Original Article

HEALTH SCIENCES
MEDICINE

DOI: 10.32322/jhsm.1752483
J Health Sci Med. 2025;8(5):879-883

Evaluation of the performance of current Artificial Intelligence Chatbots regarding patient information after coronary artery bypass surgery

Gökhan Yüksel[1], Selami Gürkan[2]

[1]Department of Nursing, Faculty of Health Sciences, Tekirdağ Namık Kemal University, Tekirdağ, Turkiye
[2]Department of Cardiovascular Surgery, Tekirdag Namık Kemal University, Tekirdağ, Turkiye

Cite this article as: Yüksel G, Gürkan S. Evaluation of the performance of current Artificial Intelligence Chatbots regarding patient information after coronary artery bypass surgery. *J Health Sci Med.* 2025;8(5):879-883.

Received: 28.07.2025 ◆ **Accepted**: 25.08.2025 ◆ **Published**: 16.09.2025

ABSTRACT

Aims: This study aims to evaluate the performance of existing Artificial Intelligence (AI) driven Chatbots regarding patient information after coronary artery bypass (CABG) surgery.

Methods: On July 1, 2025, a standardized medical prompt concerning the recovery process after CABG was submitted to ten prominent AI Chatbots: GPT-4o, GPT 4.1, Grok-4, Claude Opus-4, DeepSeek R1, Gemini Pro, Microsoft Copilot, Llama 4, Mistral Large 2, and Perplexity Sonar. Each response was assessed using two validated scoring systems; the modified Ensuring Quality Information for Patients (mEQIP) and the newly developed Quality Analysis of Medical Artificial Intelligence (QAMAI). Readability was evaluated using the average reading level consensus (ARLC) calculator, which aggregates eight standard readability formulas.

Results: Among the tested Chatbots, Perplexity Sonar achieved the highest mEQIP score (91.7%) and the highest QAMAI score (29/30), while Gemini Pro received the lowest scores in both evaluations (72.2% mEQIP, 25/30 QAMAI). The average mEQIP score across all platforms was 80.43%, and the mean QAMAI score was 27/30, indicating generally high-quality responses. Readability assessment revealed that DeepSeek R1 provided the most comprehensible content (ARLC: 9.92, equivalent to a reading age of 15-16 years), while Llama 4 produced the most complex output (ARLC: 14.69, age 23+). The average ARLC across all Chatbots was 11.9, which corresponds to a college-level reading difficulty and exceeds the recommended sixth to eighth-grade readability level for patient education materials.

Conclusion: AI Chatbots show promising capabilities in delivering post-CABG patient information, often achieving high scores in quality assessments. However, inconsistencies remain in readability, completeness, and source transparency. Despite the increasing sophistication of AI-generated health information, the elevated reading levels and inconsistent citation practices may hinder accessibility for general patient populations. To enhance their role in patient education, future Chatbot iterations should prioritize user-centered design, medical guideline compliance, and content simplification.

Keywords: Coronary artery bypass grafting, Artificial Intelligence, Chatbots

INTRODUCTION

Coronary artery disease (CAD) continues to be one of the most important health problems of our time. In 2013, CAD was the predominant cause of death in Turkiye, accounting for 38.8% of mortalities.[1] This national burden is similar to trends in other high-income areas of the world: 2.3 million individuals in the United Kingdom (UK) have CAD. Coronary artery disease is the second leading cause of death in the UK.[2] Coronary Artery Bypass Grafting (CABG) is the primary intervention for severe CAD, including triple vessel disease and substantial left main stenosis.[3] While there has been a recent decline in all cardiac revascularization procedures, over 200,000 CABG surgeries are performed in the United States annually.[4] Given the serious consequences of incorrect or delayed treatment for CAD, ensuring the quality and accessibility of online information about CABG is crucial for helping many patients better understand the topic and for the postoperative care process for those undergoing this surgery.[5]

The internet has become a primary source of patient information, particularly during and after the COVID-19 pandemic, as people increasingly turn to it for timely health updates, and access to hospitals remains highly limited. More than 5% of internet searches are related to health. It is common for people to research symptoms or treatment plans online before consulting a doctor[6]. Although internet information is presented in several formats to accommodate individuals from varied backgrounds, the reliability of the sources remains uncertain. Misinformation may exacerbate

Corresponding Author: Gökhan Yüksel, gyuksel@nku.edu.tr

worry and uncertainty, impede patients from seeking prompt medical care, and erode trust in healthcare personnel.[7] Artificial Intelligence (AI)-driven Chatbots have emerged as a significant source of health information in the digital era.[8] These systems offer swift, anonymous assistance and practical guidance, making them particularly attractive to patients who are unable to consult a healthcare expert directly. Their accessibility and easy-to-use and accessible features have contributed to their increasing appeal in addressing health-related concerns. However, the accuracy, consistency, and quality of the information provided must be rigorously ensured to enable patients to use it effectively. AI-driven Chatbots may offer reliable disease-related information, evidence-based treatment options, and insights into potential issues.[9,10]

This study aims to evaluate and compare the performance, readability, and text quality of current AI Chatbots in patient information after CABG surgery.

METHODS

On July 1, 2025, the following prompt was entered one by one into 10 current Chatbots at Tekirdağ Namık Kemal University, Department of Cardiovascular Surgery, and the responses were recorded. Since no living or human data were used in this study, ethics committee approval was not obtained. All procedures were carried out in accordance with the ethical rules and the principles. Chatbots used: GPT-4o, GPT 4.1, Grok-4, Claude Opus-4, DeepSeek R1, Gemini Pro, Microsoft Copilot, Llama 4, Mistral Large 2, and Perplexity Sonar. New, unused accounts were opened, and access to licensed versions of the previously used ones was provided. A detailed prompt was established for CABG, directing each Chatbot to produce a medically precise and comprehensible text aimed at a public readership. The prompt was organized as follows: 'What should be done during the recovery process after cardiac bypass surgery (CABG)? Please provide information on the necessary steps to be taken during the recovery process. Kindly verify that the passage is medically precise and adheres to the most recent norms and recommendations in cardiovascular surgery. Can you use medical terminology while maintaining clarity for a general audience?'

The modified Ensuring Quality of Information for Patients (mEQIP) and Quality Analysis of Medical Artificial Intelligence (QAMAI) tools were used to assess the quality of the produced texts. The EQIP tool was developed in 2004 and comprises 36 items that assess specific characteristics of dependability, content quality, and structural readability.[11] It has been widely applied in recent years to evaluate information sources of different subspecialties in medicine, including surgery, dermatology, internal medicine, urology, and infectious diseases.[12-14] Only two yes-or-no alternatives were offered for each issue to eliminate assessor subjectivity in incomplete responses. The option "N/A" was offered for elements that were not relevant to the kind of source. The score is categorized as follows: 0-25%: critical quality issues; 26-50%: significant quality concerns; 51-75%: satisfactory quality with minor deficiencies; 76-100%: proficiently composed. This systematic scoring framework guarantees an impartial and uniform assessment of answer quality.

The QAMAI tool was created using the Modified DISCERN (mDISCERN) instrument. The mDISCERN is a well-established and extensively used instrument for evaluating the quality of health information disseminated via websites, social networks, YouTube, and other multimedia platforms. The use of mDISCERN for assessing information generated by AI is infeasible, as the tool considers specific human attributes, such as board certification and the credibility of the content provider, which do not apply to AI. In parallel with the mDISCERN, a panel of experts decided to develop a unidimensional construct of the instrument with six items, assessed by Likert Scales. The six domains of information quality were posited to be interrelated, like mDISCERN, reflecting a single dimension: the quality of the information's content. Each aspect was assessed using a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The QAMAI included six items: accuracy, clarity, relevance, completeness, sources, and usefulness. The scores were aggregated into a comprehensive metric (QAMAI score) that assessed the quality of the information. The total QAMAI score was evaluated between 6 and 30. Quality score assessments were scored, and consensus was reached by an experienced cardiovascular surgeon (SG) and an experienced cardiovascular clinic nurse (GY). Kappa statistics were used to assess the degree of concordance among the two evaluators (Pearson correlation coefficient r=0.976).

After the quality consensus among cardiovascular specialists, each passage generated by the Chatbot was assessed for reading difficulty using the average reading level consensus (ARLC) calculator. This online calculator, available at https://readabilityformulas.com/calculator-arlic-formula.php, computes the average of eight established readability formulae. The formulae include the Automated Readability Index, Flesch-Kincaid grade level, Flesch reading ease, Gunning Fog Index, Coleman-Liau Readability Index, SMOG Index, FORCAST Readability Formula, and Linsear Write Readability Index. The calculator then produces a difficulty score based on educational levels, spanning from very simple (first grade, ages 6-7) to highly tough (college graduate, age 23 and above).[15] Since the number of words, syllables, and sentences was used in these formulas, these were also noted separately.

RESULTS

Among the 10 Chatbots compared, Perplexity Sonar had the highest mEQIP score (91.7%) and Gemini Pro had the lowest (72.2%). The distribution of mEQIP scores by category is shown in **Table 1**.

Among the 10 Chatbots compared, like mEQIP scores, Perplexity Sonar had the highest QAMAI score (29) and again Gemini Pro had the lowest (25) score. The distribution of mEQIP scores by category is shown in **Table 2**.

When we examine the readability evaluation, Deepseek produced the most easily understandable text (ARLC: 9.92), while Llama was the Chatbot with the most complex language to understand (ARLC: 14.69) (**Table 3**).

Table 1. Comparison of mEQIP scores of Chatbots with subgroups

	Content (out of 36)	Identification (out of 12)	Structure (out of 24)	Total	%
GPT-4o	34	2	23	59	73.6
GPT 4.1	35	5	23	63	87.5
Grok-4	34	4	23	60	83.3
Claude Opus-4	31	2	22	55	76.4
DeepSeek R1	34	2	22	58	80.6
Gemini Pro	30	2	20	52	72.2
Microsoft Copilot	33	2	22	57	79.2
Llama 4	33	2	23	58	80.6
Mistral Large 2	32	2	23	57	79.2
Perplexity Sonar	34	10	22	66	91.7

mEQIP: Modified Ensuring Quality Information for Patients

Table 3. Comparison of language complexity criteria of Chatbots

	ARLC	Reading level	Age range
DeepSeek R1	9.92	Somewhat difficult	15-16
GPT 4.1	10.56	Fairly difficult	16-17
GPT-4o	10.94	Fairly difficult	16-17
Perplexity Sonar	11.15	Fairly difficult	16-17
Claude Opus-4	11.20	Fairly difficult	17-18
Grok-4	11.77	Difficult	17-18
Microsoft Copilot	12.14	Difficult	17-18
Mistral Large 2	12.66	Very difficult	18-20
Gemini Pro	13.80	Professional	21-22
Llama 4	14.69	Extremely difficult	23+

ARLC: Average reading level consensus

DISCUSSION

This study is among the first investigations in the literature to assess the quality of replies generated by various AI Chatbot models in the postoperative CABG surgery era. Our study's primary finding is that the Chatbots exhibit a wide range of reading and quality levels. These levels pertain to works comprehensible and readable by persons with a university-level education. The second primary result is that the Chatbots exhibit significant variations in the mEQIP and QAMAI scores. The Perplexity Sonar achieved a substantially superior mEQIP and QAMAI score compared to the other Chatbots, ChatGPT and Gemini. The mEQIP scores revealed that all Chatbots attained a mean score of 80.43%, and a QAMAI score of 27 (out of 30). When evaluating the quality of the texts, we revealed that current Chatbots produce more qualified content than their predecessors, and many of them exceed the upper quality threshold of 75%.[15,16] However, when we examine the language simplicity, the average ARLC score of 11.9 suggests that the texts are too academic and difficult to understand for the majority of the patients. If the text is written for patient education or public information purposes, further simplification is recommended. The American Medical Association (AMA) and the National Institutes of Health (NIH) recommend that health materials should be at a 6-8 grade level.[17] This is because many adults may struggle to understand complex terms or long sentence structures. As such, current Chatbots still use language structures that are too complex for patient information. Earlier research indicates that health literacy plays a crucial role in adherence to treatment and overall health outcomes.[18] This implies that the AI-generated materials currently available might not be beneficial for everyone, especially older adults or individuals with lower levels of education.

We assessed the quality, precision, and readability of the information offered by ten of the leading Chatbots in response to a standardized medical inquiry concerning postoperative care after CABG surgery. The results indicated that performance differed significantly among the platforms. Perplexity Sonar received the highest ratings in both the mEQIP and QAMAI assessments, while Gemini Pro consistently ranked last. These findings illustrate that the medical content produced by Chatbots is not always uniform, highlighting the necessity for caution when using AI-driven tools for patient education.

The mEQIP tool indicated that the majority of Chatbots delivered information ranging from good to excellent quality. Nevertheless, there were significant inconsistencies in the identification and structural aspects that impacted the overall usability. For instance, GPT 4.1 excelled in maintaining structure, while Gemini Pro struggled with clarity and completeness. It is crucial to recognize that if the information

Table 2. Comparison of QAMAI scores of Chatbots with subcategories

	Accuracy	Clarity	Relevance	Completeness	Sources	Usefulness	Total
GPT-4o	5	5	5	4	2	5	26
GPT 4.1	5	5	5	5	2	5	27
Grok-4	5	5	5	5	2	5	27
Claude Opus-4	5	4	5	5	1	5	25
DeepSeek R1	5	5	5	5	2	5	28
Gemini Pro	5	4	5	4	2	5	25
Microsoft Copilot	5	5	5	5	3	5	28
Llama 4	5	5	5	5	3	5	28
Mistral Large 2	5	5	5	4	3	5	27
Perplexity Sonar	5	4	5	5	5	5	29

QAMAI: Quality Analysis of Medical Artificial Intelligence

provided is not well-structured or sufficiently comprehensive, patients may struggle to understand or retain important recovery guidance.

Additionally, while some Chatbots provided well-structured and research-backed information, most responses lacked citations for their sources. Various studies examining AI Chatbots across different medical fields have also highlighted this issue.[19] This suggests that although AI-generated information might appear valuable, it is often difficult to authenticate. Only Perplexity Sonar consistently provided sources, with only a handful of others attempting to support their statements with references. Consequently, users had no clear way to verify the accuracy or recency of the information presented. This research builds on earlier findings that indicate AI-driven health communication tools are evolving rapidly, yet remain far from perfect. Chatbots such as GPT-4o and Grok-4 demonstrated strong performance in terms of accuracy and clarity; however, variations among different platforms highlight the necessity of ongoing monitoring.[5]

To improve current Chatbots' designs, various actions can be taken. Adaptive readability algorithms that adjust the text's difficulty according to the patient's reading level, ensuring medical accuracy is not compromised, should be implemented. Clear citation methods are required so that users can check the accuracy and timeliness of the material. Organized checklists or bullet-point summaries for important postoperative directives should be added to make the information more thorough. The ability to adapt to diverse cultures and languages, facilitating patient education that is sensitive to these differences, should be included. Lastly, user feedback loops should be utilized to enable Chatbots to continually improve based on patients' understanding and satisfaction with them.

Our findings indicate that while current Chatbots offer valuable information about post-CABG surgery, their reliability is compromised by several factors; they are often difficult to understand, fail to cite sources consistently, and lack consistent material. Targeted enhancements, such as adaptive readability restrictions, verified sources, and organized, patient-centered material, might make them an important part of postoperative treatment. Following health literacy standards and evidence-based criteria for outputs is a clear way to help patients understand, follow, and get better results. As digital health becomes more popular, improving the design of Chatbots is not simply a technological improvement; it is also an important step toward safe and fair patient education.

Limitations

Firstly, it assessed a singular standardized prompt so that outcomes may vary with alternative inquiries or clinical scenarios. Secondly, despite the enhanced interrater reliability, the scoring method retains an element of subjectivity. Third, readability formulae may inadequately reflect the clarity or cultural suitability of the content. Moreover, Chatbot replies may fluctuate over time, and the inclusion of only ten prevalent English-language models constrains generalizability. Subsequent research should examine various prompts, incorporate additional Chatbot systems, and analyze outcomes in diverse languages.

Although text quality measures like BLEU, ROUGE-L, BERTScore, and METEOR, along with factual accuracy and hallucination rates, are significant in NLP research, their use in patient-facing medical information remains unstandardized and unvalidated. The fact that these scores were not evaluated in our study is another limitation. Likewise, measurements of technical efficiency, such as response length, generating time, and repetition rate, were excluded, as these factors are not within the scope of our emphasis on the quality and readability of medical material. Finally, differences in model architecture, training data, and techniques such as reinforcement learning from human feedback (RLHF) or retrieval-augmented generation were not analyzed due to limited transparency from developers, and thus, performance differences could not be correlated with these technical specifications. Reproducibility is also limited because Chatbot APIs (Application Programming Interface) are frequently updated, and developers do not consistently disclose technical details such as API version or parameter settings.

CONCLUSION

AI Chatbots demonstrate potential in delivering postoperative information following CABG surgery; yet, their responses differ in quality, comprehensiveness, and readability. Although the majority provide generally dependable content, the reading level exceeds the suggested standards for patient education, and the citation of sources is variable. Future Chatbot development should prioritize enhanced clarity of language, consistent reference, adherence to therapeutic guidelines, and human oversight to augment their use.

ETHICAL DECLARATIONS

Ethics Committee Approval

No ethical approval was needed because this is not a human study, but only online information was used.

Informed Consent

No informed consent was obtained because no patient or their information was used.

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. Turkey Statistical Agency 2013 Statistical Analysis, 1 April 2014 issue: 16162.
2. Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. *Heart*. 2016;102(24): 1945-1952. doi:10.1136/heartjnl-2016-309573
3. Melly L, Torregrossa G, Lee T, Jansens JL, Puskas JD. Fifty years of coronary artery bypass grafting. *J Thorac Dis*. 2018;10(3):1960-1967. doi: 10.21037/jtd.2018.02.43
4. Montrief T, Koyfman A, Long B. Coronary artery bypass graft surgery complications: a review for emergency clinicians. *Am J Emerg Med*. 2018;36(12):2289-2297. doi:10.1016/j.ajem.2018.09.014
5. Wai Tai C, Wing Kwan L, Chan J, Angelini GD. Ensuring quality information for patients tool to assess patient information on CABG websites: systemic search and evaluation. *Perfusion*. 2025;40(6):1468-1476. doi:10.1177/02676591241303842
6. Hesse BW, Nelson DE, Kreps GL, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Arch Intern Med*. 2005;165(22):2618-2624. doi:10.1001/archinte.165.22.2618
7. Silver MP. Patient perspectives on online health information and communication with doctors: a qualitative study of patients 50 years old and over. *J Med Internet Res*. 2015;17(1):e19. doi:10.2196/jmir.3588
8. King MR. The future of AI in medicine: a perspective from a Chatbot. *Ann Biomed Eng*. 2023;51(2):291-295. doi:10.1007/s10439-022-03121-w
9. Şahin MF, Doğan Ç, Topkaç EC, et al. Which current Chatbot is more competent in urological theoretical knowledge? A comparative analysis by the European board of urology in-service assessment. *World J Urol*. 2025;43(1):116. doi:10.1007/s00345-025-05499-3
10. Şahin MF, Topkaç EC, Doğan Ç, et al. Still using only ChatGPT? The comparison of five different Artificial Intelligence Chatbots' answers to the most common questions about kidney stones. *J Endourol*. 2024; 38(11):1172-1177. doi:10.1089/end.2024.0474
11. Moult 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):165-175. doi:10.1111/j.1369-7625.2004.00273.x
12. Anıl H, Kayra MV. The digital dialogue on premature ejaculation: evaluating the efficacy of Artificial Intelligence-driven responses. *Int Urol Nephrol*. 2025;57(9):2829-2836. doi:10.1007/s11255-025-04461-x
13. Melloul E, Raptis DA, Oberkofler CE, Dutkowski P, Lesurtel M, Clavien PA. Donor information for living donor liver transplantation: where can comprehensive information be found? *Liver Transpl*. 2012;18(8):892-900. doi:10.1002/lt.23442
14. McCool ME, Wahl J, Schlecht I, Apfelbacher C. Evaluating written patient information for eczema in German: comparing the reliability of two instruments, DISCERN and EQIP. *PLoS One*. 2015;10(10):e0139895. doi:10.1371/journal.pone.0139895
15. Malak A, Şahin MF. How Useful are current Chatbots regarding urology patient information? Comparison of the ten most popular Chatbots' responses about female urinary incontinence. *J Med Syst*. 2024;48(1): 102. doi:10.1007/s10916-024-02125-4
16. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of Chatbot responses to physician questions. *JAMA Netw Open*. 2023;6(10): e2336483. doi:10.1001/jamanetworkopen.2023.36483
17. Wilson EA, Makoul G, Bojarski EA, et al. Comparative analysis of print and multimedia health materials: a review of the literature. *Patient Educ Couns*. 2012;89(1):7-14. doi:10.1016/j.pec.2012.06.007
18. 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(2):97-107. doi:10.7326/0003-4819-155-2-201107190-00005
19. 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(9):819-824. doi:10.1001/jamaophthalmol.2023.3119