Original Article

HEALTH SCIENCES
MEDICINE

DOI: 10.32322/jhsm.1660603
J Health Sci Med. 2025;8(3):524-528

Evaluation of artificial intelligence in thoracic surgery internship education: accuracy and usability of AI-generated exam questions

İsmail Dal
Department of Thoracic Surgery, Faculty of Medicine, Kastamonu University, Kastamonu, Turkiye

Cite this article as: Dal İ. Evaluation of artificial intelligence in thoracic surgery internship education: accuracy and usability of AI-generated exam questions. *J Health Sci Med.* 2025;8(3):524-528.

Received: 18.03.2025 ◆ **Accepted**: 27.05.2025 ◆ **Published**: 30.05.2025

ABSTRACT

Aims: This study aims to evaluate the usefulness and reliability of artificial intelligence (AI) applications in thoracic surgery internship education and exam preparation.

Methods: Claude Sonnet 3.7 AI was provided with core topics covered in the 5th-year thoracic surgery internship and was instructed to generate a 20-question multiple-choice exam, including an answer key. Four thoracic surgery specialists assessed the AI-generated questions using the Delphi panel method, classifying them as correct, minor error, or major error. Major errors included the absence of the correct answer among choices, incorrect AI-marked answers, or contradictions with established medical knowledge. A second exam was manually created by a thoracic surgery specialist and evaluated using the same methodology. Seven volunteer 5th-year medical students completed both exams, and the correlation between their scores was statistically analyzed.

Results: Among AI-generated questions, 8 (40%) contained major errors, while 1 (5%) had a minor error. The expert-generated exam had a perfect accuracy rate, whereas the AI-generated exam had significantly lower accuracy (p=0.001). Median scores were 75 (67-100) for the AI exam and 85 (70-95) for the expert exam. No significant correlation was found between students' scores (r=0.042, p=0.929).

Conclusion: AI-generated questions had a high error rate (40% major, 5% minor), making them unreliable for unsupervised use in medical education. While AI may provide partial benefits under expert supervision, it currently lacks the accuracy required for independent implementation in thoracic surgery education.

Keywords: Artificial ıntelligence, thoracic surgery education, multiple choice tests, delphi technique

INTRODUCTION

Artificial intelligence (AI) applications have rapidly evolved, demonstrating significant potential in various domains of medical education, including clinical decision support, diagnostic accuracy improvement, and personalized learning experiences. Recent studies suggest that AI-generated content can be useful in medical training by automating question generation, simulating clinical cases, and enhancing student engagement.[1,2] However, concerns regarding the accuracy, reliability, and ethical implications of AI-generated educational materials remain.[3,4]

Thoracic surgery is a highly specialized field that requires a deep understanding of complex surgical procedures, anatomical structures, and perioperative management. The effectiveness of AI in generating thoracic surgery-related multiple-choice questions (MCQs) for medical students has not been extensively studied. Prior research has demonstrated that AI-generated MCQs can sometimes contain factual inaccuracies or misleading information, necessitating thorough expert review before implementation. While artificial intelligence holds significant potential in the future of medical education, the persistent reliance on traditional teaching methods presents challenges for integrating such innovative tools. Moreover, although the content of this study is not strictly limited to thoracic surgery knowledge, it was conducted within the context of thoracic surgery education and assessed by experts in the field. Therefore, the study aims to contribute not only to surgical education but also to the broader conversation on the role of AI in developing assessment tools for specialized medical domains.

The Delphi method has been widely used to assess the validity of educational content by utilizing expert consensus.[5] This approach ensures that medical assessments maintain high accuracy and educational value. Given the increasing reliance on AI in medical education, it is essential to evaluate its role in thoracic surgery training, particularly regarding its ability to generate reliable and high-quality exam questions.

Corresponding Author: İsmail Dal, idal@kastamonu.edu.tr

In this study, we assess the quality of AI-generated MCQs for thoracic surgery internships and compare them to expert-generated questions. We aim to determine whether AI can provide a valuable tool for medical education or if its current limitations make it unsuitable for unsupervised use. The results of this study could contribute to understanding the feasibility of AI-assisted education in thoracic surgery and inform best practices for AI integration into medical curricula.

METHODS

Ethical Considerations

The study was conducted following ethical guidelines for educational research and the principles of the Helsinki Declaration. Informed consent was obtained from all student participants. As the study did not involve patient data, ethics committee approval was not required under the Helsinki Declaration.

Study Design

This study was designed as a comparative analysis to evaluate the accuracy and usability of artificial intelligence (AI)-generated multiple-choice questions (MCQs) in thoracic surgery internship education. The study included an AI-generated exam and an expert-generated exam, both assessed for content accuracy by a panel of thoracic surgery specialists. The study also investigated the correlation between medical students' scores on both exams.

AI-Generated Exam

Claude Sonnet 3.7 AI was provided with the key topics covered in the 5th-year medical school thoracic surgery internship curriculum. The AI was instructed to generate a 20-question multiple-choice exam with five answer choices per question and an answer key. No additional instructions regarding difficulty level or question style were given.

Expert Evaluation and Classification of AI-Generated Questions

Three thoracic surgery specialists independently evaluated the AI-generated questions. The questions were classified into three categories:

Correct: No errors detected.

Minor error: Small mistakes that did not alter the meaning of the question or the correct answer.

Major Error: Errors that invalidated the question, including:

- Presence of multiple correct answers
- Incorrect AI-marked correct answer
- Contradictions with established medical knowledge
- Absence of the correct answer in the options

In addition to accuracy, the experts also classified each question by difficulty level (on a scale of 1 to 4) and topic (e.g., Pneumothorax, Pleural Effusion). Finally, a Delphi panel was conducted to reach a consensus on the classification of each question.

Expert-Generated Exam

A second 20-question MCQ exam was independently created by a thoracic surgery specialist, following the same curriculum and format as the AI-generated exam. This exam was also reviewed by the same three thoracic surgery experts using the Delphi method to ensure question validity.

Student Participation and Examination Process

Seven 5th-year medical students voluntarily participated in the study. Each student completed both the AI-generated and expert-generated exams under standardized testing conditions. A minimum 24-hour gap was maintained between the two exams to minimize recall bias.

Statistical Analysis

The median and interquartile range (IQR) of student scores were calculated for both exams. The correlation between students' scores on the AI-generated and expert-generated exams was assessed using Pearson's correlation coefficient. Fischer's exact test was used to compare the accuracy rates of the two exams. Statistical significance was set at $p<0.05$.

RESULTS

Evaluation of AI-Generated Questions

Out of the 20 multiple-choice questions (MCQs) generated by Claude Sonnet 3.7 AI, 8 questions (40%) contained major errors, while 1 question (5%) had a minor error. The breakdown of major errors is as follows:

- 3 questions (15%) contained two correct answers.
- 3 questions (15%) had a correct question, but the AI incorrectly marked the answer key.
- 2 questions (10%) presented medically inaccurate information, and the correct answer was missing from the answer choices.

The remaining 11 questions (55%) were classified as correct and free from any errors. The one minor error (5%) did not contradict general medical knowledge and did not change the correct answer. Therefore, it was included in the student assessment. In total, the student exam was conducted using 12 questions-11 correct and 1 with a minor error.

Evaluation of Expert-Generated Questions

The expert-generated exam underwent the same review process by the panel of three thoracic surgery specialists. No major or minor errors were identified in any of the 20 questions, indicating a perfect accuracy rate. As detailed in Table 4, the questions addressed a broad range of thoracic surgery topics, including pleural effusion, blunt trauma, pneumothorax, and esophageal cancer. The difficulty levels varied between 2 and 4, reflecting an appropriate range of complexity. These results underscore the reliability, clinical accuracy, and content diversity of expert-generated questions in medical education.

Student Performance Comparison

Seven 5th-year medical students participated in the study and completed both exams. The scores were analyzed as follows:

- The median score for the AI-generated exam was 75 (IQR: 67-100).
- The median score for the expert-generated exam was 85 (IQR: 70-95).

Although the AI-generated exam resulted in a slightly lower median score, there was variability among student performances.

Correlation Between AI and Expert-Generated Exam Scores

Statistical analysis using Pearson's correlation coefficient revealed no significant correlation between students' scores on the AI-generated and expert-generated exams (r=0.042, p=0.929) (Table 1). This suggests that the AI-generated exam did not consistently measure students' knowledge in a manner comparable to the expert-created exam.

Table 1. Comparison of AI-generated and expert-generated exam scores in thoracic surgery internship

	AI-generated exam score	Expert-generated exam score	p-value
Student 1	83	90	0.929
Student 2	67	95	
Student 3	92	90	
Student 4	100	85	
Student 5	67	85	
Student 6	75	70	
Student 7	67	85	

AI: Artificial intelligence

The expert-generated exam had a perfect accuracy rate, while the AI-generated exam showed significantly lower accuracy, with a statistically significant difference (p=0.001) (Table 2).

Table 2. Comparison of accuracy between expert-generated and AI-generated exams

	Correct questions (n)	Incorrect questions (n)	p-value
Expert-generated exam	20	0	0.001
AI-generated exam	11	9	

AI: Artificial intelligence

Out of 20 AI-generated questions, 11 were correct, 1 had a minor error, and 8 had major errors. Errors were more frequent in questions with higher difficulty levels (3-4), particularly in topics like lung cancer and pleural effusion. This indicates potential limitations of AI in complex or specialized medical domains (Table 3, 4).

DISCUSSION

AI has gained increasing attention in medical education, particularly in question generation, personalized learning, and decision support systems.[6-8] AI-driven educational tools, such as large language models (LLMs), have demonstrated potential in creating medical assessments, but their reliability remains a concern.[9,10] Our study evaluated the accuracy of

Table 3. Evaluation of AI-generated questions based on accuracy, difficulty, and topic

Question number	Accuracy	Difficulty level (1-4)	Topic
1	Major error	3	Blunt trauma
2	Correct	2	Blunt trauma
3	Major error	2	Penetrating trauma
4	Correct	2	Blunt trauma
5	Correct	2	Blunt trauma
6	Correct	3	Penetrating trauma
7	Correct	3	Pneumothorax
8	Correct	2	Pneumothorax
9	Correct	2	Pneumothorax
10	Minor error	2	Pneumothorax
11	Correct	3	Pneumothorax
12	Major error	2	Pneumothorax
13	Major error	2	Pleural effusion
14	Correct	3	Pleural effusion
15	Major error	3	Pleural effusion
16	Major error	3	Pleural effusion
17	Major error	4	Lung cancer
18	Major error	4	Lung cancer
19	Correct	2	Esophageal cancer
20	Correct	3	Esophageal cancer

AI: Artificial intelligence

Table 4. Evaluation of expert-generated questions based on accuracy, difficulty, and topic

Question number	Accuracy	Difficulty level (1-4)	Topic
1	Correct	2	Pleural effusion
2	Correct	3	Blunt trauma
3	Correct	2	Pneumothorax
4	Correct	2	Benign lung tumors
5	Correct	2	Esophageal cancer
6	Correct	2	Pneumothorax
7	Correct	3	Pleural effusion
8	Correct	2	Esophageal cancer
9	Correct	2	Pneumothorax
10	Correct	3	Blunt trauma
11	Correct	4	Blunt trauma
12	Correct	3	Pleural effusion
13	Correct	3	Foreign body aspiration
14	Correct	3	Lung abcess
15	Correct	3	Blunt trauma
16	Correct	3	Blunt trauma
17	Correct	4	Blunt trauma
18	Correct	3	Pneumothorax
19	Correct	3	Primary hyperhidrosis
20	Correct	2	Pleural effusion

AI-generated multiple-choice questions (MCQs) in thoracic surgery and compared student performance on AI-generated versus expert-generated exams. The results revealed a high major error rate (40%) in AI-generated questions, raising significant concerns about its unsupervised use in medical education.

Accuracy of AI-Generated Exam Questions

AI models have been praised for their ability to process vast amounts of medical knowledge quickly, yet their tendency to generate factually incorrect or misleading content limits their effectiveness.[11] Our findings align with previous studies that identified hallucinations (fabricated information presented as fact) in AI-generated medical content. The presence of multiple correct answers, incorrect answer keys, and medically inaccurate statements suggests that AI lacks contextual understanding and struggles with precise question formulation. The Delphi method, used in this study to assess question quality, confirmed that AI-generated exams contain errors that could mislead students and compromise medical training standards.

Another important finding of this study is the observed limitations of the AI model when generating questions related to real-life clinical reasoning and practical medical knowledge. The majority of major errors occurred in questions addressing applied clinical scenarios rather than purely theoretical content. This supports concerns that large language models, while effective in generating grammatically correct and seemingly plausible questions, may still fall short in domains requiring context-specific judgment or experiential understanding-especially in areas essential for junior medical assistants. Therefore, AI-generated content should be carefully reviewed before use in high-stakes educational settings, particularly when clinical decision-making is involved.

Comparison with Expert-Generated Questions

The expert-created exam had no major or minor errors, highlighting the superiority of human oversight in medical education. Expert validation ensures that questions align with evidence-based medicine, guidelines, and clinically relevant scenarios. The significantly lower error rate in expert-created exams reinforces the necessity of subject matter expertise in medical assessments.

Student Performance and Reliability of AI-Generated Exams

Despite the error-prone nature of AI-generated questions, student scores did not significantly correlate between the AI-generated and expert-generated exams ($r=0.042$, $p=0.929$). This suggests that AI-generated questions did not assess students' knowledge in the same manner as expert-designed exams. In contrast, studies have shown that expert-curated exams are better aligned with curriculum learning objectives and clinical competencies.[12,13] AI-based test generation tools must be refined to create consistent and standardized assessments.

Potential Role of AI in Medical Education

While AI-generated questions had a high error rate, AI could still be a valuable tool under expert supervision. AI may assist in generating a first draft of questions, which experts can refine for accuracy and clinical relevance. Previous research has demonstrated that AI can be useful for creating adaptive learning experiences and identifying knowledge gaps in students.[12,13] However, current AI technology is not yet reliable enough for unsupervised use in medical education.

Limitations

One limitation of this study is the small sample size of medical students (n=7), which may not fully represent the broader population. Additionally, only one AI model (Claude Sonnet 3.7 AI) was tested, and other LLMs, such as GPT-4 or Med-PaLM, might yield different results. In this context, it is important to note that the Claude Sonnet 3.7 model was not specifically trained by the authors using medical content or example questions; all generated questions were produced based solely on the model's pre-existing capabilities. This limits control over the content generation process and highlights the need for external validation mechanisms.

Future studies should explore larger student cohorts, test multiple AI models, and assess longitudinal performance improvements with AI-generated content. Furthermore, developing AI-guided question verification systems could mitigate the risk of erroneous content and enhance reliability in educational settings.

CONCLUSION

Our study demonstrates that AI-generated MCQs have a high error rate (40% major errors, 5% minor errors), making them unsuitable for standalone use in medical education. However, AI may have potential as a supplementary tool for question generation under expert supervision. Future advancements in AI technology, combined with rigorous human validation, could enhance the accuracy, reliability, and educational utility of AI-generated assessments. Until then, expert oversight remains essential to ensure high-quality medical education and patient safety.

ETHICAL DECLARATIONS

Ethics Committee Approval

As the study did not involve patient data, ethics committee approval was not required under the Helsinki Declaration.

Informed Consent

Informed consent was obtained from all student participants.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ*. 2025;25:129. doi:10.1186/s12909-025-06719-5
2. Ennab F, Farhan H, Zary N. Generative artificial intelligence and its role in the development of clinical cases in medical education: a scoping review protocol. *Preprints*. 2025. doi:10.20944/preprints202501.1031.v1
3. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ*. 2023;9:e48785. doi:10.2196/48785
4. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagn Interv Radiol*. 2025;31(2):75-88. doi:10.4274/dir.2024.242854
5. Colton S, Hatcher T. The web-based Delphi research technique as a method for content validation in HRD and adult education research. *Online Submission*. 2004.
6. Hicke Y, Geathers J, Rajashekar N, et al. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. *arXiv preprint arXiv:2503.05793*. 2025. doi:10.48550/arXiv.2503.05793
7. Hersh W. Generative artificial intelligence: implications for biomedical and health professions education. *arXiv preprint arXiv:2501.10186*. 2025. doi:10.48550/arXiv.2501.10186
8. Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. *J Adv Med Educ Prof*. 2023;11(3):133-140. doi:10.30476/JAMP.2023.98655.1803
9. Barile J, Margolis A, Cason G, et al. Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatr*. 2024;178(3):313-315. doi:10.1001/jamapediatrics.2023.5750
10. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. *Cureus*. 2023;15(11):e49604. doi:10.7759/cureus.49604
11. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq [Preprint]*. 2023:rs.3.rs-2566942. doi:10.21203/rs.3.rs-2566942/v1
12. Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ*. 2025;25(1):208. doi:10.1186/s12909-025-06796-6
13. Al Shuraiqi S, Aal Abdulsalam A, Masters K, Zidoum H, AlZaabi A. Automatic generation of medical case-based multiple-choice questions (MCQs): a review of methodologies, applications, evaluation, and future directions. *Big Data Cogn Comput*. 2024;8(10):139. doi: 10.3390/bdcc8100139