

Can Artificial Intelligence-Based Large Language Models Provide Accurate and Reliable Information to Asthma Patients? A Comparative Analysis with Expert Insights

Sevgi COLAK^{1,2} , Pamir CERCI³ , Betül DUMANOGU⁴ , Ozge CAN BOSTAN⁵ 

¹ Department of Internal Medicine, Division of Immunology and Allergy, Ankara University School of Medicine, Ankara, Türkiye

² Department of Immunology and Allergy, Ankara Etlik City Hospital, Ankara, Türkiye

³ Department of Internal Medicine, Division of Allergy and Immunology, Eskişehir City Hospital, Eskişehir, Türkiye

⁴ Department of Immunology and Allergy, Ağrı Training and Research Hospital, Ağrı, Türkiye

⁵ Department of Allergy and Immunology, Çanakkale Mehmet Akif Ersoy State Hospital, Çanakkale, Türkiye

Corresponding Author: Sevgi Colak ✉ drsevgicolak@gmail.com

This work has been accepted for presentation at the Thematic Poster Session (TPS) at the EAACI Congress 2025, to be held in Glasgow, UK

ABSTRACT

Objective: Providing patients with accurate and reliable information significantly improves their quality of life and reduces the burden on healthcare services. Recent advancements in artificial intelligence (AI), particularly Large Language Models (LLMs), offer opportunities for enhanced patient education. This study aimed to compare an asthma patient information text generated by an LLM with one provided by the Turkish National Society of Allergy and Clinical Immunology (TNSACI), as evaluated by specialized physicians.

Materials and Methods: Physicians with a minimum of five years of experience in Allergy and Immunology were recruited to assess blinded versions of two asthma educational texts: one generated by Chat Generative Pre-trained Transformer (ChatGPT) and the other sourced from TNSACI's website. Participants evaluated the texts using a Likert scale, assessing accuracy, comprehensiveness, level of detail, comprehensibility, consistency, reliability, and overall satisfaction. Additionally, readability scores were determined using the Flesch-Kincaid formulas.

Results: A total of 21 physicians participated (mean age: 38.4 ± 4.9 years; mean professional experience: 6.6 ± 2.7 years). The ChatGPT text contained 973 words with a readability score of 56.3 (10th-12th grade level), while the TNSACI text contained 1,603 words with a readability score of 48.5 (college level). Likert scale evaluations showed no significant difference in accuracy, comprehensiveness, consistency, or reliability between the two texts. However, the ChatGPT text was rated significantly higher for comprehensibility ($p=0.003$) and was considered less overly detailed ($p=0.001$). Regarding overall preference, 57.1% of physicians favored the ChatGPT text, 4.8% preferred the TNSACI text, and 38.1% rated them equally.


Conclusion: Specialist physicians found the ChatGPT-generated asthma information text to be more comprehensible and preferable. These results suggest that AI-based educational content could enhance patient information materials and contribute to more effective patient education.

Keywords: Asthma, patient education, artificial intelligence, large language models, readability

INTRODUCTION

Providing reliable and sufficient sources to patients seeking information about their health conditions plays a crucial role in reducing the burden on healthcare ser-

vices while significantly improving patients' quality of life. The advancement of medical technologies and enhanced accessibility to knowledge empower patients to better understand and manage their health conditions (1).

ORCID  Sevgi Colak / 0000-0001-9042-9668, Pamir Cerci / 0000-0002-0844-6352, Betül Dumanoglu / 0000-0002-4320-2425, Ozge Can Bostan / 0000-0002-4528-5404

In recent years, AI has been increasingly utilized in various domains within the medical field. From early diagnosis of diseases to the treatment plans, as well as patient management and medical research, AI has become a versatile tool in healthcare (2). AI-based advanced language models have shown great potential in patient education. These sophisticated models have the capability to answer patients' questions, provide information about symptoms, and deliver accurate and timely insights on general health topics. By offering reliable information efficiently, they help alleviate the burden on healthcare services while enabling patients to make informed decisions about their health (3).

A critical challenge in healthcare is patient retention of medical information. Research shows that patients typically forget approximately 50% of the information provided during consultations shortly after their visit (4). Written educational materials can effectively address this issue (5). Providing patients with structured written information about their condition—including definitions, causes, symptoms, diagnostic processes, treatment strategies, and emergency recommendations—helps transform abstract medical concepts into tangible and understandable information. These resources bridge the gap between technical medical terminology and patient comprehension, enabling patients to transition from passive recipients to active participants in managing their health conditions in collaboration with healthcare providers (6-9).

This approach is particularly significant in the management of chronic diseases such as asthma, which affects over 300 million people worldwide. Asthma is characterized by heterogeneous and variable respiratory symptoms, and nonadherence to treatment is often associated with misunderstandings or forgetting medical instructions (10). Informed patients who engage in shared decision-making with their physicians demonstrate an improved quality of life (11).

Accordingly, this study aims to evaluate the capacity of AI-based large language models to provide accurate and reliable information to asthma patients and to contribute to the improvement of patient information texts. For this purpose, a patient information text generated by ChatGPT-4o was compared with asthma information texts available on the website of TNSACI. The comparison

was conducted by specialized physicians based on accuracy, comprehensiveness, level of detail, comprehensibility, consistency, reliability, and overall satisfaction. Additionally, the readability of the texts was assessed using the Flesch-Kincaid formulas.

MATERIALS and METHODS

Study Design and Participants

This study involved the evaluation of patient information texts generated by AI-based language models by physicians specializing in Allergy and Clinical Immunology, each with at least five years of professional experience. The AI-generated texts were compared with the existing information materials available on the official website of TNSACI.

Text Generation by ChatGPT-4o

ChatGPT-4o, developed by OpenAI, is a large artificial intelligence language model capable of performing tasks such as text generation, language comprehension, and content creation. In this study, a prompt was developed through an iterative refinement process. Initially, a basic prompt requesting asthma patient information was created, and through multiple testing cycles, it was progressively refined to ensure comprehensive and patient-appropriate content. The final prompt instructed the model to prepare a detailed and understandable text for asthma patients, organized under the following headings: general information about asthma (definition, etiology, brief pathophysiology), symptoms and signs of asthma, methods used in the diagnosis of asthma, and treatment and follow-up of asthma. It emphasized that the informational content, flow, and readability should allow patients to easily comprehend the material, and the depth and quality should be comparable to standard patient education texts. The final prompt was: "Prepare a comprehensive, medically accurate, and easily understandable patient information text about asthma. The text should be structured for patient education and include the following sections: 1) General information about asthma (definition, causes, brief pathophysiology), 2) Symptoms and signs, 3) Diagnosis methods, 4) Treatment and follow-up. Use clear, plain language suitable for adult patients. Ensure logical flow, reliability, and readability comparable to standard patient education materials prepared by professional organizations."

Selection of TNSACI Material

Turkish National Society of Allergy and Clinical Immunology is a professional organization that supports scientific research in the field of allergy and immunology, raises public awareness, and provides guidance to health-care professionals and patients. The asthma information text selected from the TNSACI website was intended to meet the basic educational needs of asthma patients and served as the reference standard for comparison. The TNSACI asthma information text used in this study was obtained from their publicly accessible website in January 2025. However, the document does not provide information about the author(s), date of publication, or referenced sources.

Evaluation Process

The text generated by ChatGPT-4o was reviewed by the study researchers to ensure the absence of ethically questionable statements, biased information, or scientifically unverified content. No issues were identified, and no modifications were required. For the evaluation, both texts were presented anonymously via Google Forms (labeled neutrally as Text A and Text B) to maintain objectivity. Participants were unaware of the origin of each text and were instructed to assess them based on predefined criteria.

The assessment was conducted using a Likert scale, a self-report tool that measures the degree of agreement with specific statements, ranging from 1 (strongly disagree) to 5 (strongly agree) (12). Participants evaluated the texts on accuracy, comprehensiveness, level of detail, comprehensibility, consistency, reliability, and overall satisfaction. Additionally, three direct comparison questions were included regarding comprehensibility, reliability, and overall preference between the two texts.

Readability Assessment

Readability of the texts was assessed using the Flesch-Kincaid formula(13), which estimates the education level required for a reader to understand a text. The formula is expressed as:

$$\text{Flesch-Kincaid: } 0.39 \times (W/S) + 11.8 \times (B/W) - 15.59,$$

where W represents the number of words, S the number of sentences, and B the number of syllables. Higher scores

indicate easier readability. Both the TNSACI and the AI-generated texts were evaluated in their original Turkish language. The Flesch-Kincaid formula, while originally designed for English, was used as an approximate index to allow comparative readability assessment.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 25.0 (IBM Corp., Armonk, NY, USA). Continuous variables are presented as median values with minimum and maximum ranges, while categorical variables are expressed as numbers (n) and percentages (%). Categorical data from Likert scale responses were analyzed using the Wilcoxon signed-rank test. For direct comparison questions, differences in response distributions were initially evaluated using the chi-square test. Where applicable, post-hoc pairwise comparisons were conducted using Fisher's exact test. A p-value < 0.05 was considered statistically significant.

Ethical Considerations

The study was approved by the Ethics Committee of Ağrı İbrahim Çeçen University Hospital (Approval No: 438/28.11.2024). Written informed consent was obtained online from all participants before the initiation of study procedures. The study was conducted in accordance with the principles outlined in the World Medical Association Declaration of Helsinki.

RESULTS

Twenty-one specialists participated in the study. The mean age was 38.4 ± 4.9 years and mean experience 6.6 ± 2.7 years.

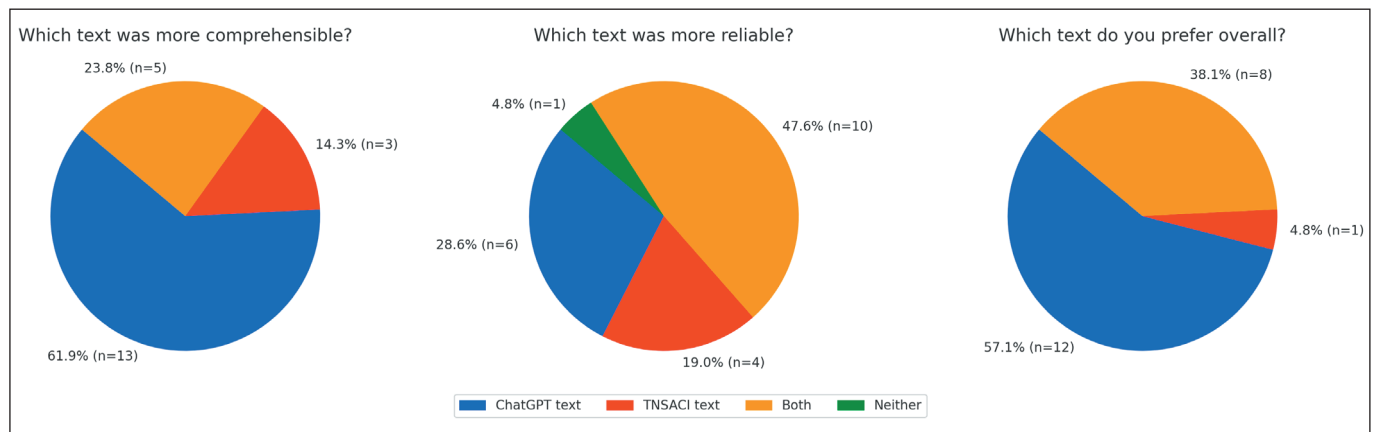
Results for the Questions with a Likert Scale

Seven questions were scored using a Likert scale ranging from 1 to 5, and the questions along with the median scores of the responses are presented in Table I. Physicians rated both texts as accurate, comprehensive, reliable, and consistent, with no statistically significant differences between the two texts in these aspects. However, the Likert scores for being overly detailed and confusing were significantly higher for the TNSACI text ($p=0.001$, median: 2 for ChatGPT versus 4 for TNSACI). Additionally, the ChatGPT text was rated significantly higher for comprehensibility ($p=0.03$, median: 4 for ChatGPT versus 3 for TNSACI). Overall satisfaction was similar for both texts.

Table I: Comparison of responses to Likert scale questions regarding the Chatgpt-4.0 and TNSACI asthma patient information texts.

Evaluation Statement	ChatGPT-4o Text*	TNSACI Text*	P
I believe the content of the text is accurate and error-free.	4 (3-5)	4 (1-5)	NS
The text addresses the topic in a sufficiently detailed and comprehensive manner.	4 (3-5)	4 (1-5)	NS
I think the text is overly detailed and confusing.	2 (1-4)	4 (2-5)	0.001
The text is comprehensible.	4 (3-5)	3 (1-5)	0.003
The information presented in the text is reliable.	4 (3-5)	4 (2-5)	NS
The information presented in the text is consistent.	4 (3-5)	4 (2-5)	NS
Overall, I am satisfied with the text for asthma patient education.	4 (3-5)	4 (2-5)	NS

*Median (min-max)

ChatGPT: Chat Generative Pre-Trained Transformer, **TNSACI:** Turkish National Society of Allergy and Clinical Immunology**Figure 1:** Results for the direct comparison questions between the TNSACI and ChatGPT text.**ChatGPT:** Chat Generative Pre-trained Transformer, **TNSACI:** Turkish National Society of Allergy and Clinical Immunology

Results for the Direct Comparison Questions Between the Two Texts

In response to the question “Which text was more comprehensible?”, the majority of participants (61.9%) indicated that the ChatGPT text was more comprehensible (Figure 1). Regarding reliability, most participants (47.6%) found both texts equally reliable, while 28.6% favored the ChatGPT text. For overall preference, more than half of the participants (57.1%) preferred the ChatGPT text over the TNSACI text. Although the overall chi-square test indicated a statistically significant difference in response distributions, post-hoc pairwise comparisons using Fisher’s exact test did not reveal significant differences between individual groups.

Results for Readability Index Analysis

The ChatGPT text included 973 words, with a Flesch-Kincaid Grade Level of 9.1 and a Flesch Reading Ease Score of 56.3, indicating a readability level corresponding to 10th-12th grade (fairly difficult to read). The TNSACI text contained 1,603 words, with a Flesch-Kincaid Grade Level of 10.0 and a Flesch Reading Ease Score of 48.5, corresponding to a college-level reading difficulty (difficult to read).

DISCUSSION

In this study, patient information text for asthma from the TNSACI website and AI-generated text by ChatGPT were compared by experienced specialist physicians. Both

texts were found to be similar in terms of accuracy, consistency, and reliability; however, the ChatGPT text was considered more comprehensible. While both texts were deemed sufficiently detailed and comprehensive, the TNSACI text was found to be overly detailed. The ChatGPT text was preferred by 57% of the participants. According to Flesch-Kincaid assessments, the ChatGPT text was easier to read. This research is a valuable addition to the literature, as it compared pre-existing patient information text and AI-generated text, evaluated and scored by specialized physicians in a semi-blinded manner. The study highlights the potential of LLMs to assist medical professionals in generating patient-friendly informational materials.

In recent years, there has been a growing number of publications in the medical literature on the use of AI in patient education. Although studies in this field are methodologically quite heterogeneous, they predominantly focus on evaluating AI-generated responses to frequently asked questions about diseases (14-16). Additionally, there are publications comparing responses from different AI chatbots and assessing the readability of pre-existing informational texts that have been restructured by AI (17-19). However, publications directly comparing pre-existing informational texts with AI-generated texts remain limited.

The Likert scale is widely used to evaluate the quality of information provided by AI (20). Another tool, "Ensuring Quality Information for Patients" (EQIP), is also valuable for structured evaluation of a text's quality, clarity, readability, and usability (21,22). DISCERN was designed to assess the quality of written informational texts presented to patients about treatment options and has been used in evaluating AI-generated texts (23,24). Among these tools, the Likert scale allows for quantitative measurement of participants' subjective opinions regarding predetermined criteria for assessing text quality (12). In a recent study designed with a methodology similar to ours, a panel of 10 experts evaluated the responses of AI chatbots to frequently asked questions about adolescent idiopathic scoliosis using a Likert scale. The findings indicated that ChatGPT-4.0 was rated as more satisfactory compared to ChatGPT-3.5 and Google Bard (25). Due to its ease of application and flexibility, which allows researchers to select specific parameters for evaluation, we chose the Likert scale for our study.

In allergy and immunology practice, AI chatbots can significantly contribute by improving patient education, streamlining diagnostic support, and providing personalized treatment recommendations (26). Considering that asthma patient education directly impacts treatment success, the potential role of AI in this field is not surprising; however, the literature on this topic remains quite limited. In a study by Ghazali, questions from the "Asthma General Knowledge Questionnaire for Adults" were posed to ChatGPT, and the responses were evaluated by three general practitioners using a four-level quantitative method. ChatGPT's answers were found to be over 90% accurate in most categories, while the accuracy for medication-related questions was 70% (27). In another study, ChatGPT's responses to frequently asked questions about asthma were evaluated by five internal medicine specialists using a similar four-level evaluation method. The information was found to be over 80% reliable and over 70% acceptable (28). In a study by Høj et al., 26 asthma-related questions were asked to ChatGPT, and five healthcare professionals scored the answers on a scale from 1 to 5, representing different levels of accuracy: 1 for very poor with unacceptable inaccuracies, 2 for poor with minor potentially harmful inaccuracies, 3 for moderate with potentially misinterpretable inaccuracies, 4 for good with only minor non-harmful inaccuracies, and 5 for very good with no inaccuracies. The majority of the responses (81%) received a score of 4 or higher, while five responses scored 3 or below, indicating minor but potentially harmful inaccuracies (29). The researchers concluded that while ChatGPT shows promise as a helpful assistant, it cannot replace healthcare professionals.

It is generally recommended that informative medical texts be written at a sixth to seventh-grade reading level, and AI can assist in enhancing the readability and understandability of such materials (19). With the canvas feature of ChatGPT, users can now more easily adjust the readability level of any text. In our study, both the TNSACI text and the AI-generated text had readability levels higher than the recommended threshold. The higher readability of the ChatGPT-generated text may be attributed to the comprehensive prompt designed to match the depth and structure of the TNSACI material.

This study has several limitations. Firstly although we provided the AI model with a detailed prompt that aimed to replicate the depth and comprehensiveness of the TNSACI text, the resulting content was significantly

shorter. This discrepancy may have influenced participants' perceptions regarding readability and overall preference in favor of the ChatGPT text, and thus should be considered a limitation. The difference in word count is likely due to the fact that the TNSACI text itself was not provided to the AI model. Instead, the model was asked to generate a completely original patient information text from scratch, based on specific educational headings. Our intention was not to optimize or shorten an existing text, but to evaluate an independently created AI-generated text. The fact that the AI-generated text was shorter may also indicate that the essential information can be delivered in a more concise and focused manner, without unnecessary elaboration. Second, since the TNSACI text was publicly available, complete blinding of the evaluators may not have been fully achieved, meaning that our study should be considered semi-blinded. Additionally, although the sample size of 21 physicians is consistent with previous studies involving expert evaluations, a formal sample size calculation was not conducted. This limits the statistical power and should be taken into account when interpreting the results.

To the best of our knowledge, our study is among the first to compare AI-generated asthma patient education texts with pre-existing materials, offering a novel contribution to the literature. Future studies should aim to refine and standardize methodologies. In light of these findings, updating existing educational materials holds promise for improving health literacy and providing better information to asthma patients.

Conflict of Interest

The authors declare that they have no conflicts of interest relevant to this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Concept: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Design: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Data collection or processing: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Analysis or Interpretation: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Literature search: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Writing: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**, Approval: **Sevgi Colak, Pamir Cerci, Betul Dumanoglu, Ozge Can Bostan**.

REFERENCES

1. Bhattad PB, Pacifico L. Empowering Patients: Promoting Patient Education and Health Literacy. *Cureus* 2022;14(7):e27336.
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022;28(1):31-8.
3. Mucci A, Green W, Hill LH. Incorporation of artificial intelligence in healthcare professions and patient education for fostering effective patient care. *New Directions for Adult and Continuing Education* 2024;2024(181):51-62.
4. Kessels RP. Patients' memory for medical information. *J R Soc Med* 2003;96(5):219-22.
5. Nicholson Thomas E, Edwards L, McArdle P. Knowledge is Power. A quality improvement project to increase patient understanding of their hospital stay. *BMJ Qual Improv Rep* 2017;6(1):u207103.w3042.
6. Oermann MH. Effects of educational intervention in waiting room on patient satisfaction. *J Ambul Care Manage* 2003;26(2):150-8.
7. Friedman AJ, Cosby R, Boyko S, Hatton-Bauer J, Turnbull G. Effective teaching strategies and methods of delivery for patient education: a systematic review and practice guideline recommendations. *J Cancer Educ* 2011;26(1):12-21.
8. Low M, Burgess LC, Wainwright TW. Patient Information Leaflets for Lumbar Spine Surgery: A Missed Opportunity. *J Patient Exp* 2020;7(6):1403-9.
9. Ginat DT, Christoforidis G. A printed information leaflet about MRI and radiologists improves neuroradiology patient health literacy. *Neuroradiol J* 2018;31(6):609-13.
10. 2024 GINA Main Report - Global Initiative for Asthma - GINA. Available from: <https://ginasthma.org/2024-report/>
11. Wilson SR, Strub P, Buist AS, Knowles SB, Lavori PW, Lapidus J, et al. Shared treatment decision making improves adherence and outcomes in poorly controlled asthma. *Am J Respir Crit Care Med* 2010;181(6):566-77.
12. Jebb AT, Ng V, Tay L. A Review of Key Likert Scale Development Advances: 1995-2019. *Front Psychol* 2021;12:637547.
13. Flesch R. A new readability yardstick. *J Appl Psychol* 1948;32(3):221-33.
14. Şenoymak İ, Erbatur NH, Şenoymak MC, Egici MT. Evaluating the accuracy and adequacy of ChatGPT in responding to queries of diabetes patients in primary healthcare. *Int J Diabetes Dev Ctries* 2025;45: 619-26.
15. Ye Z, Zhang B, Zhang K, Méndez MJG, Yan H, Wu T, et al. An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer. *BMC Womens Health* 2024;24(1):1-10.
16. Kuşçu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol* 2023;13:1256459.

17. Gondode P, Duggal S, Garg N, Lohakare P, Jakhar J, Bharti S, et al. Comparative Analysis of Accuracy, Readability, Sentiment, and Actionability: Artificial Intelligence Chatbots (ChatGPT and Google Gemini) versus Traditional Patient Information Leaflets for Local Anesthesia in Eye Surgery. *Br Ir Orthopt J*. 2024;20(1):183-92.
18. Yau JYS, Saadat S, Hsu E, Murphy LSL, Roh JS, Suchard J, et al. Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study. *J Med Internet Res*. 2024;26:e60291.
19. Nasra M, Jaffri R, Pavlin-Premrl D, Kok HK, Khabaza A, Barras C, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. *Intern Med J* 2025;55(1):20-34.
20. Høj S, Thomsen SF, Meteran H, Sigsgaard T, Meteran H. Artificial intelligence and allergic rhinitis: does ChatGPT increase or impair the knowledge? *J Public Health (Oxf)* 2024;46(1):123-6.
21. Moulton B, Franck LS, Brady H. Ensuring quality information for patients: Development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect* 2004;7(2):165-75.
22. 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* 2023;25:e47479.
23. Warn M, Meller LLT, Chan D, Torabi SJ, Bitner BF, Tajudeen BA, et al. Assessing the Readability, Reliability, and Quality of AI-Modified and Generated Patient Education Materials for Endoscopic Skull Base Surgery. *Am J Rhinol Allergy* 2024;38(6):396-402.
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53(2):105-11.
25. Lang S, Vitale J, Galbusera F, Fekete T, Boissiere L, Charles YP, et al. Is the information provided by large language models valid in educating patients about adolescent idiopathic scoliosis? An evaluation of content, clarity, and empathy : The perspective of the European Spine Study Group. *Spine Deform* 2025;13(2):361-72.
26. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial Intelligence Chatbots in Allergy and Immunology Practice: Where Have We Been and Where Are We Going? *J Allergy Clin Immunol Pract* 2023;11(9):2697-700.
27. Ghazali MT. Assessing ChatGPT's accuracy and reliability in asthma general knowledge: implications for artificial intelligence use in public health education. *J Asthma* 2025;62(6):975-83.
28. Alabdulmohsen DM, Almahmudi MA, Alhashim JN, Almahdi MH, Alkishy EF, Almossabeh MJ, et al. Is ChatGPT a Reliable Source of Patient Information on Asthma? *Cureus* 2024;16(7):e64114.
29. Høj S, Thomsen SF, Ulrik CS, Meteran H, Sigsgaard T, Meteran H. Evaluating the Scientific Reliability of ChatGPT as a Source of Information on Asthma. *J Allergy Clin Immunol Glob* 2024;3(4):100330.