel olarak anlamlı fark göstermiştir. Öte yandan, Gemini Advanced ise Faz 1'de Lysholm skoru ($p=0.044$) ve Faz 3'te kuadriseps güçlendirme egzersizi ($p=0.047$) ile yalnızca 2 parametrede istatistiksel olarak anlamlı fark göstermiştir. ChatGPT-4o'nun hasta değerlendirme verilerini daha etkin entegre ettiği görülmüştür ancak her iki modelin de kişiselleştirme konusunda sınırlılıkları mevcuttur.
Sonuçlar: ChatGPT-4o ve Gemini Advanced, kişiselleştirilmiş diz osteoartriti rehabilitasyon programları tasarlama potansiyeli taşısa da önerileri halen sınırlıdır. Klinik faydalarının artırılması için veri seti kalitesinin iyileştirilmesi, gerçek zamanlı tıbbi bilgi entegrasyonu ve alana özgü eğitimlerle desteklenmeleri gerekmektedir.

Anahtar Kelimeler: Yapay zeka, ChatGPT, Gemini, Büyük dil modelleri, Fizyoterapi, Rehabilitasyon programı, Diz osteoartriti

Introduction

The use of artificial intelligence (AI) in the healthcare sector has rapidly expanded with the advancement of large language models (LLMs) such as ChatGPT. While AI previously lacked sufficient effectiveness in clinical decision-making processes, LLMs trained on vast amounts of human-generated texts have become a focal point of research due to their potential to assist clinicians in treatment planning, outcome prediction, and clinical workflows (1-6). LLMs possess the capability to process and analyze complex, large-scale healthcare data, offering critical insights that can reduce cognitive load, improve patient care, and contribute significantly to the healthcare field (7-9).

LLMs can support both patients and physiotherapists by enhancing comprehension, improving treatment techniques, and optimizing outcomes for conditions such as osteoarthritis, which is one of the most common musculoskeletal disorders while addressing the needs

Peer-Review: Double anonymized - Two External
Plagiarism Checks: Yes - intihal.net
Complaints: geneltip@selcuk.edu.tr

Selcuk University Press
Genel Tıp Dergisi | e-ISSN: 2602-3741
https://dergipark.org.tr/tr/pub/geneltip
https://yayinevi.selcuk.edu.tr/

of diverse patient populations (10). Previous studies have evaluated the potential of LLMs to develop rehabilitation prescriptions and perform case-based clinical reasoning. These studies suggest that LLMs hold promise in creating physiotherapy programs tailored to patients' individual needs and conditions (11, 12).

Among the LLM examples, ChatGPT and Gemini are recognized as leading models (13). A review of the literature reveals a paucity of studies examining the applicability of large language models (LLMs) in the field of physiotherapy and rehabilitation. A thorough analysis of research on LLMs and physiotherapy reveals a conspicuous absence of adequate investigation into the performance of the latest versions, demonstrating superior success compared to their predecessors. Considering the advanced capabilities of models like ChatGPT-4o and Gemini Advanced to analyze and synthesize diverse patient data, we hypothesized that LLMs could provide personalized treatment recommendations based on patient assessment data. This study aimed to explore whether ChatGPT-4o and Gemini Advanced differ in their tendencies to suggest specific exercises in response to varying patient assessment data. Our research compares the performance of ChatGPT-4o and Gemini Advanced in providing exercise recommendations based on patient data. This approach offers valuable insights into the potential roles of these models in delivering personalized recommendations within rehabilitation practices.

Materials and Methods

Study Design

This observational study was designed to evaluate the differences in exercise program recommendation tendencies between two large language models, ChatGPT-4o and Gemini Advanced, for patients with knee osteoarthritis (OA). The study was conducted in a physiotherapy outpatient clinic between August and October 2024. The study adhered to the STROBE guidelines.

Setting and Sample

The sample size for the study was calculated using G*Power software (version 3.1.9.7). Based on previous studies in a similar field, a sample size of 40 participants was determined to be sufficient with 80% power and a 5% significance level (11). The study included 40 patients diagnosed with knee OA by a physical medicine specialist according to the American College of Rheumatology criteria. Participants were sedentary individuals aged 40–65 years with stage 2 or 3 OA based on the Kellgren-Lawrence classification.

Exclusion criteria included having undergone knee surgery or joint injection in the last six months, prior participation in any physiotherapy program, cognitive impairment, systemic diseases, or neurological or orthopedic conditions affecting the lower extremities.

Demographic information such as age, gender, body mass index (BMI), and educational level was recorded for all participants. Pain was assessed using a Numeric Pain Rating Scale (NPRS). Range of motion (ROM) for the hip and knee joints was measured in all directions using a universal goniometer, and quadriceps and hamstring muscle strength was assessed with an isometric dynamometer. Functional status was evaluated using the WOMAC and Lysholm scores, while physical performance was assessed through the Timed up-and-go test, the 40-meter fast-walking test, the 30-second sit-to-stand test, and the stair climb test. Static balance was measured using the single-leg stance test and dynamic balance was evaluated with the Four Square Step Test. The selection of these performance tests was based on the standards recommended by the Osteoarthritis Research Society International (OARSI) (14).

Procedure

ChatGPT-4o and Gemini Advanced were instructed to create a three-phase exercise program based on the patient's assessment results. The data were shared with the models using commands written in Turkish. Twelve parameters were analyzed based on the patient's assessment results and the models' recommendations. As supported by established guidelines and previous literature, exercise recommendations with minimal clinical relevance or those rarely used in knee OA rehabilitation were excluded from the study scope (15, 16).

Statistical Analysis

All statistical analyses were performed using the Statistical Package for Social Sciences (SPSS) for Windows, version 25.0. The normality of the data distribution was assessed using the Shapiro-Wilk test, revealing that none of the variables followed a normal distribution. The Mann-Whitney U test was used for comparisons between groups for continuous variables. Spearman's correlation coefficient was applied to evaluate the strength of the relationship between

two variables. In all analyses, p<0.05 was considered statistically significant.

Ethical Considerations

All participants provided written and verbal informed consent, and the study was approved by the Ethics Committee of Kirsehir Ahi Evran University Faculty of Medicine Health Sciences Scientific Research(2024-13/110).

Results

The study initially involved 52 patients; however, 9 patients did not meet the inclusion criteria, and three patients declined to participate. A total of 40 patients completed the evaluation procedures within the scope of the study. Patient evaluation data are shown in Table 1.

In Phase 1 with ChatGPT-4o, statistically significant

Table 1. Statistical summaries of patient data based on clinical test results

Evaluation Parameters	X±SD	Evaluation Parameters	X±SD
Age (years)	53.30±7.17	Hip abduction angle (degrees)	35.35±2.38
Height (cm)	166.18±9.05	Hip adduction angle (degrees)	18.15±1.57
Weight (kg)	67.22±11.70	WOMAC scale score	40.28±12.65
Body mass index (kg/m2)	24.55±4.99	Lysholm scale score	49.48±9.44
NPRS	4.40±1.53	30 sec Sit-to-Stand Test score (times)	11.68±1.59
Pain catastrophizing scale score	31.88±8.21	Stair Climbing Test score (sec)	12.80±3.07
Quadriceps muscle strength (N)	64.73±5.99	40 m Speed Walking Test score (sec)	33.38±3.76
Hamstring muscle strength (N)	39.90±4.30	Timed Up and Go Test score (times)	14.05±2.02
Knee flexion angle (degrees)	101±3.34	Four Step Square Test score (sec)	12.58±2.53
Knee extension angle (degrees)	-3.00 ± 1.50	One-Legged Standing Test score (sec)	16.40±5.51
Hip flexion angle (degrees)	62.15±3.46	FES-I scale score	35.65±6.90
Hip extension angle (degrees)	19.93±2.21		

X: Mean; SD: Standard deviation; y: year; cm: centimeter; kg: kilogram; m: meter; sec: second; N: Newton; NPRS: Numeric pain rating scale; WOMAC: Western Ontario and McMaster Universities osteoarthritis index; SF-12: Short Form 12; FES-I : Falls efficacy scale international

Table 2: Parameter-based statistical comparison of phase 1 exercise recommended and not recommended groups (*p<0.05).

	ChatGPT-4o				Gemini Advanced			
Recommended exercises / Evaluation parameters	Knee Range of Motion	Quadriceps Isometric	Hip Mobili-zation	Hip Abduction and Hamstring Curl	Knee Range of Motion	Quadriceps Isometric	Hip Mobi-lization	Hip Abduction and Hamstring Curl
Numeric Pain Rating Scale	0.179	0.231	0.208	0.235	0.248	0.135	0.278	0.390
Quadriceps Muscle Strength	**0.017***	0.664	0.750	0.361	0.487	0.964	0.659	0.159
Hamstring Muscle Strength	0.142	0.338	0.394	0.229	0.338	0.786	0.396	0.754
Knee Flexion Angle	**0.012***	0.257	0.755	0.237	0.136	0.140	0.483	0.240
Knee Extension Angle	0.708	**0.033***	0.839	0.721	0.479	0.572	0.543	0.860
WOMAC Scale	0.629	0.603	0.098	0.259	0.242	0.269	0.852	0.427
Lysholm Scale	0.281	0.965	0.141	0.754	0.278	0.652	0.685	**0.044***
30 Sec. Sit-to-Stand Test	0.594	0.113	0.406	0.456	0.135	0.234	0.877	0.134
Stair Climbing Test	0.152	0.088	0.215	0.602	0.861	0.945	0.166	0.152
40 m. Fast-Walking Test	0.475	0.089	0.681	0.180	0.206	0.525	0.622	0.291
Timed Up and Go Test	0.251	0.569	0.276	0.237	0.483	0.294	0.894	0.989
Four-Step Square Test	0.065	0.096	0.239	**0.043***	0.204	0.665	0.365	0.557
Single Leg Stance Test	0.467	0.098	0.116	0.404	0.965	0.188	0.307	0.540
Falls Efficacy Scale International	0.281	0.896	0.727	0.488	0.208	0.542	0.498	0.943
Pain Catastrophi-zing Scale	0.105	0.179	0.739	0.253	0.543	0.874	0.839	0.765

WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

differences were found in quadriceps muscle strength, knee flexion angle, knee extension angle and four-step square test parameters in knee ROM, quadriceps isometric and hamstring curl exercises, respectively; in Gemini Advanced, only in Lysholm scale score and hip abduction and hamstring curl exercise, between the groups recommended and not recommended for exercise (p=0.017*, p=0.012*, p=0.033*, p=0.043*, p=0.044*). The data from Phase 1 are shown in Table 2. In phase 2, a statistically significant difference was found between the recommended and non-recommended groups in ChatGPT-4o quadriceps muscle strength and hip stabilization exercises and balance and proprioception exercise with Lysholm scale (p=0.032*, p=0.040*). Phase 2 data are shown in Table 3.

In Phase 3, statistically significant differences were found only in quadriceps muscle strength and lower extremity strengthening exercises for ChatGPT-4o, and

Table 3. Parameter-based statistical comparison of phase 2 exercise recommended and not recommended groups (*p<0.05).

	ChatGPT-4o				Gemini Advanced			
Recommended exercises / Evaluation parameters	Quadriceps CKC strengthening	Balance and proprioception	Hip stabilization	Hip abduction and hamstring curl	Quadriceps CKC strengthening	Balance and proprioception	Hip stabilization	Hip abduction and hamstring curl
Numeric pain rating scale	0.281	0.251	0.204	0.294	0.488	0.251	0.307	0.515
Quadriceps muscle strength	0.828	0.240	**0.032***	0.169	0.437	0.575	0.487	0.745
Hamstring muscle strength	0.708	0.874	0.052	0.505	0.249	0.190	0.338	0.425
Knee flexion angle	0.106	0.267	0.328	0.162	0.477	0.489	0.654	0.497
Knee extension angle	0.756	0.615	0.604	0.867	0.614	0.724	0.543	0.565
WOMAC scale	0.384	0.701	0.647	0.908	0.082	0.419	0.242	0.390
Lysholm scale	0.238	**0.040***	0.574	0.573	0.136	0.555	0.278	0.456
30 seconds sit-to-stand test	0.313	0.450	0.187	0.660	0.231	0.449	0.135	0.349
Stair climbing test	0.572	0.802	0.263	0.343	0.510	0.286	0.861	0.719
40-meter fast-walking test	0.950	0.602	0.835	0.196	0.399	0.382	0.206	0.961
Timed up and go test	0.221	0.315	0.429	0.237	0.572	0.140	0.483	0.589
Four-step square test	0.316	0.425	0.121	0.715	0.069	0.730	0.204	0.482
Single-leg stance test	0.399	0.073	0.755	0.611	0.950	0.454	0.965	0.147
Falls Efficacy Scale ınternational	0.576	0.058	0.204	0.479	0.072	0.238	0.208	0.548
Pain catastrophizing scale	0.901	0.443	0.615	0.470	0.732	0.087	0.543	0.721

CKC: Closed kinetic Chain; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

Table 4. Parameter-based statistical comparison of the groups recommended and not recommended for phase 3 exercise (*p<0.05).

	ChatGPT-4o				Gemini Advanced			
Recommended exercises / Evaluation parameters	Quadriceps strengthening	Lower extremity strengthening	Dynamic balance	Adaptation to ADL	Quadriceps strengthening	Lower extremity strengthening	Dynamic balance	Adaptation to ADL
Numeric pain rating scale	0.221	0.078	0.121	0.421	0.462	0.333	0.307	0.665
Quadriceps muscle strength	0.144	**0.007***	0.918	0.172	0.395	0.985	0.167	0.931
Hamstring muscle strength	0.143	0.096	0.570	0.126	0.536	0.662	0.088	0.240
Knee flexion angle	0.062	0.126	0.302	0.573	**0.047***	0.878	0.666	0.457
Knee extension angle	0.561	0.456	0.682	0.799	0.611	0.720	0.540	0.601
WOMAC scale	0.975	0.257	0.625	0.923	0.411	0.691	0.225	0.209
Lysholm scale	0.877	0.761	0.054	0.092	0.292	0.483	0.682	0.515
30 seconds sit-to-stand test	0.231	0.303	0.958	0.818	0.465	0.577	0.451	0.860
Stair climbing test	0.826	0.182	0.161	0.368	0.979	0.848	0.291	0.861
40-meter fast-walking test	0.211	0.110	0.326	0.282	0.660	0.269	0.867	0.206
Timed up and go test	0.346	0.161	0.659	0.385	0.795	0.235	0.902	0.483
Four-step square test	0.572	0.313	0.938	0.567	0.756	0.924	0.679	0.431
Single-leg stance test	0.851	0.721	0.071	0.134	0.918	0.760	0.070	0.191
Falls Efficacy Scale International	0.756	0.165	0.157	0.436	0.304	0.149	0.430	0.896
Pain catastrophizing scale	0.384	0.132	0.817	0.604	0.817	0.865	0.439	0.435

ADL: Activities of daily living; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

in knee flexion angle and quadriceps strengthening exercises for Gemini Advanced (p=0.007, p=0.047). Phase 3 data are shown in Table 4.

Discussion

Studies investigating the use of artificial intelligence-based LLMs in the field of physiotherapy and rehabilitation appear to be quite limited. A review of the literature reveals insufficient focus on the performance of newer versions of LLMs and a notable lack of rehabilitation-focused studies. This study aims to address this gap by being the first to evaluate the performance of large language models, ChatGPT-4o and Gemini Advanced, in recommending personalized exercise programs based on varying patient data for knee osteoarthritis.

The findings indicate that ChatGPT-4o demonstrated statistically significant differences in more parameters compared to Gemini Advanced. Whether these recommendations exhibited systematic tendencies based on assessment parameters has highlighted the differences in the models' approaches.

In the literature, studies evaluating the clinical decision-making capabilities of ChatGPT and Gemini assess both the potential and limitations of these LLMs. For instance, in a study investigating ChatGPT's performance in clinical reasoning and treatment management for simple and complex scenarios in physiotherapy, it was suggested that ChatGPT holds the potential as a valuable resource for clinical decision support (11). Similarly, another study evaluating performance in solving complex clinical questions requiring clinical reasoning skills in vestibular rehabilitation found that ChatGPT outperformed Gemini but advised caution due to their limited accuracy (17). Chen et al. explored the applicability of LLMs in knee OA treatment and found that ChatGPT-3.5 and ChatGPT-4.0 exhibited inadequate performance when transitioning from general information provision to generating personalized solutions. However, the study noted that integrating accurate resources could significantly enhance their performance. This research highlighted the limitations of general-purpose LLMs in clinical contexts and underscored the importance of developing domain-specific LLMs through specialized approaches (18). Additionally, a 2024 study reported that ChatGPT-4.0 demonstrated higher accuracy and consistency rates in providing general information and designing rehabilitation programs compared to ChatGPT-3.5 and other language models (10).

A study comparing the current performance of the two most common LLMs, ChatGPT-4 and Gemini Pro (1.0 Pro), in providing intraoperative decision support for plastic and reconstructive surgery procedures found that ChatGPT-4 significantly outperformed Gemini in delivering accurate and relevant responses. However, Gemini was noted for producing more concise and readable responses and having a faster average response time compared to ChatGPT-4. The study emphasized that both models require further training and optimization to address performance inconsistencies across different procedures and enhance their reliability as decision-support tools (19). In another study evaluating the accuracy of recommendations generated by LLMs for common pediatric orthopedic conditions, ChatGPT achieved an accuracy rate of 67%, while Gemini reached 69%. However, it was highlighted that neither model could reliably represent the most up-to-date sources of medical knowledge (20).

In our study, ChatGPT-4o demonstrated statistically significant differences in seven different parameters: Four in Phase 1, two in Phase 2, and one in Phase 3. In contrast, Gemini Advanced showed statistically significant differences in only one parameter each in Phase 1 and Phase 3. These findings suggest that ChatGPT-4o may take assessment data into account more effectively in exercise recommendations compared to Gemini Advanced. This difference may be related to the logical reasoning foundation, one of the key distinctions between the two language models. For instance, ChatGPT's o1 model has been shown to outperform Gemini in functions such as solving codes and puzzles, analyzing complex problems, and correctly answering mathematical questions through the chain-of-thought method (21). However, both language models showed statistically significant differences in only a limited number of groups. This finding highlights the continued limitations of language models in providing personalized exercise recommendations based on patient data in knee osteoarthritis rehabilitation. In the future, the development of a language model specialized in physiotherapy and rehabilitation has the potential to provide more personalized and effective solutions tailored to the specific needs and challenges of this discipline.

Limitations of the Study

The study's limitations are evident in two respects: firstly, the number of patients evaluated was small;

secondly, it was exclusively conducted in Turkish. These factors restrict the generalizability of the results to other languages and larger patient populations.

Conclusions and Recommendations

In conclusion, providing extensive, high-quality, and complex datasets encompassing diverse clinical scenarios, along with the real-time integration of updated medical knowledge, could enhance the clinical reasoning performance of LLMs. Given the potential for errors in LLMs, establishing feedback loops to identify and optimize these errors could be effective in enabling these models to offer personalized recommendations that account for individual patient needs and values. Moreover, improving the performance of language models across different languages and designing domain-specific models tailored to clinical fields such as physiotherapy could expand their general applicability and facilitate more specialized and effective solutions.

Conflict of interest

The authors declare no conflicts of interest.

Financial support

This study received no financial support.

Acknowledgment

The authors express their gratitude to all participants.

References

1.Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2024.

2.Goldberg CB, Adams L, Blumenthal D, et al. To do no harm—and the most good—with AI in health care. NejmAi. 2024. p. Alp2400036.

3.Kohane IS. Injecting artificial intelligence into medicine. NejmAi. 2024. p. Ale2300197.

4.Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. JMIR. 2023;25:e48659.

5.Stafie CS, Sufaru I-G, Ghiciuc CM, et al. Exploring the intersection of artificial intelligence and clinical healthcare: a multidisciplinary review. Diagnostics. 2023;13:1995.

6.Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? JAMA. 2024;331:65-9.

7.Nazi ZA, Peng W, editors. Large language models in healthcare and medical domain: A review. Informatics; 2024: MDPI.

8.Duran A, Cortuk O, Ok B. Future Perspective of Risk Prediction in Aesthetic Surgery: Is Artificial Intelligence Reliable? Aesthet Surg J. 2024;44:NP839-NP49.

9.Güneş YC, Cesur T, Çamur E. Comparative Analysis of Large Language Models in Simplifying Turkish Ultrasound Reports to Enhance Patient Understanding. EurJTher. 2024;30:714-23.

10.Cao M, Wang Q, Zhang X, et al. Large language models' performances regarding common patient questions about osteoarthritis: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and perplexity. J Sport Health Sci. 2024:101016.

11.Bilika P, Stefanouli V, Strimpakos N, Kapreli EV. Clinical reasoning using ChatGPT: Is it beyond credibility for physiotherapists to use? Physiother Theory Pract. 2024;40:2943-62.

12.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.

13.Nazir T, Ahmad U, Mal M, et al. Microsoft Bing vs Google Bard in Neurology: A comparative study of AI-generated patient education material. medRxiv. 2023:2023.08. 25.23294641.

14.Dobson F, Hinman RS, Roos EM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis cartilage. 2013;21:1042-52.

15.Fransen M, McConnell S, Harmer AR, et al. Exercise for osteoarthritis of the knee. Cochrane database of systematic reviews. 2015.

16.McAlindon TE, Bannuru RR, Sullivan M, et al. OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthritis cartilage. 2014;22:363-88.

17.Arbel Y, Gimmon Y, Shmueli L. Evaluating the Potential of Large Language Models for Vestibular Rehabilitation Education: A Comparison of ChatGPT, Google Gemini, and Clinicians. medRxiv. 2024:2024.01. 24.24301737.

18.Chen X, You M, Wang L, et al. Evaluating and Enhancing Large Language Models Performance in Domain-specific Medicine: Osteoarthritis Management with DocOA. arXiv preprint arXiv:240112998. 2024.

19.Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large Language Models for Intraoperative Decision Support in Plastic Surgery: A Comparison between ChatGPT-4 and Gemini. Medicina. 2024;60:957.

20.Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini Provide Appropriate Recommendations for Pediatric Orthopaedic Conditions? J Pediatr Orthop. 2025;45:e66-e71.

21.Lau J. Gemini vs. ChatGPT: What's the difference? [2025]. In: Zapier, editor. 2024.