DOI: 10.37990/medr.1581104 Med Records 2025;7(1):120-8

MEDICAL RECORDS-International Medical Journal

Research Article

A New Approach: Generative Artificial Intelligence in Physiatry Resident Education

Selkin Yilmaz Muluk, Vedat Altuntas, Zehra Duman Sahin

Antalya City Hospital, Department of Physical Medicine and Rehabilitation, Antalya, Türkiye

Abstract

Aim: This study assessed the effectiveness of ChatGPT-4o, an artificial intelligence (AI) platform, in creating a therapeutic exercises presentation for physiatry residents' education. The aim was to compare the quality of content created by ChatGPT-4o with that of an expert, exploring the potential of AI in healthcare education.
Material and Method: Both an expert and AI created 24 PowerPoint slides across six topics, using same reputable sources. Two other experts assessed these slides according to CLEAR criteria: completeness, lack of false information, appropriateness, and relevance and scored as excellent, 5; very good=4, good=3, satisfactory/fair=2, or poor, 1.
Results: Interrater reliability was confirmed. Average scores (calculated from the two raters' scores) for each topic were significantly lower for AI than for the expert, although whole presentation scores did not differ between the two. Overall scores (calculated from the average scores of all items) for each topic were good to excellent for AI, excellent for the expert. The overall score for whole presentation was good for AI, excellent for the expert. Highest ranks for individual criteria was relevance for AI, lack of false information for the expert. Some AI-generated elements were later integrated into the expert work, enhancing the content.
Conclusion: ChatGPT-4o can generate effective educational content, though expert outperforms it, highlighting the need for professional oversight. Collaboration between humans and AI may further enhance educational outcomes.

Keywords: Healthcare education, physiatry residency, therapeutic exercises, ChatGPT, generative AI

INTRODUCTION

Residency education in medicine is a period of specialized training that takes place between graduating from medical school and becoming a certified independent specialist. Unlike basic medical education, it is shaped by specific healthcare settings and needs, and serves as a bridge between academic learning and real-world medical practice (1). Structured training programs that prioritize skill building and formal education are essential for supporting junior doctors and ensuring their competence (2).

When determining educational strategies, it is essential to consider institutional preferences, target populations, and learning style. A variety of methods are widely used, including simulation-based approaches, scenarios, standardized patients, research, mentoring, journal clubs, seminars, lectures, case discussions, bedside discussions, courses, games, and portfolios (3). In physiatry clinics, educational sessions such as seminars, lectures, case discussions, and journal reviews are designed for postgraduate physiatry residents and practicing physiatrists. While these programs also include bedside discussions, direct patient care, and invasive procedures, the knowledge gained from lectures and seminars serves as the basis for effective real-time practice.

In Türkiye, the Medical Specialty Regulation defines the curriculum, planning, programming, and implementation principles for each specialty (4). 'Therapeutic exercises' is one of the topics that is generally included in the curriculum of Physical Medicine and Rehabilitation programs. This topic involves movements prescribed to correct impairments, restore muscular and skeletal functions, and maintain a state of well-being in patients. They are beneficial for quality of life, and overall health (5).

CITATION
Yilmaz Muluk S, Altuntas V, Duman Sahin Z. A New Approach: Generative Artificial Intelligence in Physiatry Resident Education. Med Records. 2025;7(1):120-8. DOI:1037990/medr.1581104

Received: 07.11.2024 **Accepted:** 04.12.2024 **Published:** 14.01.2025
Corresponding Author: Selkin Yilmaz Muluk, Antalya City Hospital, Department of Physical Medicine and Rehabilitation, Antalya, Türkiye
E-mail: selkinylmaz@yahoo.com

Generative artificial intelligence (AI) platforms are systems that can generate relevant responses by drawing on vast amounts of knowledge and information and mimicking human-like conversations. They can be applied in numerous medical fields, including image analysis, clinical diagnostics, drug development, patient assistance and education, remote monitoring, tailored treatment plans, administrative functions, and medical documentation (6).

A prominent example of generative AIs is ChatGPT created by OpenAI (7). ChatGPT can be used to prepare medical letters, imaging reports, and patient discharge documents (8). It can also play a role in summarizing drug labeling documents and creating safety protocols for invasive procedures (9,10). ChatGPT-4o is one of the fastest and most developed versions of the ChatGPT.

As generative AIs have been successful in various medical fields and are able to write, summarize, and create medical texts, thats very possible that they can also summarize articles or texts and create educational slides out of them.

In this research, we aimed to evaluate the effectiveness of ChatGPT-4o in creating PowerPoint slides by comparing its slides with those prepared by a physiatrist. Both sets of slides were developed using the same reputable sources to ensure consistency. Our approach consisted of two steps: first, a physiatrist and ChatGPT-4o created 35 slides for a presentation on therapeutic exercises using relevant articles; then, two additional physiatry experts reviewed and scored both slide sets for completeness, lack of false information, appropriateness, and relevance.

The objective of this study is to evaluate ChatGPT-4o's effectiveness in preparing a therapeutic exercises presentation for an educational session for physiatry residents.

Our research has two hypotheses. The effectiveness hypothesis suggests that ChatGPT-4o will prepare a presentation rated above moderate effectiveness. The null hypothesis for this asserts that ChatGPT-4o will not effectively prepare the presentation and that it will be rated below moderate. The performance hypothesis predicts that the expert, based on subject knowledge, will outperform ChatGPT-4o. The null hypothesis states that there will be no variation in performance between the expert and ChatGPT-4o.

These hypotheses emphasize the necessity of evaluating the effectiveness of AI tools in contributing to lecture hours to meet the growing demands of health education.

MATERIAL AND METHOD

Study Design

This study integrated qualitative and quantitative components, categorizing it as a mixed-methods study. It is reported according to METRICS checklist that involves model used and its settings, evaluation approach, timing, transparency, range of tested topics, randomization, individual factors and interrater reliability, count of requests, and specificity of the prompts and language used (11). Since the study did not involve human participation and centered on engagements with conversational AI systems, ethical approval was not necessary.

Model Used and its Exact Setting

ChatGPT is a conversational AI system powered by large language models. We chose the 4o version created by OpenAI due to its status as one of the most advanced versions available at the time of our search (7). The system was assessed using standard default configurations to ensure consistent replication of the generated content. Since ChatGPT-4o does not remember information from prior interactions, each conversation begins without any reference to previous questions or answers, thereby eliminating the possibility of learning or feedback loops. However, all slide requests for each topic were made using a new session, and the regenerate button was not utilized.

Evaluation Approach for the Generated Content

The physiatry specialist initially created a presentation based on information from six selected articles about therapeutic exercises (12-17). Subsequently, the same six articles were submitted to ChatGPT-4o and it was tasked with generating an equivalent number of slides suitable for a PowerPoint presentation. Once both sets of slides were ready, the specialist anonymized them by assigning letter codes, making it impossible to tell which set was AI-generated and which was expert-made. Then, two other physiatry specialists, who had previously reviewed the articles, independently evaluated both sets of slides without knowing their sources. To maintain objectivity, the raters were also unaware of each other's scores during the assessment.

Given that the presentations shared the same underlying material, we focused on the C (Completeness of the Content), L (Lack of False Information), A (Appropriateness of the Content), and R (Relevance) elements of the CLEAR scoring system. The "E" component, which evaluates evidence supporting content, was deemed less distinct between the two presentations due to the uniformity of the sources used. By concentrating on the remaining CLEAR criteria, we aimed to provide a comprehensive analysis of how each presentation utilized shared evidence and conveyed information effectively to the audience. In this tool, items are scored as follows: excellent, 5; very good=4, good=3, satisfactory/fair=2, or poor, 1 (18).

Timing of Testing and Transparency of the Data

AI model was tested on September 21, 2024, at local time 12:20-12.35, in Istanbul zone. The conversations have been recorded in the public data archive Zenodo (19).

Range of Tested Topics and Randomization

The authors selected a specialized issue of a reputable local journal, published in Turkish, that focused on

therapeutic exercises. This issue included a range of relevant topics, such as joint range of motion exercises, stretching exercises, peripheral joint mobilization and manipulation, muscle performance exercises, aerobic exercises, aquatic exercises, posture exercises, and relaxation exercises (12-17). Articles that were not aligned with the focus of the study were excluded. For example, the article entitled "Exercise for Healthy Living and Prevention of Chronic Diseases" was omitted because of its emphasis on preventive rather than therapeutic interventions. Likewise, specialized discussions on exercises for specific conditions, such as soft tissue injuries, orthopedic surgeries, and respiratory issues, were left out to keep the core content general. Since all six essential therapeutic exercise topics were covered, randomization was deemed unnecessary.

Individual Factors in Selecting the Topic and Interrater Reliability

Therapeutic exercises subject is a foundational component of the first-year curriculum for resident physiatrists. To ensure thorough training, the authors concentrated on general therapeutic exercises for initial learning, reserving more specialized topics for subsequent sessions. The use of a reputable journal known for its thorough coverage of essential physiatry subjects ensured that personal preferences or biases did not influence the selection of topics, leading to a comprehensive introduction ideal for newly trained physicians.

To enhance objectivity in the assessment process, two independent raters, physiatrists working in an outpatient clinic and rehabilitation service, evaluated the content. Statistical measures indicated significant agreement between the raters. The inter-rater reliability confirmed that the assessments were consistent and had minimal influence from individual biases, thereby enhancing the validity of the study's findings.

Count of Slides Requested from the Model

Of the six topics selected, ten slides were requested for the first, as it was more comprehensive, while five slides were requested for each of the remaining five topics. The topics of the articles used by the physician and AI to prepare the training outline are presented in Table 1.

Table 1. Topics of the articles used in preparation of educational slide content

	Therapeutic exercises
1	Joint range of motion exercises, stretching exercises, peripheral joint mobilization and manipulation
2	Muscle performance exercises
3	Aerobic exercises
4	Aquatic exercises
5	Posture and posture exercises
6	Relaxation exercises

Specificity of the Prompts and Language Used

The questions followed a consistent methodology, using the following prompt (in Turkish): 'I plan to prepare a PowerPoint presentation on "..." for assistant doctors who are in the first year of their specialty training in the Physical Medicine and Rehabilitation Department. This presentation will be used during an educational hour. Please take the provided text and create a detailed PowerPoint presentation consisting of ... slides formatted in Turkish. Each slide should contain complete, cohesive content that I can read directly to an audience without headings.

I specifically want the slides to be organized in a listing pattern with bullet points or numbered lists to enhance readability and fit well into the PowerPoint slide format. The slides should cover essential aspects and explanations for each section, reflecting the key points suitable for a physiatry seminar. I want to clarify that I want the slides to be prepared according to the content of the text I will be loading. Please wait for the text to be loaded before starting to create the slides.' We deliberately crafted this request to resemble that of a physician. This strategy aimed to reflect professional tone of a professional requesting assistance for an upcoming educational mission. By doing so, we aimed to make our interactions with the AI platform similar to a genuine academic scenario, making the generated content more relevant to actual needs. The prompt was designed according to the recommendations of Meskó B (20).

Statistics and Data Analysis

Data analysis was carried out utilizing IBM SPSS Statistics for Windows version 29.0.2.0 (IBM Corp., Armonk, NY), with a significance level set at $p < 0.050$. Two physiatrists independently evaluated the presentations, referred to as rater 1 and rater 2.

The strength and direction of the association between the two ordinal variables were measured using Kendall's tau-b statistic. Kendall's Tau-b values are categorized as follows: 0.00 to ±0.10: very weak or no correlation, ±0.11 to ±0.30: weak correlation, ±0.31 to ±0.50: moderate correlation, ±0.51 to ±0.70: strong correlation, ±0.71 to ±1.00: very strong correlation. Kendall's tau-b values ranged from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no association (21).

Agreement between the two independent evaluators was assessed utilizing Cohen's kappa method. Cohen's Kappa quantifies inter-rater agreement for categorical data. The categorization of Cohen's kappa values is as follows: values below 0.20 indicate poor agreement, 0.21-0.40 signify fair agreement, 0.41-0.60 reflect moderate agreement, 0.61-0.80 represent substantial agreement, and 0.81-1.00 indicate nearly perfect agreement (22).

After measuring inter-rater reliability, the scores for each topic from both evaluators were totalled and divided by two. These results were accepted as 'average scores.' For example, the C score for topic 1 from evaluator 1 and the C score for topic 1 from evaluater 2 were summed and then divided by two, resulting in the 'average C score' for topic 1. Next, the average scores of the four CLEAR items for each topic were totalled and divided by four, with the resulting value accepted as the 'overall score' for each topic. For instance, the C, L, A, and R average scores for topic 1 were summed and divided by four, resulting in the 'overall score' for topic 1.

Additionally, the raters evaluated the entire presentation, considering all slides collectively, to provide 'average scores' and 'overall scores' for the whole presentation. The overall scores were organized into the following categories: scores of 1-1.79 as "poor", 1.80-2.59 as "satisfactory", 2.60-3.39 as "good", 3.40-4.19 as "very good", and 4.20-5.00 as "excellent" (18).

A paired samples t-test was used to compare average scores, as the Shapiro-Wilk test confirmed normality. For the entire presentation scores, the Wilcoxon signed-rank test was applied, as the data did not meet the normality assumption. Further analysis assessed performance on CLEAR items (excluding E) by examining within-model variability.

RESULTS

The results of the inter-rater correlation indicated a very strong correlation for the AI-generated content and a moderate correlation for the expert-generated content according to Kendall's Tau-b statistics. Similarly, the results of the inter-rater agreement indicated substantial agreement for the AI-generated content and fair agreement for the expert-generated content according to Cohen's Kappa statistics. Notably, although the raters exhibited stronger agreement for the AI-generated slides than for the expert-generated slides, the p-values confirmed that the correlation and the agreements were statistically significant in both groups. Given this significance, it was appropriate to calculate the average of the two raters' scores, as utilizing these averaged values would provide a single, more reliable score for subsequent analyses. The results of these tests are presented in Table 2.

Table 2. Interrater correlation and agreement for ai and expert-generated presentations

	Kendall's Tau-b	p-value (Kendall's Tau-b)	Categorization (Kendall's Tau-b)	Cohen's Kappa	p-value (Cohen's Kappa)	Categorization (Cohen's Kappa)
AI	0.723	<0.001	Very strong correlation	0.514	<0.001	Substantial agreement
Expert	0.513	<0.001	Moderate correlation	0.401	0.016	Fair agreement

Table 3 presents descriptive statistics of the cumulative average scores (derived from the initial average scores) for the AI-generated and expert-prepared slides. The cumulative values were calculated from 24 average scores, with each of the six topics contributing four scores. For AI-generated slides, the cumulative score was 3.77. In contrast, expert-prepared slides achieved a higher cumumulative score of 4.52.

Table 3. Descriptive statistics for cumulative average scores of AI and expert-generated slides

	N	Min	Max	Mean	SD
Cumulative scores of AI-prepared slides	24	2.50	5.00	3.77	0.79
Cumulative scores of expert-prepared slides	24	3.50	5.00	4.52	0.48

N: number of scores, Min: minimum, Max: maximum, SD: standard deviation

Table 4 presents the descriptive statistics of the overall scores for each of the six topics and for the entire presentation. The overall scores of the AI-prepared slides were rated as good to excellent, whereas the expert-prepared slides were consistently rated as excellent. The expert presentations excelled in all categories, while the AI-generated content performed notably well in the "Aquatic Exercises" category. A comparison of overall scores of AI and Expert-generated content per each topic is presented in Figure 1.

Table 4. Descriptive statistics for overall scores of AI and expert-generated slides

Therapeutic exercises subheadings	Prepared by	Min	Max	Mean	SD	Category
JRME, SE, PJMM	AI	3.00	4.00	3.25	0.50	Good
	Expert	4.00	5.00	4.75	0.50	Excellent
Muscle Performance E	AI	2.50	5.00	3.50	1.08	Very Good
	Expert	4.00	5.00	4.38	0.48	Excellent
Aerobic E	AI	3.00	4.00	3.63	0.48	Very Good
	Expert	4.00	5.00	4.50	0.41	Excellent
Aquatic E	AI	4.00	5.00	4.50	0.41	Excellent
	Expert	3.50	5.00	4.50	0.71	Excellent
Posture and Posture E	AI	3.00	4.50	3.63	0.75	Very Good
	Expert	4.00	5.00	4.50	0.58	Excellent
Relaxation E	AI	3.00	5.00	4.13	1.03	Very Good
	Expert	4.00	5.00	4.50	0.41	Excellent
Whole Presentation	AI	3.00	4.00	3.38	0.48	Good
	Expert	4.00	4.50	4.25	0.29	Excellent

JRME, SE, PJMM: joint range of motion exercises, stretching exercises, peripheral joint mobilization and manipulation, E: exercises, Min: minimum, Max: maximum, SD: standard deviation

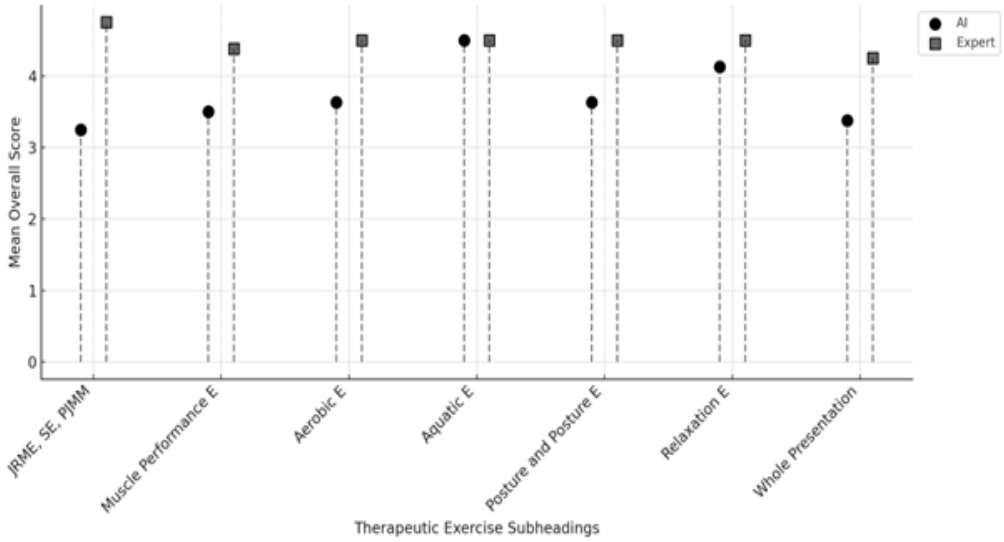

Figure 1. Comparison of overall scores of AI and expert-generated content per each topic

After demographic statistics, we planned to compare the average scores of AI-generated and expert-generated slides. Before the comparison, the normality distribution was tested using the Shapiro-Wilk test. The results showed no significant deviation from normality ($W(24)=0.924$, $p=0.071$; $W(24)=0.924$, $p=0.071$), indicating that conducting a parametric test was suitable. The paired-samples t-test revealed a statistically significant difference between the

average scores of AI and expert-prepared slides (p<0.001). The negative mean difference indicates that the AI-made slides' scores were significantly lower than those of human-made slides'. Furthermore, the effect size was large (Cohen's d=-0.998), highlighting a large difference between the groups (Table 5).

Table 5. Comparison of mean of average scores of ai and expert presentations

Measure	Mean difference	SD	SEM	95% CI Lower	95% CI Upper	t	df	p-value (two-sided)	Cohen's d (effect size)
AI - expert	-0.75	0.75	0.15	-1.07	-0.43	-4.89	23	< 0.01	-0.998

SD: standard deviation, SEM: standars error mean, CI: confidence interval

Later, we planned to compare the average scores of the whole presentation of AI and expert. The cumulative score for the expert-prepared entire presentation (4.25) was higher than AI-generated presentation (3.38). Before the comparison, the normality distribution of the scores was tested using the Shapiro-Wilk test. The results showed a significant deviation from normality (W=0.630, p=0.001; W=0.630, p=0.001). Therefore, we performed a nonparametric test, the Wilcoxon signed-rank test, for the scores of the entire presentation. The results of this test indicated a test statistic (Z) of -1.890, based on negative ranks. The asymptotic significance (2-tailed p-value) was 0.059. This p-value suggested that the difference in scores was not statistically significant at the conventional alpha level of 0.05.

Table 6 presents the final analysis of performance for each CLEAR item, examining within-model variability. The Friedman test revealed a statistically significant difference in ranks across the four CLEAR items for both AI-prepared content ($\chi^2(3)=10.39$, p=0.016) and expert-prepared content ($\chi^2(3)=11.70$, p=0.008). For AI-prepared slides, "relevance" and "lack of false information" received higher ranks, while for expert-prepared slides, "lack of false information" and "appropriateness" were ranked higher. "Completeness" had the lowest rank in both AI and expert presentations. The within-model variability in AI and expert mean ranks across the items is presented in Figure 2.

Table 6. Within-model variability in mean ranks and statistical results across CLEAR items for AI and expert-generated content

	Mean Rank (AI)	Mean Rank (Expert)
C	1.50	1.17
L	3.00	3.42
A	2.00	2.75
R	3.50	2.67
Chi-Square (χ^2)	10.39	11.70
p-value	0.016	0.008

C: Completeness of the Content, L: Lack of False Information, A: Appropriateness of the Content, and R: Relevance

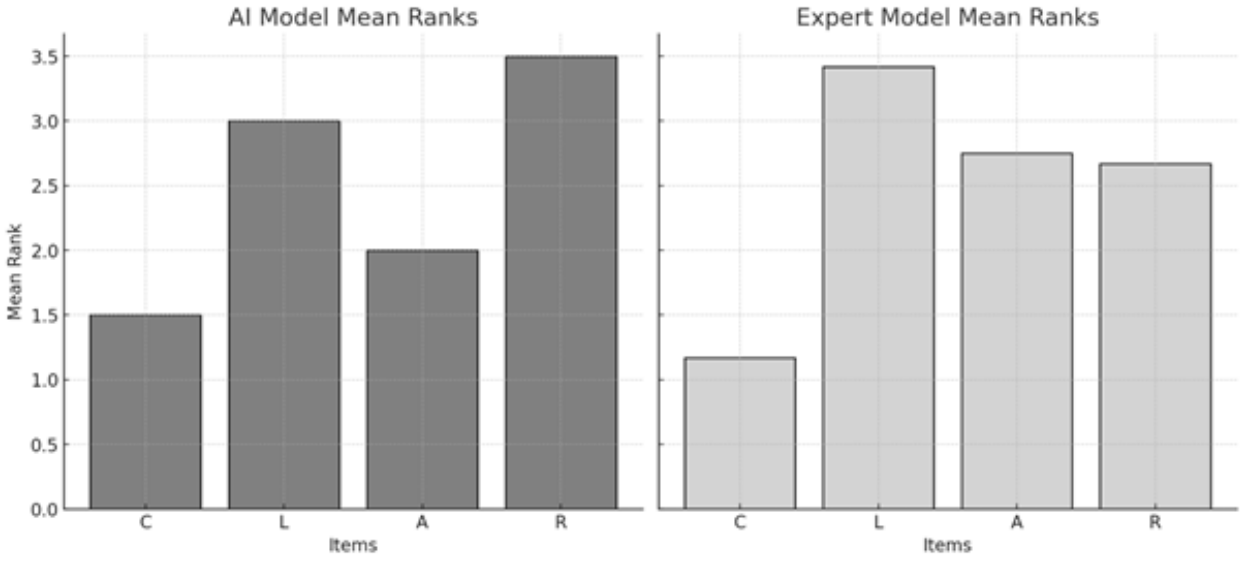

Figure 2. Within-Model Variability in Mean Ranks Across CLEAR Items for AI- and Expert-generated; C: completeness of the content, L: lack of false information, A: appropriateness of the content, and R: relevance

DISCUSSION

This study explored a novel approach to preparing educational presentations aimed at enhancing the educational hours of physiatry residents. Specifically, it evaluated the effectiveness of ChatGPT-4o in creating PowerPoint slides for therapeutic exercises. Presentations generated by both AI and the expert were assessed by blinded evaluators.

In this study, the cumulative average score for AI-generated presentations was 3.77 out of 5, while expert-prepared presentations achieved a higher mean score of 4.52 out of 5, indicating superior performance of the expert. Notably, the cumulative scores for the expert-prepared slides were significantly higher than those for the AI slides ($p<0.001$). However, when the presentations were assessed as a whole, no statistically significant difference was found between the cumulative scores ($p=0.059$).

Furthermore, the overall scores for individual topics indicated that the AI's performance ranged from good to excellent, whereas the expert received excellent ratings across all topics. When considering the entire presentation, the AI received a score of 3.38 out of 5, classified as "good", while the expert received a score of 4.25 out of 5, classified as "excellent".

These findings support both our hypotheses, suggesting that ChatGPT-4o would prepare a presentation rated above moderate and that the physiatrist would outperform ChatGPT-4o because of expertise.

As far as we know, this study is the first to investigate the capability of an AI system to create PowerPoint slides for a presentation based on medical texts. It represents a pioneering effort to evaluate the effectiveness of AI in preparing educational materials for residency programs.

To contextualize these findings, it is important to consider the existing literature on the use of AI in medicine and healthcare.

Previous studies have suggested that conversational AI platforms can enhance healthcare by reducing the daily burden on professionals. For example, ChatGPT has demonstrated its capability as an effective tool for improving medical documentation, such as clinical letters, imaging reports, and discharge reports (8). In a simulated case study, ChatGPT reviewed a conversation between a patient and a doctor, generated medical records, suggested differential diagnoses, and provided treatment recommendations; the results closely aligned with the physician's summaries (23). Likewise, ChatGPT was employed to facilitate the process of writing clinical letters for prior authorization requests from insurance providers. This innovative approach was noted to potentially save physicians considerable time, enabling them to concentrate more on patient care and clinical decision-making (24). Our findings support these results, demonstrating that the 4o version of ChatGPT can contribute to preparing medical content, even though expert performed better.

Literature also suggests that AI systems can play a beneficial role in health education initiatives. A systematic review pointed out that ChatGPT has promising uses in healthcare education, research, and practice, including improving scientific writing, streamlining workflows, and enhancing personalized learning (25). Munaf et al. recommended that resident doctors use ChatGPT for generating reports, creating mnemonics, and simulating clinical scenarios, thereby reducing administrative tasks, enhancing learning, and improving patient interaction as it streamlines workflows and presents information clearly (26). It has been shown that AI tools like ChatGPT can aid in content creation, support learning, and offer new opportunities for assessment and research in medical and postgraduate education (27). AI serves multiple roles in medical education, including enhancing clinical specialty training, facilitating personalized and adaptive learning, and improving decision-making through advanced data analysis. Additionally, AI integration promotes increased efficiency and accuracy in educational processes, driving the modernization and diversification of medical curricula (28). Our research supports these findings on educational effectiveness, as the AI-generated slides in our study were rated from good to excellent. This suggests that ChatGPT-4o could effectively reduce the workload for educator physicians, allowing them more time for other tasks.

Nonetheless a systematic review highlighted that ChatGPT only achieved moderate or 'passing' performance across various tests, deeming it unreliable for clinical deployment due to its nonclinical design (29). In line with this, our study highlighted the need for professional support when managing AI-generated content, as the expert's performance consistently surpassed that of the AI on cumulative and overall. This suggests two potential approaches: implementing professional oversight of AI-generated drafts or fostering collaboration between AI and experts.

Additionally, the use of AI in educational contexts is not without challenges. There are critical concerns regarding the potential for inaccurate information, inherent biases, and the necessity for robust privacy and security measures (26). Bajwa et al. asserted that attention must be given to ensuring ethical access to data, possessing the necessary expertise in medical fields, having sufficient computing power, and addressing the challenges associated with implementing AI in real-world settings (30). Moreover, ChatGPT should also be used with caution due to ethical, copyright, transparency, and legal issues, as well as risks of bias, plagiarism, lack of originality, inaccurate content with hallucination risks, limited knowledge, incorrect citations, cybersecurity concerns, and the potential for infodemics (25). Besides inaccuracies and misinformation, there may also be risks of over-reliance on AI for medical purposes (31).

In our study, there were no inaccuracies or misinformation present in the AI-prepared content. However completeness and appropriateness of the content received lower ranks.

Additionally, we did not encounter any inherent biases in the AI-generated slides, as the material was derived from reputable sources. Given that no sensitive information was included, privacy and data protection were also not a concern. Furthermore, the author utilizing the AI system had the necessary expertise, which mitigated potential challenges related to implementation. We did not observe any cases of hallucination in the generated content. While there is a possibility of limited knowledge in AI outputs, this was addressed by ensuring that the content was reviewed by an expert prior to use. Additionally, we provided the citations ourselves, as this was part of the study design, further ensuring the reliability of the information presented

Even though the raters did not identify the mentioned problems they claimed that there was a notable distinction in the presentation style. After completing the statistical analysis and being informed about the group assignments the raters remarked that the AI-generated slides appeared more mechanical, whereas the human-generated slides conveyed a friendlier tone.

Another notable observation in our study was that 'relevance' was highest in the AI-made presentation, while 'lack of false information' was highest in the expert-made presentation. The high relevance in the AI presentation likely resulted from its ability to efficiently process and filter large datasets, effectively aligning content with specific objectives or keywords. In contrast, the expert's superior performance in ensuring the lack of false information suggests a strong emphasis on content reliability, which can be attributed to extensive domain knowledge, critical thinking skills, and experience in fact-checking.

Additionally, both AI and expert presentations scored lowest in completeness, indicating challenges faced by each in fully addressing all necessary details. For the AI, this limitation likely arisen from the inherent constraints of its training data and algorithms, which may not capture the depth and breadth of a topic as comprehensively as a human could. On the other hand, the expert's completeness might be affected by assumptions about the audience's prior knowledge particularly since they are first-year residents or by practical constraints such as limited time in educational settings.

Looking forward, advancements in artificial intelligence are likely to address some of the challenges currently faced. As AI systems continue to evolve, there will be an increase in their potential benefits, leading to higher accuracy in their outputs. A research suggests that in the coming decade, AI will become increasingly advanced, enabling healthcare to move away from a one-size-fits-all model toward a more personalized, preventive, and data-driven approach. These changes could improve patient outcomes and clinical experiences while also reducing costs, resulting in a more efficient and tailored healthcare system (30).

Finally, based on our findings, we recommend requiring specialist oversight when using AI-generated material, as expert-prepared presentations outperformed AI-generated one. Additionally, we propose that collaboration between experts and AI could yield more refined results by leveraging their respective strengths.

Integration of AI in Educational Preparation

After completion of the study, we aimed to integrate the AI-generated presentation into the expert-made presentation. To achieve this, the AI-generated slides were meticulously reviewed and some necessary information were selectively included in the expert-made slides. This approach resulted in an enriched content, but also increased the time required for the presentation. Nevertheless, the resulting presentation was stored for use in the upcoming education hour for physiatry residents (19). This experience provided us with a unique opportunity to combine AI and human efforts to achieve better education material and to incorporate AI-generated insights into practical educational settings.

Limitations of the Study

The study has several limitations. First, the number of presentations could have been greater; however, the existing presentation covered the six most important topics related to therapeutic exercises, with each topic scored separately. While increasing the number of raters could have provided more varied feedback, both raters were experienced physiatrists, and their agreement was confirmed. Additionally, while it would have been beneficial for assistant doctors in the residency program to participate in the scoring, this was not feasible since they were in their first year of specialization and lacked sufficient knowledge on the subject. As new learners, they would have likely evaluated aspects such as fluency and clarity rather than the content itself. Lastly, although evaluating a larger number of slides could have enriched the analysis, the presentation was designed to simulate a real training session, which was intended to last approximately 40 minutes.

CONCLUSION

This study demonstrated that ChatGPT-4o can generate educational slide content on therapeutic exercises at a level above moderate for physiatry residents. However, the expert consistently outperformed the AI due to their specialization. This indicates that generative AI tools can be valuable for creating educational materials, but they should complement rather than replace human expertise. Careful integration of AI-generated health education content with professional oversight is essential to ensure the accuracy and appropriateness of the information presented.

There is a continued need for ongoing research and awareness of the practical challenges associated with integrating AI into healthcare education. Future studies should explore strategies to enhance AI's performance in its weaker areas and investigate how collaboration between AI and experts can be optimized to improve educational outcomes.

Financial disclosures: *The authors declared that this study has received no financial support.*

Conflict of interest: *The authors have no conflicts of interest to declare.*

Ethical approval: *Since the study did not involve human participation and centered on engagements with conversational AI systems, ethical approval was not necessary.*

REFERENCES

1. World Federation for Medical Education. WFME standards for postgraduate medical education, the 2023 revision: https://wfme.org/standards/pgme/ access date 01.09.2024.
2. Sierocinski E, Mathias L, Freyer Martins Pereira J, Chenot JF. Postgraduate medical training in Germany: a narrative review. GMS J Med Educ. 2022;39:Doc49.
3. Kayhan Z. Teaching our students, our residents, and ourselves. Turk J Anaesthesiol Reanim. 2014;42:1-5.
4. Yüksek Öğretim Kurulu. Tıpta Uzmanlık Tüzüğü (Mülga). https://www.yok.gov.tr/Sayfalar/Kurumsal/mevzuat/tipta_uzmanlik_tuzugu_mulga.aspx access date 01.09.2024.
5. Bielecki JE, Tadi P. Therapeutic exercise. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023.
6. Katwaroo AR, Adesh VS, Lowtan A, Umakanthan S. The diagnostic, therapeutic, and ethical impact of artificial intelligence in modern medicine. Postgrad Med J. 2024;100:289-96.
7. OpenAI. ChatGPT. https://chat.openai.com. access date 01.09.2024.
8. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568.
9. Ying L, Liu Z, Fang H, et al. Text summarization with ChatGPT for drug labeling documents. Drug Discov Today. 2024;29:104018.
10. Yilmaz Muluk S. Enhancing musculoskeletal injection safety: evaluating checklists generated by artificial intelligence and revising the preformed checklist. Cureus. 2024;16:e59708.
11. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: Development study involving a literature review. Interact J Med Res. 2024;13:e54704.
12. Özdemir H, Demirbağ Kabayel D. Range of Motion Exercises, Stretching Exercises and Peripheral Joint Mobilization/Manipulation. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st edition. Ankara: Türkiye Klinikleri; 2019:8-14.
13. Eker Büyükşireci D, Meray J. Muscle perfomance exercises. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st edition. Ankara: Türkiye Klinikleri. 2019:15-20.
14. Kurtaiş Aytür Y. Aerobic exercises. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st edition. Ankara: Türkiye Klinikleri; 2019:21-5.
15. Alp A. Aquatic exercises. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st edition. Ankara: Türkiye Klinikleri; 2019:26-32.
16. Üzümcügil Karapolat H, Akgöl I. Posture and sportive performance. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st edition. Ankara: Türkiye Klinikleri; 2019:33-39.
17. Demirsoy N. Relaxation exercises. In: Durmaz B, ed. Tedavi Edici Egzersizler. 1st ed. Ankara: Türkiye Klinikleri; 2019:40-6.
18. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus. 2023;15:e49373.
19. Yilmaz Muluk S, Altuntas V, Duman Sahin Z. Preparing educational presentations about therapeutic exercises for resident physiatrists (Version 2) [Data set]. Zenodo. Published October 28, 2024. doi:10.5281/zenodo.14003512.
20. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 2023;25:e50638.
21. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. 2nd ed. New York: Wiley; 1999;101-3.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.
23. Kaneda Y, Takita M, Hamaki T, et al. ChatGPT's potential in enhancing physician efficiency: a Japanese case study. Cureus. 2023;15:e48235.
24. Diane A, Gencarelli P Jr, Lee JM Jr, Mittal R. Utilizing ChatGPT to Streamline the Generation of Prior Authorization Letters and Enhance Clerical Workflow in Orthopedic Surgery Practice: A Case Report. Cureus. 2023;15:e49680.
25. Sallam M. ChatGPT utility in healthcare education, research, and practice: A systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11:887.
26. Munaf U, Ul-Haque I, Arif TB. ChatGPT: A helpful tool for resident physicians?. Acad Med. 2023;98:868-9.
27. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: Potential impact and opportunity. Acad Med. 2024;99:22-7.
28. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res. 2023;15:4820-8.
29. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. 2024;245:108013.
30. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8:e188-e194.
31. Kleesiek J, Wu Y, Stiglic G, et al. An opinion on ChatGPT in healthcare: Written by humans only. J Nucl Med. 2023;64:701-3.