PREDICTABILITY OF THE NEED FOR CARDIOLOGY CONSULTATION WITH ARTIFICIAL INTELLIGENCE MODELS IN PATIENTS WHO ARE PLANNED FOR NON-CARDIAC SURGERY IN THE PREOPERATIVE PERIOD

INTRODUCTION

Preoperative cardiac risk assessment plays a critical role in reducing perioperative morbidity and mortality in patients undergoing non-cardiac surgery. This evaluation may require cardiology consultation and further examinations to determine the suitability of patients for surgery and to prevent possible cardiac complications (1). However, unnecessary requests for cardiology consultations may lead to inefficient use of health resources and delay of surgical procedures (2).

Modern artificial intelligence techniques, especially natural language processing and machine learning applications, offer more effective tools in preoperative risk assessment compared to traditional methods (3). In particular, large language models show successful results in perioperative risk stratification and postoperative outcome prediction using clinical notes and procedure descriptions obtained from the patient's electronic health records (4). In this context, preoperative risk assessment guidelines (e.g.European Society of Cardiology (ESC) 2024 non-cardiac surgery guideline) can help standardize this process (5). In light of these developments, AI-powered decision support systems can optimize anesthesia residents' decisions on the necessity of cardiology consultations, both improving patient safety and rationalizing resource utilization (6).

Large language models may experience complex situations in comprehending the given command and communicating with humans. In order to overcome this problem, language models are given prompts and learn on the database of the given texts, comprehend the relationships between the data and create patterns. In this way, the learning pattern obtained is more appropriate for the language models to make more appropriate inferences for the given questions.(7)

In this study, the effectiveness of artificial intelligence models in predicting the need for cardiology consultation in patients who are scheduled for non-cardiac surgery is evaluated and the effects of prompted and unprompted data entries on model performance are examined.

METHOD

This study was followed up in accordance with ethical standards and the Declaration of Helsinki. The study was approved by the ethics committee of the Ministry of Health University Bursa City Hospital on 11.12.2024.. .ASA class 1-4 patients evaluated in the anesthesia clinic in the preoperative period were included in the study. Patients over the age of 18 who were scheduled for non-cardiac surgery, who were treated with less than 2 years of experience, and who agreed to participate in the study for whom cardiology consultation was required by anesthesia assistants trained in anesthesiology were included in the study. Pediatric, geriatric and unwilling patient groups were excluded from the study.

The following data obtained during the standard perioperative evaluation process of each patient were recorded. Demographic information (age, gender, weight, height, BMI), medical history (existing diseases, allergy status, smoking, alcohol, substance use, surgeries, routinely used medications), METs (Metabolic Equivalents) score, ECG (Electrocardiography) findings, PA (Posterior Anterior) chest X-ray findings and the size of the surgery to be performed were recorded.

The patient data obtained were sent to Chat Gpt 4.5-5, Gemini 2.5 Flash-Pro, Deep seek, Co Pilot, Claude, Grok language model versions in the form of a patient scenario, first with prompts (you are an experienced anesthesiologist of 10 years, you met this patient in the outpatient clinic, according to the ESC 2024 guideline, can you evaluate this patient's need for cardiology consultation with its reasons?) and without prompts (Can you evaluate the patient's cardiology need in the given scenario?) and to make inferences. It was stated that the recommendations of artificial intelligence

applications should not be considered as medical consultations. Personal data was not shared with AI applications due to privacy of personal data; The shared data included the patient's age, comorbidities, pathological findings on physical examination, planned surgery, medications used, and surgical history. The answers given afterwards were recorded in texts and the Ateşman Turkish readability scoring was calculated to measure the intelligibility of the text. (8)(Table 1) The Global Quality Scale (GQS), which has been used in many previous studies, was used to evaluate the overall quality of the response content (9). (Table 2) . GQS is a Likert-type scale, with 1 point indicating low quality and 5 points indicating excellent quality. Responses with a score of 1 have poor flow of content, lack most information, and are not useful to physicians. The answers with 5 points are very useful answers for physicians, covering all the information with the perfect flow of content.

Existing patient scenarios were evaluated by two anesthesiologists with at least 10 years of experience in terms of cardiology consultation requirement according to the 2024 ESC guideline, and the results were recorded, and the decisions of the artificial intelligence models were asked to be evaluated with a general quality scoring (GQS) from 1 to 5, depending on the relevance of the consultation requirement results and the answer text to the subject in the scenario and the parallelism of the information given with the current guideline, without knowing which model the decisions of the artificial intelligence models were. noted.

Table 1. Rating of Turkish Texts According to Readability Number in Ateşman Readability Formula

	Readability Count
It's easy	90- 100
Easy	70- 89
Medium Difficulty	50- 69
Difficult	30- 49
Very Difficult	1- 29

Readability Number = 198.825–40.175xA-2.610xW

A: Average word length

B: Average sentence length

Table 2. Overall Quality Scoring Scoring Score

Score	Definition of quality
1	The content is not suitable for use by physicians with low information quality and fluency
2	The content information quality and fluency are low, some information on the subject is basically useful for existing physicians.
3	The content information quality and fluency are moderate, it contains some important information but does not provide sufficient information, it is fundamentally useful to physicians.
4	The content is useful for physicians with good information quality and fluency, the majority of the information on the subject has been presented.

5	The content quality and fluency are excellent, it provides full competence on the subject, it is very useful for physicians.

STATISTICS

The sample size of the study was determined using the G*Power (v3.1.9.4) program. The effect size value was calculated as f=0.36, and it was decided to carry out the study for a minimum of n=176 patients at the relevant effect size value, α=0.05, 95% potency.

In the study, Cohen's Kappa analysis was applied to assess the consistency of the requirement for cardiology consultation by two independent experts. The conformity of continuous variables to normal distribution was examined with the Shapiro Wilk test. According to the results of the normality test, the variables that conform to the normal distribution are determined by using the mean ± standard deviation and the variables that do not conform to the normal distribution are determined by using the median (minimum: maximum) values; categorical variables are expressed as n (%). In the comparisons made between the two groups, the Mann Whitney U test was used if there was no conformity to the normal distribution. According to the normality test results, the Kruskal Wallis test was used if the number of groups was more than two and normal distribution was not observed. In case of general significance after the Kruskal-Wallis test, subgroup analyzes were carried out using the Dunn-Bonferroni test. Categorical variables were compared between the groups using the Pearson Chi-Square test. The performance of artificial intelligence models in classifying the necessity of cardiology consultation was evaluated with ROC (Receiver Operating Characteristic) analysis. Within the scope of ROC analysis, the discrimination power of the models was compared by calculating the area under the curve (AUC), sensitivity and specificity values. The level of agreement between the expert (Dr-1) and the artificial intelligence models (ChatGPT 4.5, ChatGPT 5, Copilot, Deepseek, Grok, Claude, Gemini Flash, Gemini Pro) was evaluated with Cohen's Kappa coefficient (κ) and the interpretation was based on the Landis and Koch (1977) classification. SPSS for statistical analysis (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.) program was used and the type I error rate was accepted as 5%.

CONCLUSION

A total of 182 patients were included in the study over a 6-month period. The mean age of the patients was 65.06 years.

Table 3. Agreement levels of artificial intelligence models between the common doctor's opinion of Cardiology Consultation Requirements with and without prompts

Promptlu	Kappa (95 % CI)	p value	Accuracy
• ChatGPT 4.5	0.44 (0.31–0.57)	<0.001	76.4%
• ChatGPT 5	0.44 (0.31–0.57)	<0.001	76.4%
• Copilot	0.53 (0.37–0.68)	<0.001	84.6%
• Deepseek	0.16 (0.02–0.31)	0.001	80.2%
• Grok	0.40 (0.23–0.56)	<0.001	84.6%
• Claude	0.40 (0.23–0.56)	<0.001	84.6%
• Gemini flash	0.47 (0.35–0.59)	<0.001	75.8%
• Gemini pro	0.47 (0.35–0.59)	<0.001	75.8%
Promptsuz			
• ChatGPT 4.5	0.42 (0.26–0.59)	<0.001	82.4%
• ChatGPT 5	0.42 (0.26–0.59)	<0.001	82.4%

• Copilot	0.41 (0.27–0.56)	<0.001	77.5%
• Deepseek	0.67 (0.53–0.80)	<0.001	89%
• Grok	0.20 (0.04–0.35)	<0.001	80.8%
• Claude	0.20 (0.04–0.35)	<0.001	80.8%
• Gemini flash	0.08 (−0.07–0.23)	0.266	66.5%
• Gemini pro	0.08 (−0.07–0.23)	0.266	66.5%

Kappa (95 % CI)

Prompted responses;

- The agreement between the common physician opinion and ChatGPT-4.5 and 5's classifications of cardiology consultation requirement reveals a moderate and statistically significant agreement.
- It reveals that there is a good and statistically significant agreement between the common physician opinion and the Copilot model's classifications of cardiology consultation requirement.
- It reveals that the common physician opinion and the Deepseek model show a low but statistically significant agreement as a result of the agreement between the cardiology consultation requirement classifications.
- It reveals that the common physician opinion and the Grok model show a moderate and statistically significant agreement as a result of the agreement between the cardiology consultation requirement classifications.
- It reveals that there is a moderate and statistically significant agreement between the common physician opinion and the Claude model as a result of the agreement between the classifications of the necessity of cardiology consultation.
- It reveals that there is a moderate and statistically significant agreement between the common physician opinion and the Gemini Flash and Pro model classifications of cardiology consultation requirement.

Responses without prompts;

- It reveals that the common physician opinion and the non-prompt responses of ChatGPT-4.5 and 5 show a moderate and statistically significant agreement as a result of the agreement between the cardiology consultation requirement classifications.
- It reveals that there is a moderate and statistically significant agreement between the common physician opinion and the non-prompt responses of the Copilot model as a result of the agreement between the cardiology consultation requirement classifications.
- It reveals that the unprompted responses of the common physician opinion and the Deepseek model show a good and statistically significant agreement as a result of the agreement between the classifications of the necessity of cardiology consultation.
- It reveals that the common physician opinion and the Grok model's unprompted responses show a low but statistically significant agreement as a result of the agreement between the cardiology consultation requirement classifications.
- It reveals that there is a low but statistically significant agreement between the common physician opinion and the non-prompt responses of the Claude model as a result of the agreement between the classifications of the necessity of cardiology consultation.

- It reveals that the common physician opinion and the non-prompt responses of the Gemini Flash and Pro model showed a low and statistically insignificant agreement as a result of the agreement between the cardiology consultation requirement classifications.

Table 4. Comparison of Ateşman Readability Scores with Prompted and Non-Prompt Responses

	Promptlu	Promptsuz	p value
ChatGPT 4.5	55.6(38-81.4)	55.6(38-81.4)	>0.99a
ChatGPT 5	63.4(47.6-87.4)	63.4(47.6-87.4)	>0.99a
Copilot	76.1(47-91.3)	61.1(43.6-76)	<0.001a
Deepseek	67.4(45.9-87.5)	80.2(73.4-85.4)	<0.001a
Grok	73.5(44.2-94)	66.5(57.3-76.8)	<0.001a
Claude	67.4(45.9-87.5)	67.4(45.9-87.5)	>0.99a
Gemini flash	55.6(38-81.4)	59.3(24.3-114)	<0.001a
Gemini pro	59.3(24.3-114)	65.3(45.3-98.3)	<0.001a

The data are expressed as median (minimum: maximum).
a: Mann Whitney U test

- In the evaluation made in terms of readability scores, it was determined that there was no significant difference between the prompted and non-prompted responses of ChatGPT 4.5 and ChatGPT 5 models (p>0.99).
- The median readability score of the prompted responses of the Copilot model was 76.1 (minimum: 47- maximum: 91.3) and the median readability score of the non-prompted responses was 61.1 (minimum: 43.6- maximum: 76). It was determined that the use of prompts in the Copilot model significantly increased the median readability scores (p<0.001).
- The median readability score of the prompted responses of the Deepseek model was 67.4 (minimum: 45.9- maximum: 87.5) and the median readability score of the non-prompted responses was 80.2 (minimum: 73.4- maximum: 85.4). It was determined that the use of the Deepseek model without prompts significantly increased the median readability scores (p<0.001).
- The median readability score of the prompted responses of the Grok model was 73.5 (minimum: 44.2- maximum: 94) and the median readability score of the unprompted responses was 66.5 (minimum: 57.3- maximum: 76.8). In the Grok model, it was determined that the use of prompts significantly increased the median readability scores (p<0.001).
- In the evaluation made in terms of Ateşman readability scores, it was determined that there was no significant difference between the prompted and non-prompted responses of the Claude model (p>0.99).
- The median readability score of the prompted responses of the Gemini flash model was 55.6 (minimum: 38- maximum: 81.4) and the median readability score of the non-prompted responses was 59.3 (minimum: 24.3- maximum: 114). It was determined that using the Gemini flash model without prompts significantly increased the median readability scores (p<0.001).
- The median readability score of the prompted responses of the Gemini pro model was 59.3 (minimum: 24.3- maximum: 114) and the median readability score of the

non-prompted responses was 65.3 (minimum: 45.3- maximum: 98.3). It was determined that using the Gemini pro model without prompts significantly increased the median readability scores (p<0.001).

Table 5. Comparison of Quality Scores with Prompted and Non-Prompt Responses

	Promptlu	Promptsuz	p value
ChatGPT 4.5	4(1-5)	4(1-5)	0.078a
ChatGPT 5	4(1-5) 3.31±1.11	4(1-5) 3.63±0.95	0.001a
Copilot	4(1-5)	4(1-5)	0.271a
Deepseek	4(1-5)	4(1-5)	0.080a
Grok	4(1-5)	4(1-5)	0.503a
Claude	4(1-5)	4(1-5)	0.256a
Gemini flash	4(1-5)	3(1-4)	0.001a
Gemini pro	4(1-5)	3(1-5)	0.016a

The data is mean±std. deflection and median (minimum: maximum).
a: Mann Whitney U test

- The average quality score of the prompted responses of the ChatGPT 5 model was determined as 3.31±1.11 and the average quality score of the non-prompted responses was determined as 3.63±0.95. It was determined that the use of the ChatGPT 5 model without prompts significantly increased the average quality scores (p=0.001).
- The median quality score of the prompted responses of the Gemini flash model was determined as 4 (minimum: 1- maximum: 5) and the median quality score of the non-prompted responses was determined as 3 (minimum: 1- maximum: 4). In the Gemini flash model, prompt use was found to significantly increase median quality scores (p=0.001).
- The median quality score of the prompted responses of the Gemini pro model was determined as 4 (minimum: 1- maximum: 5) and the median quality score of the non-prompted responses was determined as 3 (minimum: 1- maximum: 5). In the Gemini pro model, prompt use was found to significantly increase median quality scores (p=0.016).

Table 6. Comparison of Prompted and Non-Prompt Responses of Artificial Intelligence Models According to Ateşman Readability Scores

	Chatbot								
	ChatGPT 4.5	ChatGPT 5	Copilot	Deepseek	Grok	Claude	Gemini flash	Gemini pro	p value
Promptlu Readability	55.6(38-81.4)	63.4(47.6-87.4)	76.1(47-91.3)	67.4(45.9-87.5)	73.5(44.2-94)	67.4(45.9-87.5)	55.6(38-81.4)	59.3(24.3-114)	<0.001b
Promptsuz Readability	55.6(38-81.4)	63.4(47.6-87.4)	61.1(43.6-76)	80.2(73.4-85.4)	66.5(57.3-76.8)	67.4(45.9-87.5)	59.3(24.3-114)	65.3(45.3-98.3)	<0.001b

The data are expressed as median (minimum: maximum).
b: Kruskal Wallis test

According to the fire readability scores of the prompted responses of the artificial intelligence models, the median readability score of the ChatGPT 4.5 model was 55.6 (minimum: 38- maximum: 81.4), the median readability score of the ChatGPT 5 model was 63.4 (minimum: 47.6- maximum: 87.4), the median readability score of the Copilot model was 76.1 (minimum: 47- maximum: 91.3), the median readability score of the Deepseek model was 67.4 (minimum: 45.9- maximum: 87.5), and the median readability score of the Grok model was 73.5 (minimum: 44.2- maximum: 94), the median readability score of the Claude model was 67.4 (minimum: 45.9- maximum: 87.5), the median readability score of the Gemini flash model was 55.6 (minimum: 38- maximum: 81.4), and the median readability score of the Gemini pro model was 59.3 (minimum: 24.3- maximum: 114), and there was a significant difference between the groups ($p<0.001$). Within the scope of subgroup analyses, it was determined that the median readability scores of ChatGPT 4.5 and Gemini flash models were lower than other models ($p<0.001$). The median readability score of the Copilot model was found to be higher than the readability scores of all other models ($p<0.001$).

According to the fire readability scores of unprompted responses belonging to artificial intelligence models, the median readability score of the ChatGPT 4.5 model was 55.6 (minimum: 38- maximum: 81.4), the median readability score of the ChatGPT 5 model was 63.4 (minimum: 47.6- maximum: 87.4), the median readability score of the Copilot model was 61.1 (minimum: 43.6- maximum: 76), the median readability score of the Deepseek model was 80.2 (minimum: 73.4- maximum: 85.4), and the median readability score of the Grok model was 66.5 (minimum: 57.3- maximum: 76.8), the median readability score of the Claude model was 67.4 (minimum: 45.9- maximum: 87.5), the median readability score of the Gemini flash model was 59.3 (minimum: 24.3- maximum: 114), and the median readability score of the Gemini pro model was 65.3 (minimum: 45.3- maximum: 98.3), and there was a significant difference between the groups ($p<0.001$). Within the scope of subgroup analyses, it was determined that the median readability scores of ChatGPT 4.5 and Gemini flash models were lower than other models ($p<0.001$). It was determined that the median readability score of the Deepseek model was higher than the readability scores of all other models ($p<0.001$).

Table 7. Comparison of Prompted and Non-Prompt Responses of Artificial Intelligence Models According to Consultation Status

	Chatbot								
	ChatGPT 4.5	ChatGPT 5	Copilot	Deepseek	Grok	Claude	Gemini flash	Gemini pro	p value
Promptlu Consultation required	116(63.7%)	116(63.7%)	147(80.8%)	175(96.2%)	169(92.9%)	169(92.9%)	105(57.7%)	105(57.7%)	<0.001c
Promptsuz Consultation required	153(84.1%)	153(84.1%)	174(95.6%)	128(70.3%)	145(79.7%)	134(73.6%)	134(73.6%)	174(95.6%)	<0.001c

The data is expressed in n%.

c: Chi-Square test

When the prompted responses of the artificial intelligence models were evaluated according to the need for cardiology consultation, it was determined that there was a significant difference between the artificial intelligence groups ($p<0.001$). It has been determined that the Deepseek model's referral rate to patients requiring cardiology consultation is higher compared to all other artificial intelligence models. On the other hand, it was determined that the rate of referring patients requiring cardiology consultation of Gemini flash and Gemini pro models was lower compared to all other artificial intelligence models.

When the unprompted responses of the artificial intelligence models were evaluated according to the need for cardiology consultation, it was determined that there was a significant difference between the artificial intelligence groups ($p<0.001$). Copilot and Gemini pro moedels have been found to have a higher rate of referring patients requiring cardiology consultation compared to all other AI models.

Table 8. Comparison of Prompted and Non-Prompt Responses of Artificial Intelligence Models with GQS

	Chatbot								
	ChatGPT 4.5	ChatGPT 5	Copilot	Deepseek	Grok	Claude	Gemini flash	Gemini pro	p value
Promptlu									
GQS	4(1-5) 3.3±1.1	4(1-5) 3.3±1.1	4(1-5) 3.5±1.1	4(1-5) 3.7±1.1	4(1-5) 3.7±0.9	4(1-5) 3.7±1	4(1-5) 3.3±1.1	4(1-5) 3.4±1.2	<0.001b
Promptsuz									
GQS	4(1-5) 3.5±1	4(1-5) 3.6±0.9	4(1-5) 3.6±1.2	4(1-5) 3.9±1	4(1-5) 3.7±1.2	4(1-5) 3.6±1.1	3(1-4) 2.9±1	3(1-5) 3.1±1	<0.001b

The data are expressed as median (minimum: maximum).

b: Kruskal Wallis test

According to the GQS scores of the prompted responses of artificial intelligence models, the average GQS score of the ChatGPT 4.5 model was 3.3±1.1, the average GQS score of the ChatGPT 5 model was 3.3±1.1, the average GQS score of the Copilot model was 3.5±1.1, the average GQS score of the Deepseek model was 3.7±1.1, the average GQS score of the Grok model was 3.7±0.9, the average GQS score of the Claude model was 3.7±1, the average GQS score of the Gemini flash model was 3.3±1.1, and the average GQS score of the Gemini pro model was 3.4±1.2. significant difference ($p<0.001$). Within the scope of subgroup analyses, it was determined that the mean GQS scores of Deepseek, Grok and Claude models were higher than other models ($p<0.001$). The average GQS scores of ChatGPT 4.5, ChatGPT 5, and Gemini flash models were found to be lower than other models ($p<0.001$).

According to the GQS scores of the non-prompt responses of artificial intelligence models, the average GQS score of the ChatGPT 4.5 model was 3.5±1, the average GQS score of the ChatGPT 5 model was 3.6±0.9, the average GQS score of the Copilot model was 3.6±1.2, the average GQS score of the Deepseek model was 3.9±1, the average GQS score of the Grok model was 3.7±1.2, the average GQS score of the Claude model was 3.6±1.1, the average GQS score of the Gemini flash model was 2.9±1.1, and the average GQS score of the Gemini pro model was 3.1±1. significant difference ($p<0.001$). Within the scope of the subgroup analysis, it was determined that the mean GQS score of the Deepseek model was higher than all other models ($p<0.001$). The average GQS score of the Gemini flash model was found to be lower than the other models ($p<0.001$).

As a result of the ROC analysis, which was carried out by accepting the evaluation of an anesthesiologist with 10 years of experience as the gold standard, the performance of artificial intelligence models in classifying the necessity of cardiology consultation was examined.

Table 9. ROC Analysis Results of Prompted Artificial Intelligence Models (Gold Standard: Dr-1)

Model	AUC	Sensitivity	Specificity	p value
ChatGPT 4.5	0.775	0.755	0.795	<0.001
ChatGPT 5	0.775	0.755	0.795	<0.001
Copilot	0.753	0.916	0.590	<0.001
Deepseek	0.557	0.986	0.128	0.275
Grok	0.651	0.993	0.308	0.004
Claude	0.651	0.993	0.308	0.004
Gemini flash	0.819	0.714	0.924	<0.001
Gemini pro	0.819	0.714	0.924	<0.001

The highest AUC value was achieved in the Gemini Flash and Gemini Pro models (AUC = 0.819), which best classify Dr-1's consultation decisions. The sensitivity (0.714) and specificity (0.924) values of these models were balanced and it was determined that they had a statistically significant discrimination power (p<0.001).

Gemini Flash, Gemini Pro, ChatGPT 4.5, ChatGPT 5, Copilot, Grok, and Claude differ significantly at the AUC>0.5 level (p<0.05).

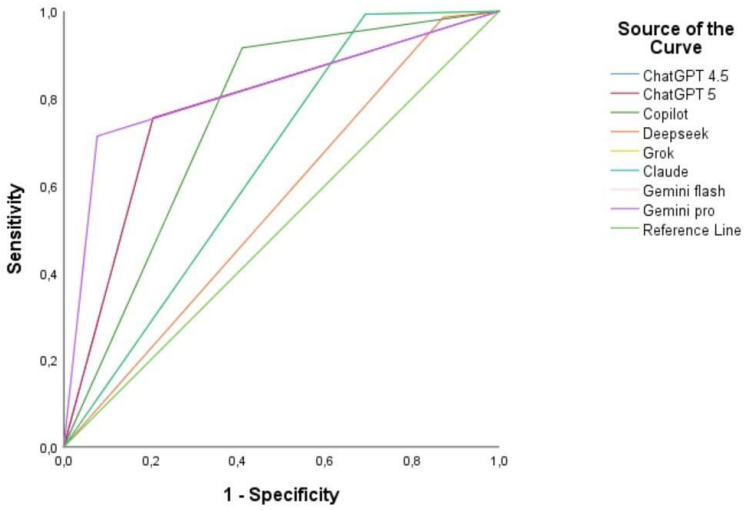

Fig 1. Promptlu….

Table 10. ROC Analysis Results of Promptless Artificial Intelligence Models (Gold Standard: Dr-1)

Model	AUC	Sensitivity	Specificity	p value
ChatGPT 4.5	0.692	0.923	0.462	<0.001
ChatGPT 5	0.692	0.923	0.462	<0.001
Copilot	0.735	0.804	0.667	<0.001
Deepseek	0.828	0.937	0.717	<0.001
Grok	0.570	0.986	0.154	0.181
Claude	0.570	0.986	0.154	0.181
Gemini flash	0.545	0.755	0.333	0.397
Gemini pro	0.545	0.755	0.333	0.397

The Deepseek model (AUC = 0.828) is the model that best classifies Dr-1's consultation decisions.

Deepseek, Copilot, ChatGPT 4.5, and ChatGPT 5 differ significantly at AUC>0.5 ($p<0.05$).

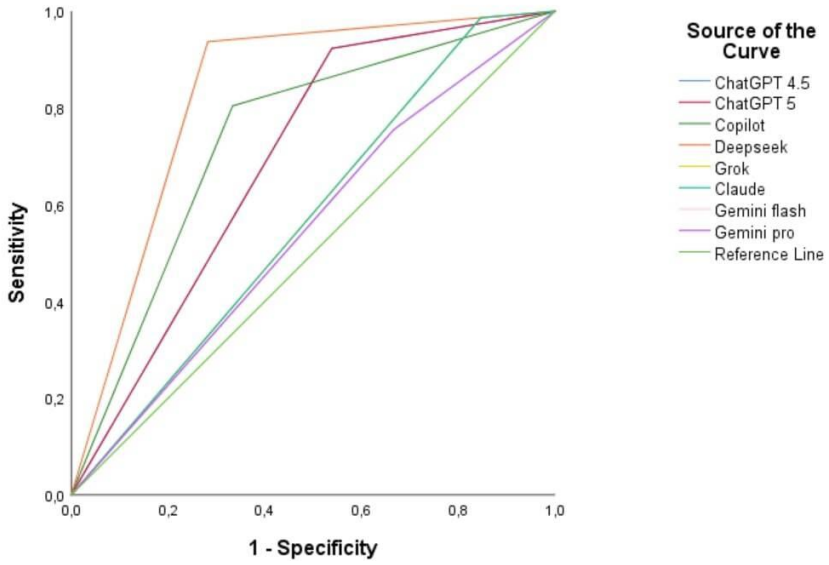

Fig 2. Promptsuz…

DISCUSSION

In this study, which evaluated the ability of artificial intelligence applications to analyze patient physical examination, symptoms, demographic data and determine the need for cardiology consultation, it was observed that these applications generally made different preferences to the preferences of anesthesiologists. This study comprehensively evaluated the performance of various AI models in assessing the necessity of perioperative cardiology consultation by comparing them with the common clinical decisions of anesthesiologists with at least 10 years of experience. Our findings revealed that different AI language models exhibit significant differences in performance, readability, and quality. Additionally, prompt usage has been observed to have variable effects on model performance and the quality of its responses.

One of the most striking findings of our study is that the Deepseek model achieved both the highest coefficient of agreement (κ=0.67) and the highest diagnostic accuracy (89%) in prompt-free use. The same model also achieved the highest AUC value (0.828) in prompt-free mode, demonstrating the

classification performance closest to expert clinical judgment. This result indicates that certain AI language models can provide a high level of clinical insight with minimal prompting. In contrast, the same model showed a low kappa value (κ=0.16) when prompted, suggesting that the performance of some models may be adversely affected by prompt intervention.

The Gemini series models (Flash and Pro) achieved the highest AUC values (0.819) and near-perfect specificity (0.924) in prompted use. This suggests that these models can be particularly valuable in clinical scenarios where unnecessary consultations are sought to be minimized. However, this high specificity is associated with a decrease in sensitivity (0.714). In the non-prompt mode, the performance of the same models loses its statistical significance (p=0.266) and the decrease in AUC values suggest that the structure of these models may change inferences depending on detailed prompts.

In terms of readability, Copilot received the highest median score (76.1) for prompted use, while Deepseek excelled for unprompted use (80.2). Considering that readability is important for rapid information learning, especially in busy clinical settings, these findings may provide guidance on which model will produce a more useful response under which conditions. In the Overall Quality Scores (GQS) assessment, the Deepseek, Grok, and Claude models consistently scored higher in both prompted and non-prompted scenarios. This shows that the answers produced by these models are of higher quality not only in terms of accuracy but also in terms of scope, usefulness, and presentation.

The wide variance in consultation referral rates (between 57.7% and 96.2% in prompted mode) is an indication that the thresholds for interpreting clinical status and predicting risk for different AI language models can be very different from each other. For instance, Deepseek exhibits a very high referral rate (96.2%) in prompted mode, suggesting that the model may adopt an "overly cautious" approach, while the Gemini series is more selective.

In conclusion, this study demonstrated that no AI language model provides absolute superiority in all metrics. The choice of model for clinical use will depend on the intended use:

In situations requiring High Sensitivity (e.g., screening high-risk patients), Deepseek or ChatGPT 4.5/5 without prompts may come to the fore.

In cases requiring High Specificity (to reduce unnecessary consultations), Gemini Pro/Flash with prompts may be more suitable.

If a balanced Performance and High Readability are desired, Deepseek without prompts or Copilot with prompts can be considered.

These are a general inference of the statistical findings obtained after our study. One of the main limitations of our study is that data can be obtained from a single center and through a specific clinical scenario (cardiology consultation). Additionally, due to the variable nature of AI language models, these results may change with the model's updates. Future studies should focus on larger patient populations, different surgical branches, and more specific prompt engineering strategies.

Confidence in AI models is likely to increase as clinical users understand the potential and limitations of AI models, especially as they observe that they are successfully applied in real patient scenarios and as it becomes easier for the user to access the data they want (10)

Kambale and Jadhaw predicted that AI would contribute to standardizing anesthesia methods and reducing human-induced errors. This highlights the potential of AI to enhance the efficiency and safety of medical procedures by leveraging consistent data-driven decision-making processes (11). This suggests that in the future, anesthesiologists will be able to use AI applications to quickly review and validate their administers' decisions.

Although there are publications in the literature on drug infusion systems, pain management, and the choice of anesthesia type based on patient data of artificial intelligence, we did not come across any studies that determined the need for preoperative cardiology consultation with artificial intelligence language models in our literature review. As noted in the study by Singh and Nath and Bellini et al., the impact of AI in clinical practice is often limited to digital screen data. As our research shows, more studies directly examining real-life applications remain needed (12, 13). For the purposes of our study, the review by Lopes et al. highlighted that despite the rapid advancement of AI in the medical field, there is still a lack of clinical application in anesthesia practice (14).

In our study, we concluded that while AI can recommend anesthesia modalities based on shared patient information, further improvements are needed in compliance with existing guidelines and the management of specific patient groups.

This study has some limitations due to its single-center design, the fact that it was only included in the study within a period of 6 months, and the exclusion of patients who declined to participate in the study. The fact that patients who did not consent to the study were not included and that there were no pediatric and geriatric age groups prevented our findings from being reflected to a wider population. Furthermore, as a potential limitation of the study, some AI language programs have used free versions of certain programs.

Artificial intelligence cannot fully master the guidelines and exceptional and special situations that arise in the medical treatment process. In addition, the process of training the user with prompts and whether they are given free space or not may differ in the learning pattern and give different answers even in the same scenarios. This will affect the length and comprehensibility of the content, flow, responses, as in our study, and will produce different results if the user makes only artificial intelligence-oriented decisions.

In addition, the other issue is data privacy. Even if the large data sets used to train artificial intelligence information sharing models are "anonymized", it is possible to re-identify the data with advanced techniques. A few data points may be sufficient to retrospectively identify a patient. When unprotected data falls into the hands of parties, such as a third party, it can form the basis for discrimination against individuals. For example, a person at risk of a genetic disease may face an increase in health insurance premiums or exclusion in recruitment processes. The presence of such problems may create distrust in the user and the patient at the point of use.

In summary, artificial intelligence is not a tool to replace an experienced anesthesiologist. However, as this study proves, it has a high potential to enrich the clinical decision process as a consistent, fast, and valuable second opinion or supportive tool in perioperative evaluation, especially when the right model and the right use case are chosen.

REFERENCES

1.M. Graeßner *et al.*, “Enabling personalized perioperative risk prediction by using a machine-learning model based on preoperative data,” *Scientific Reports*, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-33981-8.

2.B. Choi *et al.*, “Prediction Model for 30-Day Mortality after Non-Cardiac Surgery Using Machine-Learning Techniques Based on Preoperative Evaluation of Electronic Medical Records,” *Journal of Clinical Medicine*, vol. 11, no. 21, p. 6487, Nov. 2022, doi: 10.3390/jcm11216487.

3.M. Vine *et al.*, “Innovative approaches to preoperative care including feasibility, efficacy, and ethical implications: a narrative review,” *AME Surgical Journal*, vol. 4. AME Publishing Company, p. 1, Feb. 01, 2024. doi: 10.21037/asj-23-41.

4.P. Chung, C. T. Fong, A. M. Walters, N. Aghaeepour, M. Yetişgen, and V. N. O’Reilly-Shah, “Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication,” *JAMA Surgery*, vol. 159, no. 8, American Medical Association, p. 928, Jun. 05, 2024. doi: 10.1001/jamasurg.2024.1621

5.T. Yurttas, R. Hidvegi, and M. Filipovic, “Biomarker-Based Preoperative Risk Stratification for Patients Undergoing Non-Cardiac Surgery,” *Journal of Clinical Medicine*, vol. 9, no. 2, p. 351, Jan. 2020, doi: 10.3390/jcm9020351

6.J. Stones and D. Yates, “Clinical risk assessment tools in anaesthesia,” *BJA Education*, vol. 19, no. 2. Elsevier BV, p. 47, Dec. 15, 2018. doi: 10.1016/j.bjae.2018.09.009.

7. Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4; 25:e50638. doi: 10.2196/50638. PMID: 37792434; PMCID: PMC10585440.

8. ATEŞMAN, Ender. (1997). Measuring Readability in Turkish. A.U. Tömer Language Journal, issue:58, pp.171-174.

9. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology. 2023 Oct;180:35-58. doi: 10.1016/j.urology.2023.05.040. Epub 2023 Jul 4. PMID: 37406864.

10. Canales C, Lee C, Cannesson M. Science Without Conscience Is but the Ruin of the Soul: The Ethics of Big Data and Artificial Intelligence in Perioperative Medicine. Anesth Analg. 2020 May; 130(5):1234-1243. doi: 10.1213/ANE.0000000000004728. PMID: 32287130; PMCID: PMC7519874.

11.Singh M, Nath G. Artificial intelligence and anesthesia: A narrative review. Saudi J Anaesth. 2022 Jan-Mar; 16(1):86-93. doi: 10.4103/sja.sja_669_21. Epub 2022 Jan 4. PMID: 35261595; PMCID: PMC8846233.

12. Bellini V, Rafano Carnà E, Russo M, Di Vincenzo F, Berghenti M, Baciarello M, Bignami E. Artificial intelligence and anesthesia: a narrative review. Ann Transl Med. 2022 May; 10(9):528. doi: 10.21037/atm-21-7031. PMID: 35928743; PMCID: PMC9347047.

13. Kambale M, Jadhav S. Applications of artificial intelligence in anesthesia: A systematic review. Saudi J Anaesth. 2024 Apr-Jun; 18(2):249-256. doi: 10.4103/sja.sja_955_23. Epub 2024 Mar 14. PMID: 38654854; PMCID: PMC11033896.

14. Lopes S, Rocha G, Guimarães-Pereira L. Artificial intelligence and its clinical application in Anesthesiology: a systematic review. J Clin Monit Comput. 2024 Apr; 38(2):247-259. doi: 10.1007/s10877-023-01088-0. Epub 2023 Oct 21. PMID: 37864754; PMCID: PMC10995017.