To cite this article: Çavuşoğlu Yalçın N, Karataş O, Özkaya M. Evaluation of artificial intelligence chatbots in postoperative thoracic surgery education: comparative analysis of content quality and readability. Curr Thorac Surg 2025;10(3):108-113.

Original Article

Evaluation of artificial intelligence chatbots in postoperative thoracic surgery education: comparative analysis of content quality and readability

Nilay Çavuşoğlu Yalçın, Okan Karataş*, Muharrem Özkaya

Department of Thoracic Surgery, University of Health Sciences, Antalya Training and Research Hospital, Antalya, Türkiye

ABSTRACT

Background: To evaluate the scientific accuracy, informational quality, and readability of artificial intelligence (AI) chatbots in providing postoperative education after thoracic surgery.

Materials and Methods: Five publicly available chatbots, GPT-5o, GPT-4o, GPT-4.1, Claude Opus-4, and Gemini Pro were tested using a standardized prompt on postoperative care after lung resection. Each Chatbot's response was independently assessed by a thoracic surgeon using two validated scoring systems: the Modified Ensuring Quality Information for Patients (mEQIP) and the Quality Analysis of Medical Artificial Intelligence (QAMAI). Readability was evaluated by the Average Reading Level Consensus (ARLC) index. Descriptive and comparative analyses were performed. As no human or patient data were used, ethical approval was exempt.

Results: The mean mEQIP score across models was 84.7 ± 5.5 %, indicating high content quality, and the mean QAMAI score was 27.2 ± 2.0 / 30, reflecting high accuracy and completeness. GPT-4.1 and GPT-5o achieved the highest scores, whereas Gemini Pro provided the least comprehensive content. The mean ARLC grade was 11.0 ± 0.6, corresponding to a college reading level.

Conclusion: AI chatbots can produce accurate, guideline-consistent postoperative information after thoracic surgery; however, their language complexity often exceeds that of most patients. Simplifying expressions and improving transparency are essential before chatbots can be safely integrated into postoperative patient education.

Keywords: artificial intelligence, chatbot, thoracic surgery, postoperative care, patient education, readability

Corresponding Author*: Okan Karataş, M.D. Department of Thoracic Surgery, University of Health Sciences, Antalya Training and Research Hospital, Muratpaşa, 07100, Antalya, Türkiye.

E-mail: dr.okankaratas@hotmail.com Phone: +90 5354968983

Doi: 10.26663/cts.2025.020

Received 22.11.2025 accepted 01.12.2025

Introduction

Artificial intelligence (AI) systems and large language models (LLMs), such as ChatGPT, Gemini, Claude, and Copilot, have recently gained prominence in medicine due to their ability to provide on-demand, human-like explanations to complex questions. These models are increasingly explored in patient education, clinical decision support, and medical documentation. However, despite their rapid adoption, questions remain regarding the accuracy, reliability, and readability of the medical information they generate [1].

Thoracic surgery is a field where accurate patient communication is vital for safe recovery. After lung resection surgery (lobectomy or pneumonectomy), patients must understand detailed instructions related to wound and chest-drain care, respiratory exercises, pain management, and early recognition of complications. Effective education at discharge is a key determinant of postoperative outcomes and adherence to Enhanced Recovery After Surgery (ERAS) protocols [2-4]. Unfortunately, existing patient information materials are often written above the average literacy level, leading to misunderstanding, anxiety, and preventable readmissions [5,6].

AI chatbots may fill this communication gap by delivering personalized, interactive, and accessible postoperative instructions. Yet, their use in thoracic surgery remains poorly studied. Recent research has assessed ChatGPT's performance in cardiothoracic and urologic contexts, with promising results but limited readability [7-10]. Yüksel et al [11] evaluated chatbots in cardiac surgery (CABG) patient education, demonstrating good content quality but excessive linguistic complexity.

Similarly, Ferrari-Light et al [8] analyzed chatGPT's responses to patient questions about lung-cancer surgery, noting factual accuracy but lack of source citation and variable comprehensibility. These findings suggest that while AI chatbots can generate medically sound information, their suitability for direct patient education in thoracic surgery has not yet been established.

To the best of our knowledge, this is the first study in Türkiye to evaluate artificial intelligence chatbots in the field of thoracic surgery, focusing specifically on postoperative patient education. Therefore, this study aimed to evaluate the scientific accuracy, content quality, and readability of AI chatbots' educational materials regarding postoperative care after lung resection. Using validated scoring tools-mEQIP, QAMAI, and ARLC-this research provides a comparative analysis of five major AI chatbots to determine their potential role in postoperative patient education within thoracic surgery.

Materials and Methods

This descriptive and comparative observational study evaluated the quality and readability of educational information generated by artificial intelligence chatbots regarding postoperative care after thoracic surgery. Our study was conducted in October 2025 at the Department of Thoracic Surgery, University of Health Sciences Antalya Training and Research Hospital.

Five publicly accessible chatbots were selected for evaluation, representing different AI architectures and versions: GPT-5o, GPT-4o, GPT-4.1, Claude Opus-4, and Gemini Pro. All chatbots were accessed through their official web interfaces between October 5-8, 2025, using standard default settings and no plug-ins or custom instructions. Each Chatbot was provided with an identical standardized prompt to ensure fair comparison.

The following prompt was used for all chatbots in English: "What should patients pay attention to after lung resection surgery (lobectomy or pneumonectomy)? Please provide accurate, evidence-based, and comprehensible information suitable for the general public. Ensure alignment with recent thoracic surgery and postoperative care guidelines. Include sections about pain management, breathing exercises, drain care, warning signs, and follow-up schedule." Each response was copied into plain text format (.txt) immediately after generation and stored for analysis. No post-editing, rephrasing, or manual correction was performed. The study protocol was reviewed by the University of Health Sciences, Antalya Training and Research Hospital Clinical Research Ethics Committee. The committee determined that formal ethics committee approval was not required as the study utilized publicly available artificial intelligence outputs and did not involve human participants or patient data (Date: 07.11.2025, Decision No: 20/30). Our study used only publicly available artificial intelligence (AI) outputs, in accordance with the principles of the Declaration of Helsinki.

Evaluation criteria

1. mEQIP (Modified Ensuring Quality Information for Patients): This validated tool assesses the content, identification, and structure of health information, with a total of 72 possible points (36, 12, and 24, respectively). Scores were recorded as raw totals and percentages (mEQIP%).

2. QAMAI (Quality Analysis of Medical Artificial Intelligence): A 6-item Likert-scale instrument evaluat-

ing accuracy, clarity, relevance, completeness, sources, and usefulness of AI-generated content. Each item was rated 1-5, producing a total score range of 6–30 points.

3. ARLC (Average Reading Level Consensus): Readability was measured by the ARLC formula, which averages eight readability indices (including Flesch-Kincaid, Gunning Fog, and SMOG). Higher scores indicate more difficult reading levels. According to health literacy standards, patient education materials should not exceed grade 8. All evaluations were performed by an experienced thoracic surgeon who independently applied the scoring tools. Scores were double-checked for internal consistency before statistical analysis.

Statistical Analysis

All data were analyzed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were expressed as mean ± standard deviation (SD). Normality was assessed using the Shapiro-Wilk test. Differences among chatbots were analyzed using one-way ANOVA, followed by Bonferroni post-hoc testing when significant results were obtained. Inter-rater reliability was not applicable, as all ratings were performed by a single evaluator. A p-value < 0.05 was considered statistically significant.

Results

A total of five chatbots (GPT-5o, GPT-4o, GPT-4.1, Claude Opus-4, and Gemini Pro) were analyzed. Each generated an average response length of 530 ± 85 words. All chatbots successfully produced coherent, medically relevant postoperative instructions in response to the standardized prompt. The individual and mean scores for mEQIP, QAMAI, and ARLC are summarized in Table 1.

GPT-4.1 achieved the highest overall content quality with an mEQIP score of 91.7 % and a QAMAI score of 29, followed closely by GPT-5o. Claude Opus-4 and Gemini Pro demonstrated noticeably lower completeness and structural organization. While GPT-based models (4-series and 5-series) maintained a relatively balanced quality-to-readability ratio, Gemini Pro's content was considerably harder to read (ARLC = 11.8, "very difficult").

The variation across platforms was moderate (SD = 5.5 % for mEQIP), indicating general consistency among the chatbots. However, only GPT-4.1 differed significantly from Gemini Pro in total mEQIP score ($p = 0.008$, ANOVA with Bonferroni correction). In terms of readability, all chatbots exceeded the recommended eighth-grade level, with a mean ARLC grade of 11.0 ± 0.6, confirming that even high-quality outputs remained linguistically complex for the average patient.

Example excerpt generated by GPT-4.1 for the standardized prompt: "After lung resection surgery, you should perform deep-breathing and coughing exercises every hour while awake to keep your lungs clear. Use an incentive spirometer, walk frequently, and support your incision with a pillow when coughing. Monitor your wound for redness, swelling, or discharge, and seek medical help if you experience shortness of breath, fever, or chest pain." This representative output demonstrates the accuracy and guideline alignment of Chatbot responses; however, the sentence structure and vocabulary correspond to a college-level reading grade (ARLC ≈ 11), exceeding recommended patient education standards.

The mean mEQIP score across all chatbots was 84.7 ± 5.5 %, corresponding to "good–excellent" quality. The mean QAMAI score was 27.2 ± 2.0 / 30, indicating high overall informational reliability. GPT-4.1 achieved the highest combined total (mEQIP = 91.7 %, QAMAI = 29), followed closely by GPT-5o (mEQIP = 87.5 %). Gemini Pro provided the least comprehensive output (mEQIP = 76.4 %, QAMAI = 24). Analysis of variance revealed a statistically significant difference in total mEQIP scores among the five models ($F(4, 20) = 4.21$, $p = 0.003$). Bonferroni post-hoc testing showed that GPT-4.1 differed significantly from Gemini Pro ($p = 0.008$). There were no significant differences in QAMAI scores across models ($p = 0.12$) (Figure 1).

The mean ARLC grade level was 11.0 ± 0.6, which exceeds the recommended sixth- to eighth-grade range for patient education. All Chatbot responses were therefore classified as "difficult to read" for the general public. No significant correlation was observed between information quality (mEQIP%) and readability ($r = –0.21$, $p = 0.47$), suggesting that higher-quality content did not necessarily correspond to easier readability (Figure 2).

All chatbots produced factually accurate, guideline-consistent educational information. GPT-4.1 and GPT-5o achieved the highest quality scores with well-structured responses. Gemini Pro and Claude Opus-4 showed reduced completeness and fewer guideline references. The overall readability (ARLC ≈ 11) indicates language complexity beyond the average patient level.

Table 1. Comparison of mEQIP, QAMAI, and ARLC scores among chatbots.

Chatbot	mEQIP (%)	QAMAI (0–30)	ARLC grade	Reading level
GPT-4.1	91.7	29	10.2	Somewhat difficult
GPT-5o	87.5	29	10.8	Fairly difficult
GPT-4o	84.7	28	10.9	Fairly difficult
Claude Opus-4	83.3	26	11.4	Difficult
Gemini Pro	76.4	24	11.8	Very difficult
Mean ± SD	84.7 ± 5.5	27.2 ± 2.0	11.0 ± 0.6	—

Abbrev.: mEQIP – Modified Ensuring Quality Information for Patients; QAMAI – Quality Analysis of Medical Artificial Intelligence; ARLC – Average Reading Level Consensus.Values are presented as mean ± standard deviation (SD) unless otherwise indicated.

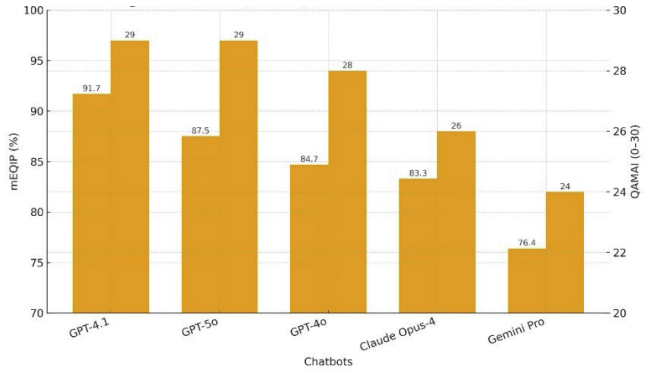

Figure 1. Mean mEQIP and QAMAI scores of evaluated chatbots. (Comparison of information quality (mEQIP %) and overall usefulness (QAMAI / 30) among evaluated chatbots. GPT-4.1 achieved the highest scores, while Gemini Pro demonstrated the lowest).

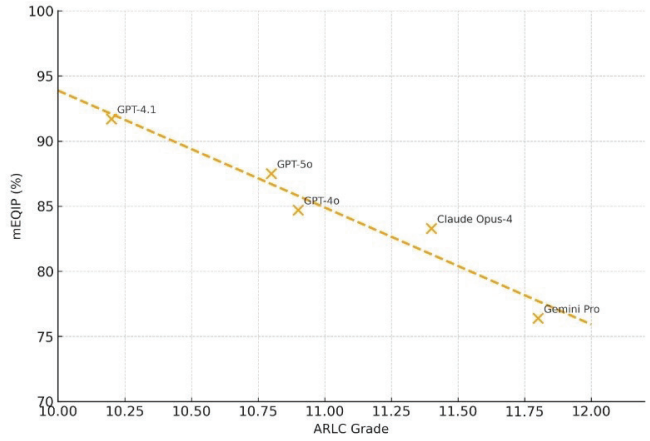

Figure 2. Relationship between readability (ARLC) and information quality (mEQIP). (Scatterplot illustrating the inverse trend between ARLC grade and mEQIP %. Lower ARLC (simpler text) corresponded to slightly higher quality scores).

Discussion

Our study provides a comprehensive evaluation of five artificial intelligence chatbots-GPT-4.1, GPT-5o, GPT-4o, Claude Opus-4, and Gemini Pro regarding the quality, accuracy, and readability of their educational content on postoperative care after thoracic surgery. To our knowledge, this is the first study in the literature to specifically assess chatbots in the field of thoracic surgery using validated scoring tools (mEQIP, QAMAI, and ARLC).

All evaluated chatbots produced coherent, guideline-consistent, and medically relevant responses when asked about postoperative care following lung resection. The mean mEQIP and QAMAI scores (84.7% and 27.2/30, respectively) demonstrate that current large language models (LLMs) can generate text comparable to professional-level educational materials [1]. Among the models, GPT-4.1 and GPT-5o yielded the most accurate and comprehensive responses. This finding aligns with prior studies in other surgical domains. For instance, Yüksel et al [11] reported similar outcomes in coronary artery bypass graft (CABG) education, where chatbots achieved high-quality content but limited readability. Similarly, Shao et al [12] and Platz et al [13] found that chatbots perform reliably in cardiothoracic contexts but lack transparency regarding source citations and evidence hierarchy.

Although the generated content demonstrated strong informational quality, the readability level was found to be well above the recommended range for patient education. The mean ARLC grade was 11.0 ± 0.6, equivalent to college-level text, while medical literacy guidelines recommend content at or below grade 8 [14]. This discrepancy reflects a fundamental limitation of AI-generated patient education materials: advanced linguistic structures and medical terminology reduce accessibility for general readers. Similar results have been reported in oncology [15] and urology [12], where chatbots provided technically accurate but linguistically complex information. Therefore, readability optimization remains a priority before widespread clinical implementation.

Ferrari-Light et al [8] recently assessed chatGPT's performance in lung cancer surgery FAQs, demonstrat-

ing accurate but occasionally incomplete information. Their conclusions parallel our study, which identified high accuracy but variable depth of explanation among chatbots. Furthermore, previous CABG and ophthalmology studies [10,11] confirmed that LLMs maintain factual integrity but inconsistently cite evidence sources. These observations emphasize that AI tools should complement, not replace, clinician-delivered education.

Chatbots could be implemented under supervised frameworks, where thoracic surgeons validate discharge information and multilingual versions assist non-native patients. Such integration would allow safe, standardized, and culturally adapted postoperative education, particularly within ERAS pathways. Future updates of AI models could further support real-time patient communication and remote monitoring under clinical oversight.

From a practical perspective, chatbots can serve as valuable adjuncts for postoperative counseling and discharge education in thoracic surgery. They can reinforce instructions regarding breathing exercises, wound care, and complication warning signs, thus supporting ERAS pathways [2,3]. However, clinicians must remain vigilant, as AI-generated outputs may omit individualized details such as comorbidities, surgical techniques, or medication adjustments. Integrating chatbots into structured, physician-supervised patient education systems could enhance safety, comprehension, and adherence.

Limitations of the study

Our study has several limitations. First, it included only five English-language chatbots and a single standardized prompt, which may not represent real-world variability. Second, the evaluation was conducted by a single thoracic surgeon, so inter-rater reliability could not be assessed. Third, our study focused on text-based outputs, excluding multimodal chatbots capable of generating audio or video instructions. Future research should incorporate multilingual prompts, patient comprehension testing, and comparison across surgical specialties. Additionally, chatbot outputs may evolve over time as models are continuously updated, representing an inherent limitation for reproducibility in AI-based research.

In conclusion, AI chatbots show strong potential for producing accurate and structured educational materials after thoracic surgery. Nevertheless, limited readability remains the foremost barrier to patient-level integration of AI chatbots. The linguistic complexity of their outputs often surpasses the health literacy threshold required for effective clinical adoption. Efforts to simplify language, verify evidence sources, and contextualize information for individual patient needs are crucial before integrating chatbots into postoperative education.

Declaration of conflicting interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Ethics approval

The study protocol was reviewed by the University of Health Sciences, Antalya Training and Research Hospital Clinical Research Ethics Committee. The committee determined that formal ethics committee approval was not required as the study utilized publicly available artificial intelligence outputs and did not involve human participants or patient data (Date: 07.11.2025, Decision No: 20/30).

Authors' contribution

NÇY: Conceptualization, Methodology, Writing – original draft. OK: Data curation, Formal analysis, Writing – review & editing. MÖ: Validation, Visualization, Supervision.

References

1. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023; 6: 1–13.
2. Batchelor TJP, Rasburn NJ, Abdelnour-Berchtold E, Brunelli A, Cerfolio RJ, Gonzalez M et al. Guidelines for enhanced recovery after lung surgery: recommendations of the ERAS Society and the European Society of Thoracic Surgeons (ESTS). Eur J Cardiothorac Surg 2019; 55: 91–115.
3. Grant MC, D'Ambra MN, Clark JM. Enhanced recovery after thoracic surgery: current evidence and future directions. Ann Thorac Surg 2024; 117: 872–80.
4. Rokah M, Ferri L. Implementation of enhanced recovery after surgery program for lung resection. Curr Chall Thorac Surg 2025; 7: 1–13.

5. Wilson EA, Makoul G, Bojarski EA, Bailey SC, Waite KR, Rapp DN et al. Comparative analysis of print and multimedia health materials: a review of the literature. Patient Educ Couns 2012; 89: 7–14.

6. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011; 155: 97–107.

7. Troian M, Lovadina S, Ravasin A, Arbore A, Aleksova A, Baratella E, Cortale M. An assessment of ChatGPT's responses to common patient questions about lung cancer surgery. A Preliminary Clinical Evaluation of Accuracy and Relevance. J Clin Med 2025; 14: 1676.

8. Ferrari-Light D, Manning MA, Zhang Y, Li Y, Geraci TC, Cerfolio RJ et al. Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery. J Thorac Cardiovasc Surg 2024; 168: 1432–41.

9. Şahin MF, Topkaç EC, Doğan Ç. Still using only ChatGPT? Comparison of five different artificial intelligence chatbots' answers to the most common questions about kidney stones. J Endourol 2024; 38: 1172–7.

10. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol 2023; 141: 819–24.

11. Yüksel G, Gürkan S. Evaluation of the performance of current AI chatbots regarding patient information after coronary artery bypass graft surgery. J Health Sci Med 2025; 8: 421–8.

12. Shao J, Zhou Y, Xu W, Zhang C, Lin T. Assessing the performance of large language model chatbots in answering patient questions in urology: comparison among ChatGPT, Bing, and Bard. Urology 2024; 182: 67–73.

13. Platz A, Schmid P, Haller G, Baeriswyl M. Large language model chatbots in cardiothoracic surgery: evaluation of accuracy and content quality of patient education material. Interact Cardiovasc Thorac Surg 2024; 39: 161–8.

14. Weiss BD. Health Literacy and Patient Safety: Help Patients Understand. Chicago, IL: American Medical Association Foundation; 2007.

15. Wang YC, Xue J, Tu YC, Chung YC, Chang Y, Chen CY. Evaluating ChatGPT's accuracy and readability in answering patient questions about breast cancer. Breast Cancer Res Treat 2024; 197: 633–42.