JOMDI

Journal of Medical and Dental Investigations

J Med Dent Invest 2025; 6: e250131
https://doi.org/10.5577/jomdi.e250131

Original Article Open Access

Evaluation of the usability of ChatGPT-4o and Gemini 2.5 Pro in patient education about lumbar disc herniation

Umut Ogün Mutlucan[1], **Ökkeş Zortuk**[2]

[1] University of Health Sciences, Antalya Research and Training Hospital, Department of Neurosurgery, Antalya, Türkiye
[2] Defne State Hospital, Department of Emergency Medicine, Hatay, Türkiye

Received: 6 November 2025
Revised: 26 November 2025
Accepted: 10 December 2025
Published: 23 December 2025

Correspondence:

Dr. Umut Ogün MUTLUCAN
SBU Antalya Research and Training Hospital, Department of Neurosurgery, Antalya, Türkiye.
E-mail: umutlucan@gmail.com

How to cite this article:

Mutlucan UO, Zortuk Ö. Evaluation of the usability of ChatGPT-4o and Gemini 2.5 Pro in patient education about lumbar disc herniation. J Med Dent Invest 2025;6:e250131. https://doi.org/10.5577/jomdi.e250131

Abstract

Aim: This study aimed to evaluate the efficacy and reliability of ChatGPT-4o and Gemini-2.5 Pro in providing patient education on lumbar disc herniation (LDH) by assessing their responses to frequently asked questions.

Methodology: A list of 35 common patient questions about LDH and its treatment was compiled from neurosurgery and orthopedic websites. These questions were translated into Turkish and posed to both ChatGPT-4o and Gemini-2.5 Pro. Two independent neurosurgeons evaluated the responses using a 5-point scoring system (5 = accurate and understandable; 1 = incorrect/irrelevant). Discrepancies were resolved through joint review and consensus. Statistical analysis, including paired t-tests, was conducted using SPSS ($p < 0.05$).

Results: Both ChatGPT and Gemini showed similar overall performance with an average score of 4.74. While ChatGPT significantly outperformed Gemini in the "Exercise and Physical Activity" category ($p = 0.039$), there were no significant differences in other categories, such as "Treatment," "Symptoms," "Surgical Treatment and Recommendations," or "Diagnosis". The findings suggest that large language models hold promise for patient information and guidance, but neither demonstrated a substantial overall superiority in educating patients about LDH.

Conclusion: The findings suggest that large language models such as ChatGPT and Google Gemini show considerable promise for providing information and guidance to patients. While ChatGPT exhibited a significant advantage in the "Exercise and Physical Activity" category, overall, neither model demonstrated substantial superiority in educating patients about LDH. Further development is necessary to ensure accuracy and reliability for medical applications.

Keywords: ChatGPT, Gemini, patient education, lumbar disc herniation

Introduction

Lumbar disc herniation (LDH) is a prevalent condition that can lead to significant pain and functional impairment, often requiring medical intervention. The condition is characterized by displacement of disc material beyond the intervertebral disc space, which can compress nerve roots and cause radiculopathy (1). Treatment options range from conservative management to surgical interventions, depending on the severity and progression of symptoms. The popularity of artificial intelligence (AI) and related technologies is increasing daily across medical science and other disciplines. One potential application of AI is in medical technology. The issue of access to patient information and its contributions to health services is a complex one. In recent years, numerous artificial intelligence models have emerged, with ChatGPT (Chat Generative Pre-Trained Transformer) and Google Gemini being the most prominent. These database-based chat programs, which are offered free of charge, facilitate access to information (1, 2).

The responses to these inquiries regarding diseases are derived from internet databases, potentially resulting in inadequate patient education or misinformation (2, 3).

The integration of AI technologies, including ChatGPT and deep learning models, within medical disciplines such as lumbar disc herniation diagnosis and treatment is gaining traction. ChatGPT, a language model, has demonstrated the capacity to provide medical information and treatment options; however, further development is necessary to ensure the model's accuracy and reliability (3, 4). Concurrently, deep learning models have been employed with considerable efficacy in the diagnosis of lumbar disc herniation from MRI scans, exhibiting a high degree of accuracy in detection and classification tasks. These advancements underscore the potential of AI to enhance medical diagnostics and patient care (4, 5).

ChatGPT has been utilized to provide medical information and treatment options for various conditions, including shoulder impingement syndrome. This has enabled the platform to offer insights into symptoms, risk factors, and treatment exercises. However, it is imperative to exercise caution when interpreting the findings, given the potential for inherent biases and inaccuracies in the information provided. In clinical practice, ChatGPT has been demonstrated to support decision-making by generating differential diagnosis lists and optimizing clinical decision support (5, 6). Nevertheless, it has not yet been established as a substitute for expert opinion. Deep learning models, including those employing convolutional neural networks (CNNs), have been developed for the diagnosis of lumbar disc herniation from MRI scans. The models in question have been shown to achieve high accuracy, with one study reporting 95.65% accuracy in detecting herniation types and 91.38% in recognizing them (7).

Despite the considerable potential of artificial intelligence (AI) technologies, such as ChatGPT and deep learning models, in medical applications, challenges remain. It is imperative to exercise caution and conduct meticulous scrutiny of ChatGPT's potential biases and inaccuracies, particularly in clinical settings. In a similar vein, the use of deep learning models depends on the availability of large datasets and rigorous validation to ensure reliability (8). These technologies have the potential to serve as valuable tools to complement the efforts of healthcare professionals; however, they cannot yet be considered substitutes for expert judgment (7-9).

Patients frequently use these artificial intelligence models to access more information. However, the effectiveness and reliability of these artificial intelligence models remain to be fully elucidated. The objective of this study was to evaluate the efficacy of ChatGPT-4o and Gemini 2.5 Pro by assessing the responses of neurosurgery specialists to patient inquiries about lumbar disc herniation.

Materials and Methods

The present study was conducted using artificial intelligence programs, and the authors have no relationship with the AI companies or sites associated with the study. A comprehensive web search was conducted using the query "Frequently Asked Questions About Lumbar Disc Herniation" to identify frequently asked questions (FAQs) from patients and their relatives regarding lumbar disc herniation and its treatment. Relevant content was retrieved from websites related to neurosurgery and orthopedics, including those of medical clinics, professional associations, and patient-oriented health forums.

The authors then conducted a thorough review of the collected questions, excluding those that were repetitive, irrelevant, non-medical, or lacked meaningful content. A comprehensive list of the 35 most frequently asked and relevant questions was subsequently identified. The aforementioned inquiries underwent grammatical refinement and were subsequently classified into the following categories: (1) Treatment, (2) Symptoms and Pain Management, (3) Exercise and Physical Activity, (4) Surgical and Other Interventional Methods, (5) Diagnosis, and (6) General Topics.

Each of the finalized questions was translated into Turkish and presented individually to two artificial intelligence models: ChatGPT-4o and Gemini-2.5 Pro. Responses were obtained in Turkish and subsequently evaluated by two independent neurosurgeon authors. The following scoring system was employed:

- **5 points** for accurate answers understandable to the general public,
- **4 points** for academically correct answers,

- **3 points** for correct but incomplete answers,
- **2 points** for partially correct but misleading or contradictory answers,
- **1 point** for completely incorrect or irrelevant answers.

The responses from both AI models were evaluated independently by each neurosurgeon. In the event that both evaluators assigned the same score to a given answer, the score was accepted. In instances where there was a discrepancy in the scores, the authors conducted a joint review and discussion of the response to reach a consensus score, which was subsequently documented as the final evaluation.

A graphical illustration of the study is shown in Figure 1.

Statistical analysis

The statistical analyses were conducted using SPSS software for Windows (Version 27.0, IBM Inc., Armonk, NY, USA).

Categorical data were defined as percentages and frequencies. The numerical data were defined, and a distribution analysis was performed. Data that conformed to the normal distribution were defined as mean ± standard deviation (SD). The interobserver test was used to assess the consistency between raters in the study. In the analysis of the relationship between numerical variables, the paired t-test was used to assess whether the data were normally distributed. Data with a p-value below 0.05 were considered to be significant.

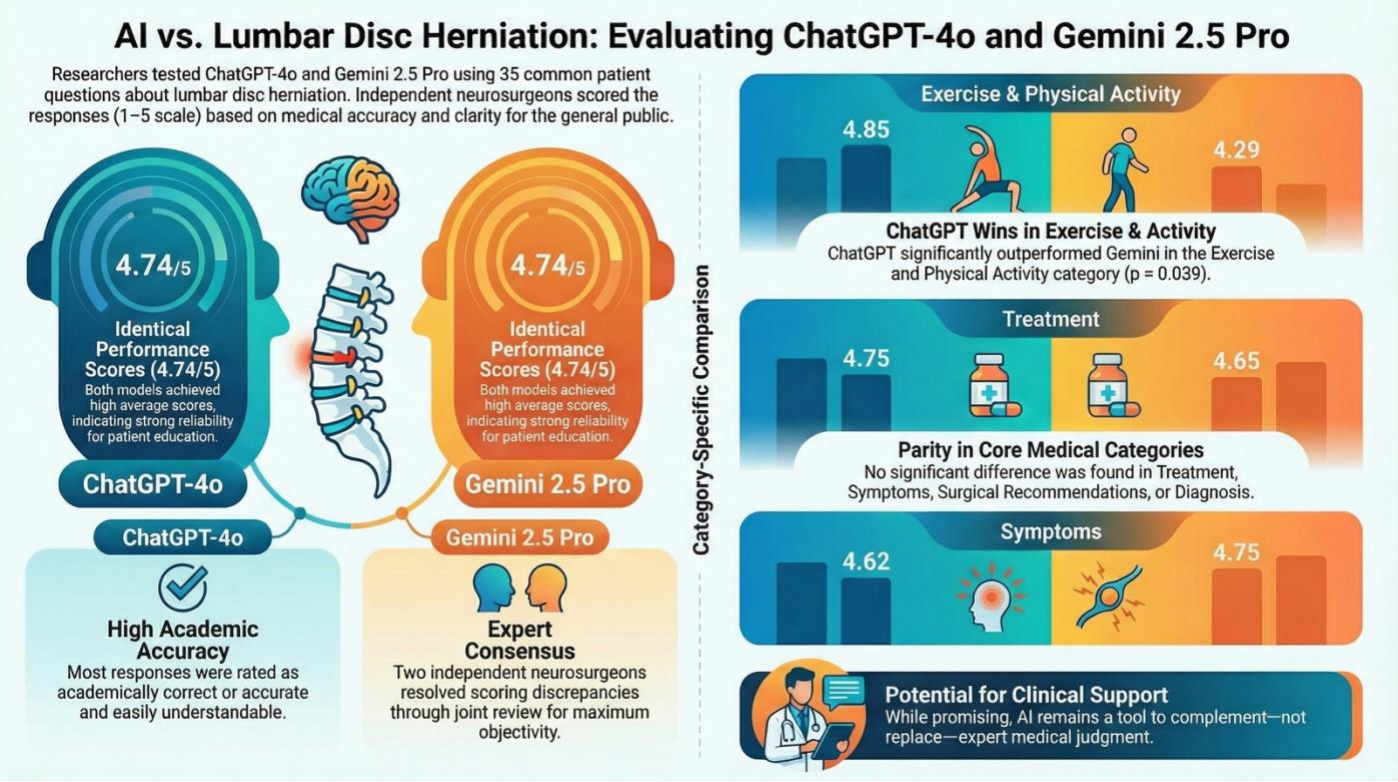

Figure 1. Graphical illustration of the study.

Results

In the study, the answers provided by LLMs were evaluated by two different experts. The evaluation analysis performed by the experts is presented in Table 1.

In the comparison for treatment, the mean score for GPT was 4.75 ± 0.46, while the mean score for GEMINI was 4.65 ± 0.74, and there was no significant difference between them (p = 0.342). In the comparison of symptoms, the mean score for GPT was 4.62 ± 0.51, while the mean score for GEMINI was 4.75 ± 0.46, and no significant difference was observed (p = 0.299). In the comparison of exercise and physical activity, the mean GPT score was 4.85 ± 0.37, while the mean GEMINI score was 4.29 ± 0.53. The GPT score was significantly higher (p = 0.039). For surgical treatment and treatment recommendations, the mean GPT score was 4.85 ± 0.37, and the mean GEMINI score was 4.85 ± 0.38, with no significant difference (p = 0.500). In the comparison of diagnoses, the mean GPT score was 4.60 ± 0.55, while the mean GEMINI score was 5.00 ± 0.00. A non-significant difference was observed between them (p = 0.089). Figure 2 presents the comparisons according to question groups.

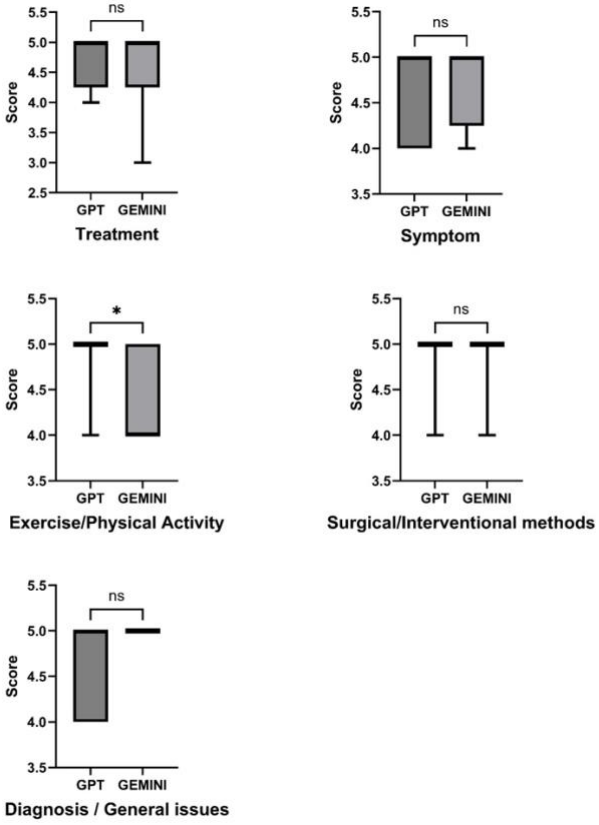

Figure 2. Comparison by question groups.

A comparison of the mean scores revealed that GPT and GEMINI exhibited similar performance, with mean scores of 4.74 and 4.74, respectively. This observation did not demonstrate a statistically significant difference between the two groups (p = 0.400). Figure 3 presents a comparison of the responses to all questions.

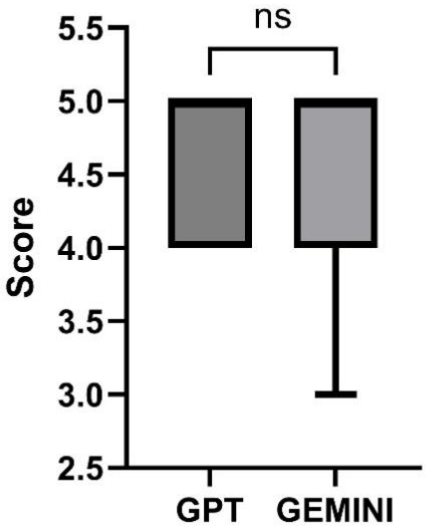

Figure 3. Comparison by overall score.

Table 1. Interobserver correlation.

	Mean ± SD	Intraclass Correlation	p-value
GPT Observer 1	4.74 ± 0.44	0.012	0.486
GPT Observer 2	4.34 ± 0.80		
GEMINI Observer 1	4.71 ± 0.51	-0.541	0.894
GEMINI Observer 2	4.82 ± 0.45		

Discussion

The inquiry pertains to the capabilities of AI models, particularly ChatGPT and Google Gemini, in the context of LDH, and their overall performance in medical and legal domains (9, 10). ChatGPT-4.0 has demonstrated its capacity to provide comprehensive clinical support in the diagnosis and treatment of LDH, exhibiting high levels of accuracy and consistency in its responses. However, concerns regarding readability persist due to the utilization of complex language. In broader medical applications, ChatGPT has been shown to consistently outperform Google Gemini in terms of accuracy and the quality of its responses across a range of medical domains, including glaucoma surgery, neuroradiology, and peripheral artery disease (10, 11). This finding suggests that ChatGPT possesses a high degree of robustness in medical diagnostics and planning.

A comparative analysis of ChatGPT-4.0 and its miniature counterpart revealed their proficiency in managing LDH, exhibiting substantial clinical support capabilities with optimal accuracy and safety metrics (12). A comparison of the responses provided by ChatGPT-4.0 and its smaller version shows that ChatGPT-4.0 produces more comprehensive responses. The model performed well in image analysis for LDH, with precision, recall, and F1 scores exceeding 0.80. However, the content was difficult for patients to understand. Regarding glaucoma surgical planning, ChatGPT-4.0 showed greater agreement with specialists than Google Gemini, especially in difficult cases. This is evident in the significantly higher Global Quality Score (GQS) for ChatGPT. In neuroradiology diagnostics, ChatGPT

versions 3.5 and 4.0 outperformed Google Gemini in diagnostic quiz solving, highlighting its potential as a diagnostic tool (12, 13). In the context of peripheral artery disease, ChatGPT demonstrated a higher degree of accuracy and satisfaction in its responses compared to Google Gemini, although the latter exhibited faster response times. In the context of Islamic law, both AI models were evaluated for their ability to provide reliable and understandable responses. ChatGPT was found to generally offer more accurate information. In the domain of pediatric radiology, ChatGPT demonstrated superior accuracy in responding to text-based inquiries when compared to Google Gemini, thereby underscoring its utility in medical education (13, 14). While ChatGPT has demonstrated superior performance in terms of accuracy and quality across various domains, it is noteworthy that Google Gemini has exhibited particular strengths, such as faster response times and perfect accuracy in debunking myths surrounding palliative care. This suggests that while ChatGPT may be more reliable for detailed and complex medical tasks, Google Gemini could be advantageous in scenarios where speed and straightforward accuracy are prioritized (15, 16).

There is a surge in interest and uptake of the integration of AI technologies, including ChatGPT and deep learning models, within medical disciplines such as lumbar disc herniation diagnosis and treatment. ChatGPT, a language model, has demonstrated the capacity to provide medical information and treatment options; however, further development is necessary to ensure the model's accuracy and reliability (17). Concurrently, deep learning models have been successfully implemented to diagnose lumbar disc herniation from MRI scans, exhibiting high precision in detection and classification tasks. These advancements underscore the potential of AI to enhance medical diagnostics and patient care (18, 19). ChatGPT has been employed to furnish medical information and treatment options for conditions such as shoulder impingement syndrome, offering insights into symptoms, risk factors, and treatment exercises. However, it is imperative to exercise caution due to the potential presence of biases and inaccuracies in the information provided (20).

In clinical practice, ChatGPT has been demonstrated to support decision-making by generating differential diagnosis lists and optimizing clinical decision support. However, it has not yet been shown to replace expert opinion. Deep learning models, including those employing convolutional neural networks (CNNs), have been developed for the diagnosis of lumbar disc herniation from MRI scans. These models have demonstrated high accuracy rates; a study reported an accuracy of 95.65% for detection and 91.38% for the recognition of herniation types (21).

The study's limitations are due to the internet-based nature of the chatbot responses. Recent findings have shown that these responses can cause confusion due to erroneous learning algorithms. Despite artificial intelligence companies' efforts to mitigate this issue, it is important to acknowledge the possibility of encountering it in chatbots. In the present study, patient questions were predominantly used and evaluated, as outlined in the appendix. The prospect of using artificial intelligence programs to address patient queries is fascinating.

Conclusion

The findings of the present study demonstrate that large language model (LLM) technologies, including ChatGPT and Google Gemini, exhibit considerable promise in the provision of information and guidance to patients. Furthermore, the investigation revealed that artificial intelligence models do not demonstrate a substantial degree of superiority in the education of patients regarding LDH.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Concept — U.O.M.; Design — U.O.M.; Supervision — Ö.Z.; Materials - U.O.M.; Data collection &/or processing — U.O.M.; Analysis and/or interpretation — Ö.Z.; Literature search — U.O.M., Ö.Z.; Writing — U.O.M.; Critical review —U.O.M., Ö.Z.

Conflict of Interest: There is no any conflict of interest.

Funding: There is no any funding.

References

1. Beinecke JM, Venderink W, Suelmann BBM, Sedelaar JPM, Huisman HJ, Fütterer JJ. Evaluation of machine learning strategies for imaging confirmed prostate cancer recurrence prediction on electronic health records. Comput Biol Med. 2022;143:105263. https://doi.org/10.1016/j.compbiomed.2022.105263
2. Ebrahim M, Al-Ayyoub M, Alsmirat M. Determine bipolar disorder level from patient interviews using bi-lstm and feature fusion. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE; 2018:182-189. https://doi.org/10.1109/SNAMS.2018.8554886

3. Najadat H, Al-Zu'bi S, Al-Zu'bi M. Investigating the classification of human recognition on heterogeneous devices using recurrent neural networks. In: Advances in Intelligent Systems and Computing. Vol 1299. Springer; 2021:67-80. https://doi.org/10.1007/978-3-030-64058-3_7
4. Wu Z, Xuan S, Xie J, Lin C, Lu C. How to ensure the confidentiality of electronic medical records on the cloud: a technical perspective. Comput Biol Med. 2022;147:105726. https://doi.org/10.1016/j.compbiomed.2022.105726
5. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):971-987. https://doi.org/10.1109/TPAMI.2002.1017623
6. Ojala T, Pietikäinen M, Mäenpää T. Gray scale and rotation invariant texture classification with local binary patterns. In Computer Vision-ECCV. 404-420. https://doi.org/10.1007/3-540-45054-8_27
7. Lowe DG. Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. Vol 2. IEEE; 1999:1150-1157. https://doi.org/10.1109/ICCV.1999.790410
8. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025;15(3):123-135. https://doi.org/10.5577/intdentres.662
9. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297. https://doi.org/10.1007/BF00994018
10. Hamilton NA, Pantelic RS, Hanson K, Teasdale RD. Fast automated cell phenotype image classification. BMC Bioinformatics. 2007;8:110. https://doi.org/10.1186/1471-2105-8-110
11. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry. 1998;33(3): 366-375. https://doi.org/10.1002/(SICI)1097-0320(19981101)33:3<366::AID-CYTO12>3.0.CO;2-R
12. Ogink PT, Groot OQ, Bindels BJJ, Tobert DG. The use of machine learning prediction models in spinal surgical outcome: an overview of current development and external validation studies. Semin Spine Surg. 2021;33(2):100872. https://doi.org/10.1016/j.semss.2021.100872
13. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. https://doi.org/10.1136/bmj.g7594
15. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976). 2000;25(22):2940-2952. https://doi.org/10.1097/00007632-200011150-00017
16. Von Korff M, Jensen MP, Karoly P. Assessing global pain severity by self-report in clinical and health services research. Spine (Phila Pa 1976). 2000;25(24):3140-3151. https://doi.org/10.1097/00007632-200012150-00009
17. Terluin B, Eekhout I, Terwee CB, de Vet HC. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol. 2015;68(12):1388-1396. https://doi.org/10.1016/j.jclinepi.2015.03.015
18. Terluin B, Eekhout I, Terwee CB. The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. J Clin Epidemiol. 2017;83:90-100. https://doi.org/10.1016/j.jclinepi.2016.12.015
19. Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol. 2010;63(7):760-766.e1. https://doi.org/10.1016/j.jclinepi.2009.09.009
20. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. https://doi.org/10.1136/bmj.m441
21. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14(1):25. https://doi.org/10.1186/1471-2288-14-25