EUROPEAN JOURNAL OF TRANSLATIONAL
AND CLINICAL MEDICINE 2026;9(2):00-00

RESEARCH ARTICLE

Can ChatGPT support patient–physician communication regarding spine disorders?

Sara Kierońska-Siwak[1], Grzegorz Meder[2], Andreas Yannopoulos[3], Maciej Dzierżanowski[4], Jakub Wiśniewski[5], Marcin Rudaś[6], Magdalena Julia Jabłońska[6,7], Oskar Puk[6,7], Dariusz Grzanka[1]

[1] Department of Clinical Pathomorphology, Nicolaus Copernicus University, Bydgoszcz, Poland
[2] Department of Interventional Radiology, Jan Biziel University Hospital No. 2, Bydgoszcz, Poland
[3] Department of Neurosurgery, General University Hospital of Heraklion, Crete, Greece
[4] Department of Physiotherapy, Nicolaus Copernicus University, Bydgoszcz, Poland
[5] Department of Neurosurgery, Copernicus Hospital, Gdańsk, Poland
[6] Department of Neurosurgery, Stereotactic and Functional Neurosurgery, Jan Biziel University Hospital No 2, Bydgoszcz, Poland
[7] Doctoral School of Medical and Health Sciences, Nicolaus Copernicus University, Bydgoszcz, Poland

Abstract

Introduction: The growing use of large language models (LLMs) such as ChatGPT is transforming the way patients and physicians access and discuss medical information. However, little is known about whether AI-generated responses can effectively foster mutual understanding and improve communication between patients and healthcare professionals, particularly in spine conditions where patients' health literacy is crucial. The aim of this study was to assess how patients and physicians perceive the clarity, usefulness and credibility of ChatGPT-generated responses to common questions about lumbar disc herniation. **Material and methods:** This study involved 70 participants (50 patients and 20 physicians) who assessed ChatGPT responses to 7 standardized questions about back pain and spinal surgery. Each response was rated on a five-point Likert scale for satisfaction, comprehensibility and usefulness. **Results:** ChatGPT responses were generally rated positively, particularly for understandability and usefulness (mean > 4/5). Patients rated ChatGPT significantly higher than physicians ($p < 0.05$), whereas physicians in training showed greater acceptance than experienced specialists. Older participants and those with a history of spine surgery found ChatGPT responses more supportive and practical. **Conclusions:** ChatGPT shows potential as a tool supporting patient-physician communication in spine conditions. Although physicians remain cautious about the medical accuracy of the generated content, patients appreciate its clarity and accessibility. Integrating AI-generated content into clinical communication can help reduce the knowledge and expectations gap between patients and healthcare professionals.

Keywords: ChatGPT · artificial intelligence · patient communication · lumbar discopathy · health education

Corresponding author:
Sara Kierońska-Siwak, Department of Clinical Pathomorphology, Nicolaus Copernicus University, Bydgoszcz, Poland;
e-mail: sara.kieronska@gmail.com
Available online: ejtcm.gumed.edu.pl

Citation

Kierońska-Siwak S, Meder G, Yannopoulos A, Dzierżanowski M, Wiśniewski J, Rudaś M, Jabłońska MJ, Puk O, Grzanka D. Can ChatGPT support patient–physician communication in spine disorders? Eur J Transl Clin Med. 2026;9(2):00-00.

DOI: 10.31373/ejtcm/225829

Abbreviations

- AI – artificial intelligence
- GPT – generative pre-trained transformer
- LLM – large language model
- p – p-value (probability value)
- Q1–Q7 – survey questions 1 to 7
- SD – standard deviation

Introduction

In recent years, large language models (LLMs), such as OpenAI's ChatGPT, have rapidly gained attention across a wide range of domains, including healthcare. These tools are capable of generating natural language responses to complex queries, offering instant access to information, explanations, and recommendations. Given the increasing demand for accessible medical knowledge and the growing reliance on digital solutions, the integration of AI-based systems in health communication presents both promising opportunities and serious challenges [1-2].

Although ChatGPT is not a certified medical tool, its use among the general public and health professionals alike has become increasingly common. Patients may turn to such models for preliminary information, second opinions, or to clarify medical terminology [3-4].

Previous research has highlighted the potential of AI chatbots to support healthcare by improving communication, enhancing patient education, and reducing barriers to access. However, concerns remain regarding the factual reliability of AI-generated information, the ethical implications of unsupervised use, and the risk of misinterpretation by users without medical training. These concerns underscore the need for empirical evaluation of AI tools in real-world health scenarios, especially from the perspective of both healthcare providers and patients [5].

The present study aims to fill this gap by systematically assessing how ChatGPT's responses to medical questions are perceived in terms of satisfaction, clarity, helpfulness, and clinical accuracy. The study further examined whether these perceptions differ across respondent groups (patients vs. physicians, specialists vs. residents) and explored the influence of demographic factors (e.g. age, sex and medical history) on the evaluation of AI-generated responses.

Material and methods

Patients and physicians were asked the same questions in surveys regarding low back pain, lumbar disc disease, and degenerative disease of the lumbar spine. All responses assessed in this study were generated using the GPT-4 version of the ChatGPT (OpenAI, San Francisco, USA) model, obtained via the official website.

Inclusion and exclusion criteria

Patients were eligible for inclusion if they were ≥ 18 years of age or older, had a history of low back pain and/or degenerative lumbar spine disease, were able to read and understand the questionnaire written in Polish language, provided informed consent to participate in the study, and completed all survey questions. Patients were excluded if they were < 18 years of age, submitted incomplete questionnaires, declined participation or withdrew consent during the study, had cognitive impairment or any other condition (somatic or psychiatric) that limited their ability to understand and interpret the survey content.

Each participant provided informed, written consent to participate in this study. The survey was completely anonymous and did not contain any personal or sensitive information. Ethical approval for this questionnaire-based study was obtained from the Bioethics Committee of the Collegium Medicum, Nicolaus Copernicus University in Bydgoszcz (KEWL No. 35/2026). The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

The survey

A survey was conducted on paper and electronically using the Forms software (Google, Mountain View, CA, USA). The 7 questions used in this study were developed based on

clinical experience and observations of the questions most frequently asked by patients during outpatient visits related to back pain and degenerative diseases of the lumbar spine. The final set of questions was reviewed and selected by neurosurgery specialists to ensure its high substantive value and relevance to daily clinical practice. The 7 questions addressed the causes, symptoms, treatment options, indications for surgery, potential complications, and management of worsening symptoms in lumbar spine disorders. The complete questionnaire and ChatGPT-generated responses are provided in the Supplementary Material.

Survey questions

Q1. What are the causes of back pain?
Q2. What symptoms require medical consultation and how urgently should a doctor be consulted?
Q3. What are the types of lumbar spine surgery?
Q4. What are the complications of lumbar spine surgery?
Q5. What are the treatment options for lower back pain?
Q6. What are the indications for lumbar spine surgery?
Q7. My back pain is getting worse. What should I do?

Statistical analyses

Statistical analyses were performed using IBM SPSS Statistics software (Armonk, NY, USA). Continuous variables were described using means, standard deviations and medians. Normality of distribution was assessed using the Shapiro-Wilk test. Comparisons between groups were performed using the Mann-Whitney U test. Associations between categorical variables were analyzed using the chi-square test. Statistical significance was set at $p < 0.05$.

Results

Patients

The patient group consisted of 50 individuals (68% female and 32% male). The largest age group included patients aged 46-60 (30%), followed by those aged 18-30 (38%), 31-45 (22%), and the smallest group were individuals over 60 years old (10%). Majority of the patients (62%) had higher education, whereas 22% had secondary education and 16% had vocational education. A total of 14% of patients (n = 7) reported having undergone spine surgery.

Physicians

This study included 20 physicians, of whom 15 were neurosurgery specialists and 5 were in neurosurgery residency training. The physicians represented a more homogeneous group in terms of education, but varied in their levels of clinical experience.

Overall evaluation of ChatGPT responses

Statistical analysis showed that ChatGPT's responses were generally evaluated positively by all study participants. The highest average scores were recorded for clarity (mean: 4.15; SD: 0.91) and helpfulness defined as the usefulness of the response in addressing the question (mean: 4.05; SD: 0.76) as shown in Table 1. All variables significantly deviated from a normal distribution (Shapiro–Wilk test, $p < 0.01$), though the distributions were approximately symmetric, with slight left skewness and kurtosis values indicating moderate flattening (Table 1).

I. Patients vs. physicians

Statistically significant differences were found between patients and physicians. Patients rated ChatGPT responses higher than physicians, particularly in terms of clarity (Table 2 and Figure 1).

II. Specialists vs. residents

Within the physician group, a trend emerged showing that residents gave higher ratings to ChatGPT responses. The most notable difference concerned Q4 – helpfulness ($p = 0.020$) (Table 3).

Influence of age, sex and medical history

Participants over 45 years old were more likely to report that ChatGPT suggested next steps, and gave higher ratings for satisfaction and helpfulness ($p < 0.05$ for multiple variables, e.g. Q2, Q3, Q5, Q6) (Table 4). Most variables showed no statistically significant differences. However, women rated Q4 – satisfaction higher than men ($p = 0.037$). Participants who had undergone spinal surgery were more likely to state that ChatGPT suggested further action (Q6, $p = 0.025$) and were more willing to continue consultations with neurosurgeon.

Findings from Chi-Square Tests of Independence

Chi-square analyses revealed no significant overall differences in declarations of willingness to continue consultations

or in perceptions of whether ChatGPT suggested further action, except for the following:

Q2 – Sex: women more frequently reported that ChatGPT indicated further action ($p = 0.039$).

Q3 – Sex: men more frequently than women declared that ChatGPT partially suggested which procedure is indicated for the particular diagnosis ($p = 0.004$).

Discussion

Overall, our findings indicate a positive reception of ChatGPT's responses, particularly in terms of clarity and helpfulness. Notably, patients rated its responses significantly higher than physicians, particularly regarding claritiy ($p < 0.01$) and helpfulness ($p < 0.05$). This suggests that ChatGPT may serve as an effective educational and informational tool for individuals without formal medical training. However, physicians evaluated the factual accuracy of the responses as moderately high, highlighting the need for ongoing validation of LLM outputs against current medical standards.

A significant finding is that neurosurgery residents in training tended to assess the responses as more helpful compared to specialists. This may reflect a greater openness to emerging technologies among younger clinicians. Furthermore, among patients, older age and medical history (e.g. past spinal surgery) were associated with higher ratings of response utility and more frequent reports that ChatGPT suggested appropriate next steps [6].

While no statistically significant differences were found between the studied groups regarding whether ChatGPT suggested further medical action, trends emerged indicating the patients' greater acceptance of that LLM. Importantly, most variables did not show sex-based differences, although isolated findings (e.g., for Q2 and Q3) suggest that individual characteristics may still influence perception.

The rapid development of LLMs has opened new possibilities for patient education and health communication. These models have significant potential to provide rapid and accessible medical information to large numbers of patients. A review of recent literature reveals both promising applications and important limitations of ChatGPT, particularly in the context of spine-related disorders [7-9]. Several studies have evaluated the accuracy, readability, and clinical relevance of ChatGPT's responses to commonly asked patient questions. For example, ChatGPT-3.5 provided generally correct but overly technical responses to questions regarding pediatric scoliosis [10]. Similar results were found in relation to cervical spine surgery and lumbar disc herniation: the LLM demonstrated moderate accuracy (up to 70%) but struggled to adjust the complexity of its language to the average patient's

Table 1. Average ratings of ChatGPT responses by all participants (n = 70)

	Mean	SD	Median	Range	Skewness	s	Shapiro–Wilk (p)
Satisfaction	3.94	0.81	4	2-5	-0.62	0.24	< 0.01
Clarity	4.15	0.91	4	2-5	-0.83	-0.13	< 0.01
Helpfulness	3.74	0.94	4	1-5	-0.61	0.12	< 0.01

Table 2. Comparison of ratings: patients vs. physicians (Mann–Whitney U Test)

	Group medians	p-value	Statistical significance
Q2 – Clarity	Patients: 5.00 Physicians: 4.00	0.001	Yes ($p < 0.01$)
Q4 – Clarity	Patients: 4.00 Physicians: 4.00	0.004	Yes ($p < 0.01$)
Q5 – Helpfulness	Patients: 4.00 Physicians: 4.00	0.039	Yes ($p < 0.05$)

Table 3. Physicians' evaluation of ChatGPT responses, by level of training

Question	Median – residents	Median – specialists	p-value	Statistical significance
Q4 – Helpfulness	4.5	4.0	0.020	Yes ($p < 0.05$)

Table 4. Age differences in satisfaction ratings

	Median ≤ 45 yrs	Median > 45 yrs	p-value	Statistical significance
Q2 – Satisfaction	4.00	5.00	0.006	Yes ($p < 0.01$)

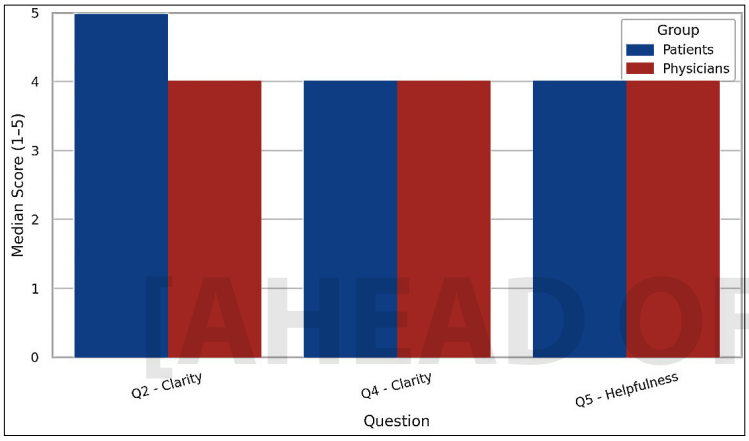

Figure 1. Comparison of median ChatGPT response ratings by patients and physicians for selected questions

reading level [11]. This highlights a significant gap between technical accuracy and practical accessibility [12].

Interestingly, attempts to simplify answers to a sixth-grade reading level resulted in only modest improvements in clarity, suggesting limitations in ChatGPT's ability to effectively adapt communication for populations with lower health literacy. This is particularly important, as effective patient education requires not only factual accuracy but also clarity, empathy, and personalization [13].

In clinical simulations, ChatGPT showed inconsistent performance in generating differential diagnoses and recommending treatment. In a comparative study involving spine surgeons, the model misinterpreted traumatic cases and proposed management strategies that were judged clinically inappropriate by the participating specialists in more than 50% of the evaluated scenarios [12]. These findings indicate that while ChatGPT may serve as a supportive educational tool, it certainly cannot replace clinical judgment by a medical professional. On the other hand, the model showed utility in research-related tasks, such as generating topics for systematic reviews or summarizing clinical guidelines, particularly regarding low back pain and spine surgery. This suggests a complementary role for LLMs in medical education and academic support [13-14].

Additionally, other studies have shown that ChatGPT-4 can provide reliable and understandable responses to patient questions related to spinal cord stimulation, highlighting the potential of more advanced versions of LLMs [15]. However, even these models require further clinical validation and patient-specific adaptation [16]. Future research should explore real-world interactions between patients and LLMs, evaluating outcomes such as comprehension, trust, behavioral impact, and emotional support. New models (e.g. GPT-4o) offer enhanced multimodal and conversational abilities, which warrant continuous, systematic evaluation regarding their potential in healthcare contexts [17].

Limitations and clinical implications

This study has several important limitations that should be considered when interpreting its results. First, the physician sample size was limited (n = 20), with only 5 of those physicians in residency training. This sample structure may impact the representativeness of the data and limit the generalizability of the results to the broader healthcare professional population. Furthermore, participants were recruited voluntarily, which carries the risk of self-selection bias: individuals more interested in new technologies or using digital tools may have been overrepresented in the sample, potentially inflating the overall assessment of the usefulness and understandability of ChatGPT responses. Another limitation is that this study was perceptual: participants rated

responses based on subjective impressions (clarity, helpfulness, satisfaction), but there was no independent assessment of the responses' compliance with current clinical guidelines or standards of medical practice. Furthermore, the study did not compare the responses generated by ChatGPT with other popular information sources, e.g. search engines, Wikipedia, traditional medical websites or advice provided by physicians. Finally, a single version (GPT-4) of the ChatGPT LLM was used in this study. These models are subject to dynamic changes, updates, and optimization, so responses generated in the future may differ from those evaluated in this study. The variability of the LLMs over time is a challenge to the repeatability and comparability of results across studies and time [18-19].

Conclusions

ChatGPT responses were generally well-rated by both physicians and patients, particularly for clarity and helpfulness. Patients provided significantly higher ratings than physicians, highlighting ChatGPT's potential as an educational support tool in patient–physician interactions. Physicians in training were more receptive to ChatGPT's outputs than experienced specialists. Age and prior medical experiences (e.g. past surgery) positively influenced perceptions of satisfaction and usefulness of the chatbot responses. Medical accuracy of ChatGPT's responses requires further verification, as physicians' evaluations were moderately positive but not conclusive.

Funding

No external funding was received for this research.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

The datasets generated and analyzed during the current study are not publicly available due to patient privacy and ethical restrictions but are available from the corresponding author on reasonable request.

References

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 [Internet]. 2025;32(1):45. Available from: https://doi.org/10.1186/s12929-025-01131-z
2. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med [Internet]. 2025;5(1):26. Available from: https://doi.org/10.1038/s43856-024-00717-2
3. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res [Internet]. 2023;25:e47479. Available from: https://www.jmir.org/2023/1/e47479
4. Lu L, Zhu Y, Yang J, Yang Y, Ye J, Ai S, et al. Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. Sci Rep [Internet]. 2025;15(1):1223. Available from: https://doi.org/10.1038/s41598-024-84512-y
5. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform [Internet]. 2024;25(1):bbad493. Available from: https://doi.org/10.1093/bib/bbad493
6. Brückner S, Brightwell C, Gilbert S. FDA launches health care at home initiative to drive equity in digital medical care. npj Digit Med [Internet]. 2024;7(1):204. Available from: https://doi.org/10.1038/s41746-024-01198-2
7. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med [Internet]. 2024;11. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2024.1477898/full
8. Lang SP, Yoseph ET, Gonzalez-Suarez AD, Kim R, Fatemi P, Wagner K, et al. Analyzing Large Language Models' Responses to Common Lumbar Spine Fusion Surgery Questions: A Comparison Between ChatGPT and Bard. Neurospine [Internet]. 2024;21(2):633–41. Available from: http://e-neurospine.org/journal/view.php?doi=10.14245/ns.2448098.049

9. Zhao Y-F, Bove A, Thompson D, Hill J, Xu Y, Ren Y, et al. Generative AI Is Not Ready for Clinical Use in Patient Education for Lower Back Pain Patients, Even With Retrieval-Augmented Generation. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]. 2025;2025:644–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/40502233
10. Lieu B, Crawford E, Laubach L, Yeramosu T, Sharps C, Horstmann J, et al. Patient education strategies in pediatric orthopaedics: using ChatGPT to answer frequently asked questions on scoliosis. Spine Deform [Internet]. 2025;13(5):1377–89. Available from: https://link.springer.com/10.1007/s43390-025-01087-y
11. Subramanian T, Araghi K, Amen TB, Kaidi A, Sosa B, Shahi P, et al. Chat Generative Pretraining Transformer Answers Patient-focused Questions in Cervical Spine Surgery. Clin Spine Surg [Internet]. 2024;37(6):E278–81. Available from: https://journals.lww.com/10.1097/BSD.0000000000001600
12. Chalhoub R, Mouawad A, Aoun M, Daher M, El-sett P, Kreichati G, et al. Will ChatGPT be Able to Replace a Spine Surgeon in the Clinical Setting? World Neurosurg [Internet]. 2024;185:e648–52. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1878875024003140
13. Shrestha N, Shen Z, Zaidat B, Duey AH, Tang JE, Ahmed W, et al. Performance of ChatGPT on NASS Clinical Guidelines for the Diagnosis and Treatment of Low Back Pain. Spine (Phila Pa 1976) [Internet]. 2024;49(9):640–51. Available from: https://journals.lww.com/10.1097/BRS.0000000000004915
14. Herzog I, Mendiratta D, Para A, Berg A, Kaushal N, Vives M. Assessing the potential role of ChatGPT in spine surgery research. J Exp Orthop [Internet]. 2024;11(3). Available from: https://esskajournals.onlinelibrary.wiley.com/doi/10.1002/jeo2.12057
15. Lo Bianco G, Cascella M, Li S, Day M, Kapural L, Robinson CL, et al. Reliability, Accuracy, and Comprehensibility of AI-Based Responses to Common Patient Questions Regarding Spinal Cord Stimulation. J Clin Med [Internet]. 2025;14(5):1453. Available from: https://www.mdpi.com/2077-0383/14/5/1453
16. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C, et al. Large language models: Are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J [Internet]. 2024;33(11):4135–43. Available from: https://doi.org/10.1007/s00586-023-07975-z
17. Wang S, Wang Y, Jiang L, Chang Y, Zhang S, Zhao K, et al. Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. Eur J Med Res [Internet]. 2025;30(1):45. Available from: https://doi.org/10.1186/s40001-025-02296-x
18. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv Prepr arXiv230313375 [Internet]. 2023; Available from: https://arxiv.org/abs/2303.13375
19. Sakaguchi K, Sakama R, Watari T. Evaluating ChatGPT in Qualitative Thematic Analysis With Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study. J Med Internet Res [Internet]. 2025;27:e71521. Available from: https://www.jmir.org/2025/1/e71521