

Artificial Intelligence and Guideline-Augmented Prompting in Assessing the Need for Preoperative Cardiology Consultation

Preoperatif Kardiyoloji Konsültasyonu Gerekliliğinin Değerlendirilmesinde Yapay Zeka ve Kılavuz Destekli Komut Yönlendirmesi

ABSTRACT

Objective: With the growing elderly population worldwide, the number of annual surgical procedures has risen substantially, leading to an increase in the demand for preoperative cardiology consultations. In parallel, recent years have witnessed remarkable innovations in cardiology driven by advances in artificial intelligence (AI) and machine learning (ML). In this study, we aimed to evaluate the performance of three widely used AI models: ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro, in assessing the necessity of cardiology consultation in preoperative patients and to explore the potential contribution of guideline-augmented prompting in this context.

Method: A council consisting of seven cardiologists and seven anesthesiologists was formed. Each physician evaluated 20 preoperative patient scenarios and provided recommendations on whether a separate cardiology consultation was necessary. For each case, the majority decision of the council was accepted as the reference standard. The same scenarios were presented to the three AI models, and their responses were recorded. Subsequently, the AI models with the highest concordance were integrated into the decision framework using guideline-augmented prompting, and the cases were re-evaluated.

Results: Although there was no statistically significant difference, ChatGPT-5 and Gemini 2.0 Pro showed higher concordance than Deepseek-V3 in preoperative consultation decisions ($\kappa = 0.706$ and $\kappa = 0.681$, respectively; 85% accuracy). Following the integration of guidelines into ChatGPT-5 and Gemini 2.0 Pro, the models were re-evaluated and demonstrated improved performance ($\kappa = 0.898$, 95% accuracy).

Conclusion: ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro demonstrated effectiveness in assessing the necessity of cardiology consultation in preoperatively evaluated patients. Moreover, the integration of guideline-augmented prompting was shown to improve the accuracy and reliability of AI model performance.

Keywords: Artificial intelligence, ChatGPT, machine learning, preoperative consultation

ÖZET

Amaç: Yaşlı nüfusun artmasıyla birlikte, tüm dünyada yıllık cerrahi işlemlerin sayısı da önemli ölçüde artmıştır. Bu durum preoperatif kardiyoloji konsültasyonlarının artışına neden olmuştur. Buna paralel olarak, son yıllarda yapay zeka (YZ) ve makine öğrenimi alanındaki gelişmelerin etkisiyle kardiyoloji alanında önemli yenilikler yaşanmıştır. Bu çalışmada, yaygın olarak kullanılan YZ modellerinden ChatGPT-5, Deepseek-V3 ve Gemini 2.0 Pro'nun preoperatif hastalarda kardiyoloji konsültasyonunun gerekliliğini değerlendirmedeki performansını değerlendirmeyi ve bu bağlamda kılavuz destekli komut yönlendirmesinin potansiyel katkısını araştırmayı amaçladık.

Yöntem: Yedi kardiyolog ve yedi anestezi uzmanından oluşan bir konsey oluşturuldu. Her hekim, 20 preoperatif hasta senaryosunu değerlendirdi ve kardiyoloji konsültasyonunun gerekli olup olmadığına dair önerilerde bulundu. Her vaka için konseyin çoğunluk kararı referans standart olarak kabul edildi. Aynı senaryolar üç YZ modeline sunuldu ve yanıtları kaydedildi. Ardından, en yüksek uyum gösteren YZ modellerine kılavuz destekli komut yönlendirmesi kullanılarak güncel kılavuzlar entegre edildi ve vakalar yeniden değerlendirildi.

Bulgular: İstatistiksel olarak anlamlı bir fark olmamasına rağmen, ChatGPT-5 ve Gemini 2.0 Pro, ameliyat öncesi konsültasyon kararında Deepseek-V3'ten daha yüksek uyum gösterdi ($\kappa = 0,706$, $\kappa = 0,681$; %85 doğruluk). Kılavuzların ChatGPT-5 ve Gemini 2.0 Pro'ya entegre edilmesinin ardından modeller yeniden değerlendirildi ve performanslarında iyileşme izlendi ($\kappa = 0,898$, %95 doğruluk).

ORIGINAL ARTICLE ARAŞTIRMA MAKALESİ

Mehmet Uğur Çalışkan¹

Ceren Yağmur Doğru Yılmaz²

Halenur Sarbaş³

Elmas Kaplan⁴

Ceren Özdemir Al⁵

Ertan Andaç Al⁵

¹Department of Cardiology, Kızılcahamam State Hospital, Ankara, Türkiye

²Department of Cardiology, Çorum Erol Olçok State Hospital, Çorum, Türkiye

³Department of Cardiology, Ardahan State Hospital, Ardahan, Türkiye

⁴Department of Cardiology, Çankırı State Hospital, Çankırı, Türkiye

⁵Department of Cardiology, Nazilli State Hospital, Aydın, Türkiye

Corresponding author:

Mehmet Uğur Çalışkan

✉ ugurkobian@gmail.com

Received: August 25, 2025

Accepted: December 25, 2025

Cite this article as: Çalışkan MU, Doğru Yılmaz CY, Sarbaş H, Kaplan E, Özdemir Al C, Al EA. Artificial Intelligence and Guideline-Augmented Prompting in Assessing the Need for Preoperative Cardiology Consultation. *Türk Kardiyol Dern Ars.* 2026;54(3):268-271.

DOI: 10.5543/tkda.2025.70041



Copyright@Author(s)

Available online at archivestsc.com.

Content of this journal is licensed under a Creative Commons Attribution - NonCommercial-NoDerivatives 4.0 International License.

Sonuç: ChatGPT-5, Deepseek-V3 ve Gemini 2.0 Pro'nun preoperatif hastalarda kardiyoloji konsültasyonu gerekliliğini değerlendirmedeki etkinliği kanıtlanmıştır. Kılavuz destekli komut yönlendirmesi YZ modellerinin doğruluğunu arttırmaktadır.

Anahtar Kelimeler: Yapay zekâ, ChatGPT, makine öğrenimi, preoperatif konsültasyon

Advances in modern medicine have significantly contributed to an increase in average life expectancy. As a consequence, the number of patients requiring major surgical procedures has also risen steadily over the years. It is estimated that worldwide, more than 300 million patients undergo surgical operations annually.¹ The number of preoperative cardiology consultations has also increased due to the higher prevalence of comorbidities in elderly patients and the associated increased risk of ischemic events.² The 2022 European Society of Cardiology (ESC) Non-Cardiac Surgery Guidelines and the 2024 American College of Cardiology (ACC) Non-Cardiac Surgery Guidelines, in conjunction with a multidisciplinary approach, outline which patients should be referred to cardiology for consultation during the preoperative assessment.^{1,3}

Artificial intelligence (AI), which has been gaining ground in recent years, is a computer technology that attempts to solve problems by means of human-like thinking abilities.⁴ AI has introduced key concepts such as machine learning (ML) and deep learning (DL). Although these terms are closely related and often used interchangeably, they represent different approaches and levels of complexity within the broader field of AI. ML is a technology that allows AI to generate insights or predictions from existing data by developing algorithms that best represent the underlying patterns within a dataset.⁵ DL can be defined as an advanced form of ML that enables more comprehensive analysis by combining more data.⁶ In the field of cardiology, AI and ML have increasingly become the focus of research, serving a variety of purposes such as the interpretation of radiological and echocardiographic images, automated analysis of electrocardiograms (ECG), and the management and risk stratification of specific patient populations.^{7,8} Several studies in the literature have investigated the feasibility and potential clinical utility of AI in the risk assessment of preoperative patients.⁹ However, to date, no clinical studies have specifically investigated the ability of AI and guideline-augmented prompting to evaluate the necessity of preoperative cardiology consultation.

In this study, we aimed to evaluate the performance of three widely used AI models (ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro) in assessing the necessity of cardiology consultation in preoperative patients and to explore the potential contribution of guideline-augmented prompting in this context.

Materials and Methods

A cardiologist developed 20 standardized patient scenarios that included detailed information on each patient's age, sex, diagnosis, planned surgical procedure, physical examination findings, medical history, and additional anamnesis questions (Supplementary Material: Patient Scenarios). These scenarios

ABBREVIATIONS

ACC	American College of Cardiology
AI	Artificial intelligence
DL	Deep learning
ECG	Electrocardiograms
ESC	European Society of Cardiology
ML	Machine learning

were designed to simulate real-world clinical conditions and served as the basis for evaluating the necessity of preoperative cardiology consultation. A council was established consisting of 14 physicians, including seven anesthesiologists from three different centers and seven cardiologists from four different centers. Each council member was individually presented with the patient scenarios and asked to provide a recommendation for each case in the format of "cardiology consultation required" or "not required." All members responded independently, indicating the decision they would make when evaluating the patient in each scenario, without following any predefined algorithm. The responses were compiled by an independent cardiologist, and for each scenario, the decision supported by the majority vote was accepted as the reference standard (Supplementary Material: Council Decision).

A standardized directive was provided to commonly used AI systems (ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro):

"I need your help. You are an anesthesiologist. I will give you 20 patient scenarios. In these scenarios, I want you to indicate which patients require preoperative consultation with cardiology."

Each model was tested separately using this standardized prompt (Supplementary Material: ChatGPT-5, DeepSeek-V3, and Gemini 2.0 Pro Answers). Following the initial evaluations, the responses obtained from the AI models were compared with the reference standard established by the expert panel. As a result of this analysis, the AI systems that demonstrated the closest agreement with the council's decisions were selected. Subsequently, the 2022 European Society of Cardiology (ESC) and 2024 American College of Cardiology/American Heart Association (ACC/AHA) Non-Cardiac Surgery Guidelines were incorporated into the models using guideline-augmented prompting.^{1,3} Integration was achieved by manually uploading the guideline documents into the system. The same 20 case scenarios were then presented again, and the model's guideline-based responses were obtained and compared with the initial responses. The models were also asked to indicate the differences compared to their previous assessments (Supplementary Material: ChatGPT-5 and Gemini 2.0 Pro with Integration of Clinical Guidelines).

Statistical Analysis

Data analysis was conducted using SPSS version 26.0 (IBM Corp., Armonk, NY, USA). For each AI model, the accuracy rate was calculated as the proportion of responses that were in complete agreement with the reference standard established by the expert council. Cohen's kappa coefficient was calculated to evaluate the models' agreement with the council's decisions. Kappa values were interpreted according to the classification of Landis and Koch (1977) as < 0.20 poor, 0.21-0.40 low, 0.41-0.60 moderate, 0.61-0.80 good, and 0.81-1.00 excellent agreement. The McNemar test was applied to compare the distribution of decisions among the AI models. In all statistical analyses, a p value < 0.05 was considered indicative of statistical significance.

Results

In the study, the expert council determined that cardiology consultation was necessary in 8 of the 20 patient scenarios. Among the artificial intelligence models, ChatGPT-5 recommended consultation in 11 cases, Deepseek-V3 in 10 cases, and Gemini 2.0 Pro in seven cases. The comparative distribution of the AI models' recommendations versus the council's reference decisions is summarized in Table 1.

The consistency of all three AI models with the council's reference decisions was found to be statistically significant (P < 0.05). Among the models, the highest consistency was observed with ChatGPT-5 (κ = 0.706, 85% accuracy). Although Gemini 2.0 Pro demonstrated the same accuracy rate as ChatGPT, its Cohen's kappa value was slightly lower (κ = 0.681, 85% accuracy). After integrating the 2022 ESC and 2024 ACC non-cardiac surgery guidelines into ChatGPT-5 and Gemini 2.0 Pro, both models were re-evaluated and showed improved performance (κ = 0.898, 95% accuracy).

When comparing the decision distributions among the models, no statistically significant differences were found between ChatGPT-5 and the other models (P > 0.05) (Table 2).

Discussion

In this study, we investigated the feasibility of using AI models and the impact of guideline-augmented prompting on assessing the necessity of preoperative cardiology consultation. Our findings demonstrated that all three widely used AI models: ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro, showed statistically significant consistency with the expert panel's decisions (P < 0.05), indicating that these models are feasible tools for supporting decision-making in this field. Although there was no statistically significant difference in performance when the models were compared directly (P > 0.05), ChatGPT-5 and Gemini 2.0 Pro demonstrated higher accuracy compared with DeepSeek-V3. The limited sample size remains a constraint and may affect the generalizability of these results. Moreover, integrating guideline-augmented prompting with current clinical guidelines further enhanced performance, increasing accuracy from 85% to 95%.

Recent clinical investigations have highlighted that ChatGPT may demonstrate superior performance compared to other general-purpose AI models, particularly in the evaluation of clinical case scenarios.^{10,11} In a study conducted by Pierri et al.,¹¹ the performance of ChatGPT-4o, Claude 3.5, and Gemini Flash 1.5 AI models were compared in the field of general cardiology using

Table 1. Comparison and compatibility analysis between artificial intelligence model decisions and council decisions

Model	Kappa	Agreement level	Accuracy (%)	P
ChatGPT-5	0.706	Good	85%	0.001
Deepseek-V3	0.600	Moderate	80%	0.006
Gemini 2.0 Pro	0.681	Good	85%	0.002
ChatGPT-5 (integrated guidelines)	0.898	Excellent	95%	<0.001
Gemini 2.0 Pro (integrated guidelines)	0.898	Excellent	95%	<0.001

Table 2. Comparison of decision distributions among artificial intelligence models

Compared models	P
ChatGPT-5 vs. Deepseek-V3	1.000
ChatGPT-5 vs. Gemini 2.0 Pro	0.125
Deepseek-V3 vs. Gemini 2.0 Pro	0.250

a set of 70 questions. The results demonstrated that ChatGPT-4o outperformed the other two models in terms of accuracy and consistency. However, no model demonstrated professional-level reliability. In another study conducted by Kozaily et al.,¹² the performances of ChatGPT-3.5 and Google Bard AI models were compared using 30 questions related to the diagnosis, prognosis, and treatment of heart failure. The findings indicated that ChatGPT provided more accurate responses overall. However, it was also observed that some recommendations generated by both models were either inconsistent with current clinical guidelines or contradicted real-world clinical practice. In our study, no statistically significant differences were found when comparing the decisions of the three models.

Machine learning and deep learning hold significant potential in the field of medicine. These approaches not only allow for comprehensive analysis of large-scale datasets but also facilitate the integration of years of accumulated clinical knowledge and experience. This enables the identification of previously unrecognized relationships within clinical decision-support processes and facilitates the development of predictive models. Moreover, these AI-based approaches not only enhance the processing and interpretation of existing data, but also support the generation of new hypotheses and the creation of predictive datasets applicable to a wide range of clinical scenarios.^{13,14} In a study conducted by Yoon et al.,¹⁵ preoperative anesthesia assessment notes for 717,389 patients were retrospectively reviewed. Using these notes, the ChatGPT-4, BioClinicalBERT, and ClinicalBigBird models were trained with ML methods, and their performances were subsequently compared with those of expert anesthesiologists and assistant anesthesiologists. The study demonstrated that AI models performed significantly better than assistant physicians and similarly well as specialist physicians. In our study, the same AI models were re-evaluated after the integration of guideline-augmented prompting. This approach demonstrated the positive impact of guideline-augmented prompting on model performance and outcome

accuracy. Both ChatGPT-5 and Gemini 2.0 Pro, when used with guideline-augmented prompting, produced identical responses. The only discrepancy from the council's decision occurred in Case 1, involving acute appendicitis. While the council did not request cardiology consultation due to the urgent nature of the surgery, the guideline-integrated AI systems recommended consultation based on the elevated perioperative cardiovascular risk. If AI models are to be consulted on specific topics, integrating relevant literature or data into the model beforehand increases the consistency and clinical applicability of the responses obtained.

Although advances in AI models are highly promising, they have also raised a number of ethical concerns and debates. In particular, sharing patient data with these systems introduces the risk of cyberattacks and data breaches, which may seriously compromise patient privacy, data security, and confidentiality.¹⁶ AI models such as ChatGPT, Deepseek, and Gemini are designed as general-purpose systems and therefore cannot be held accountable for incorrect or inappropriate decisions, a limitation that may pose a potential risk of harm to patients in clinical practice. However, although the ultimate responsibility for patient management lies with the physician, AI models can serve as valuable tools by facilitating information exchange on important cardiovascular diseases, providing clinical insights, and supporting decision-making processes.¹⁷ The results of our study also support the notion that AI systems supported by guideline-augmented prompting can assist physicians in preoperative assessment.

The primary limitation of this study is the limited sample size, which diminishes statistical power and reduces the generalizability of the findings. Larger-scale studies are needed to validate these results and further evaluate the clinical applicability of artificial intelligence-based decision support systems. Nevertheless, the novelty of being the first study to evaluate the necessity of cardiology consultation in preoperative patient assessment and to investigate the impact of guideline-augmented prompting using the same model underscores the significance and originality of our work.

Conclusion

Commonly used AI models such as ChatGPT-5, Deepseek-V3, and Gemini 2.0 Pro demonstrated effectiveness in assessing the necessity of cardiology consultation in preoperatively evaluated patients. Moreover, guideline-augmented prompting further improved the accuracy and reliability of AI model performance.

Online Supplementary Link: Supplementary may be accessed via this link.

Ethics Committee Approval: This study did not involve patient data; therefore, ethics committee approval was not required.

Informed Consent: Written informed consent was not required for this study.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: No financial support was received for the research, authorship, and/or publication of this article.

Use of AI for Writing Assistance: No use of AI-assisted technologies was declared by the authors.

Author Contributions: Concept – M.U.Ç.; Design – M.U.Ç.; Supervision – C.Y.D.Y.; Resource – E.K.; Materials – E.K.; Data Collection and/or Processing – C.Ö.A.; Analysis and/or Interpretation – H.S.; Literature Review – E.A.A.; Writing – M.U.Ç.; Critical Review – M.U.Ç.

Peer-review: Externally peer-reviewed.

Data Availability Statement: All data generated or analyzed during this study are available from the corresponding author upon reasonable request.

References

- Halvorsen S, Mehilli J, Cassese S, et al.; ESC Scientific Document Group. 2022 ESC Guidelines on cardiovascular assessment and management of patients undergoing non-cardiac surgery. *Eur Heart J*. 2022;43(39):3826-3924. Erratum in: *Eur Heart J*. 2023;44(42):4421. [CrossRef]
- Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators; Spence J, LeManach Y, Chan MTV, et al. Association between complications and death within 30 days after noncardiac surgery. *CMAJ*. 2019;191(30):E830-E837. [CrossRef]
- Writing Committee Members; Thompson A, Fleischmann KE, Smilowitz NR, et al. 2024 AHA/ACC/ACS/ASNC/HRS/SCA/SCCT/SCMR/SVM Guideline for Perioperative Cardiovascular Management for Noncardiac Surgery: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J Am Coll Cardiol*. 2024;84(19):1869-1969. Erratum in: *J Am Coll Cardiol*. 2024;84(24):2416.
- Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. 2017;69S:S36-S40. [CrossRef]
- Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol*. 2020;9(2):14.
- Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. *J Med Imaging Radiat Sci*. 2019;50(4):477-487. [CrossRef]
- Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in Cardiology. *J Am Coll Cardiol*. 2018;71(23):2668-2679. [CrossRef]
- Kaçar MN, Ulusoy İ, Yayla Ç. A Review on the Estimation of Coronary Fractional Flow Reserve Using Artificial Intelligence. *Turk Kardiyol Dern Ars*. 2025;53(4):275-280. [CrossRef]
- Abdel Malek M, van Velzen M, Dahan A, et al. Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study. *Br J Anaesth*. 2025;134(5):1333-1340. [CrossRef]
- Pay L, Yumurtaş AÇ, Çetin T, Çınar T, Hayıroğlu Mİ. Comparative Evaluation of Chatbot Responses on Coronary Artery Disease. *Turk Kardiyol Dern Ars*. 2025;53(1):35-43. [CrossRef]
- Pierri MD, Galeazzi M, D'Alessio S, et al. Evaluating Large Language Models in Cardiology: A Comparative Study of ChatGPT, Claude, and Gemini. *Hearts*. 2025;6(3):19. [CrossRef]
- Kozaily E, Geagea M, Akdogan ER, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. *Int J Cardiol*. 2024;408:132115. [CrossRef]
- Cascarano A, Mur-Petit J, Hernández-González J, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artif Intell Rev*. 2023;56(2):1711-1771. [CrossRef]
- Sadr H, Nazari M, Khodaverdian Z, et al. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *Eur J Med Res*. 2025;30(1):418. [CrossRef]
- Yoon SB, Lee J, Lee HC, Jung CW, Lee H. Comparison of NLP machine learning models with human physicians for ASA Physical Status classification. *NPJ Digit Med*. 2024;7(1):259. [CrossRef]
- Siafakas N, Vasarmidi E. Risks of Artificial Intelligence (AI) in Medicine. *Pneumon*. 2024;37(3):1-5. [CrossRef]
- Madaudo C, Parlati ALM, Di Lisi D, et al. Artificial intelligence in cardiology: a peek at the future and the role of ChatGPT in cardiology practice. *J Cardiovasc Med (Hagerstown)*. 2024;25(11):766-771. [CrossRef]

Artificial Intelligence and Guideline-Augmented Prompting in Assessing the Need for Preoperative Cardiology Consultation



20 patient scenarios



Expert Council
-7 cardiologists
-7 anesthesiologists



($\kappa=0.706$, 85% accuracy)

ChatGPT-5



($\kappa=0.600$, 80% accuracy)

Deepseek-V3



($\kappa=0.681$, 85% accuracy)

Gemini 2.0 Pro



ChatGPT-5 and Gemini 2.0 Pro
(with Guideline-Augmented Prompting)

($\kappa=0.898$, 95% accuracy)

Çalışkan, M. (2025) BioRender.