Original Paper

Assessing the Quality of AI Explanations of Atrial Fibrillation: Comparative Analysis of ChatGPT and Google Gemini

Jasneel Kahlam[1], DO; Alexander Sacher[2], DO; Kyrollos Saad[3], DO; David Zhang[4], MD; Rohit Sekandalapuram[5], MD; Quang Dai La[6], MS; David F Lo[6], BS; John P Reilly[4], MD

[1]Stony Brook Southampton Hospital, Southampton, NY, United States

[2]Department of Internal Medicine, Department of Internal Medicine, New York Hospital Queens, Flushing, NY, United States

[3]Department of Cardiology, Deborah Heart and Lung Center, Browns Mills, NJ, United States

[4]Stony Brook University Hospital, Stony Brook, NY, United States

[5]Brooklyn Hospital Center, Brooklyn, NY, United States

[6]American Preventive Screening & Education Association, Stratford, NJ, United States

Corresponding Author:
Jasneel Kahlam, DO
Stony Brook Southampton Hospital
240 Meeting House Lane
Southampton, NY, 11968
United States
Phone: 1 (631) 726 8200
Email: Jasneelkahlam@gmail.com

Abstract

Background: Atrial fibrillation (AF) is one of the most common arrhythmias and is associated with multiple risk factors, including stroke and heart failure. Therefore, it is important that accurate and comprehensive educational resources on AF are available. Artificial intelligence (AI) platforms such as Google Gemini and ChatGPT are emerging tools for medical education. To our knowledge, this is the first study to assess the quality of these platforms for AF education.

Objective: This pilot study aimed to (1) assess the quality of ChatGPT and Google Gemini explanations of AF and its treatment, (2) compare responses from both platforms, and (3) analyze differences in interpretation between cardiologists and nonmedical professionals.

Methods: The prompt “Explain atrial fibrillation and how to treat it to a patient” was entered into both ChatGPT and Google Gemini. The responses, along with a survey created in Google Forms, were sent out to potential participants between September and October 2024. The survey questions were derived from the Patient Education Materials Assessment Tool for Printable Materials and DISCERN criteria. Each survey question was multiple choice and rated on a 5-point Likert scale, where 1 indicated strong disagreement and 5 indicated strong agreement. Averages of each answer from the survey were calculated and compared using the Wilcoxon signed-rank test.

Results: The survey was completed by 28 participants, including cardiologists and cardiology fellows and 17 nonmedical professionals (college graduates not employed in the medical field). No significant overall quality differences were observed between ChatGPT and Google Gemini. When comparing participant groups, cardiologists rated AI-generated responses higher than nonmedical professionals for perceived bias (mean 3.82 vs mean 4.33; P=.04) and for explaining consequences of no treatment (mean 2.85 vs mean 1.86; P=.005). In contrast, nonmedical professionals rated structural and organizational features more favorably than cardiologists, including visual cues (mean 3.97 vs mean 2.48; P<.001), informative headers (mean 4.77 vs mean 3.95; P=.001), concise sections (mean 4.71 vs mean 4.29; P=.04), identifiable patient actions (mean 4.21 vs mean 3.19; P=.04), and references to additional resources (mean 2.65 vs mean 1.71; P=.006).

Conclusions: The comparable quality of ChatGPT and Google Gemini suggests that both are viable for AF education. Differences in ratings of organizational features versus perceptions of bias and consequences of no treatment highlight a disconnect between cardiologists and nonmedical professionals. To optimize clinical application, cardiologists should evaluate both comprehensiveness and clarity of AI-generated responses. This serves as a pilot study, and future research should focus on longitudinal studies of AI-generated responses and their effects on clinical outcomes.

XSL•FO
RenderX

(JMIR Form Res 2026;10:e73715) doi: 10.2196/73715

KEYWORDS

atrial fibrillation; medical education; ChatGPT; Google Gemini; artificial intelligence

Introduction

Atrial fibrillation (AF) is a common arrhythmia with increasing prevalence, particularly among older adults, individuals who consume alcohol, and smokers [1]. AF significantly elevates the risk of stroke, ischemic events, and hospitalizations. However, lifestyle modifications, such as regular exercise and abstinence from alcohol and smoking, can help reduce AF risk [1]. When the diagnosis of AF is confirmed, medications and interventions can control heart rate, rhythm, and clot formation. Given these risks and treatment complexities, effective patient education for AF prevention and management is essential.

Patient education plays a crucial role in improving adherence to treatment plans and lifestyle modifications. Low health literacy has been associated with increased health care use and noncompliance with medical recommendations [2]. Educational interventions directed toward patients with chronic heart failure have been shown to enhance quality of life and medication adherence [3-5]. Similarly, digital education tools have demonstrated benefits in other medical contexts; for example, postoperative knee pain management following total knee replacement was improved with a mobile app delivering consistent daily education [6]. Despite adherence improvement, patient education must be assessed for veracity and utility. For instance, 2 studies on the quality of YouTube videos for coronary artery disease and colorectal cancer received low ratings in terms of accuracy and usefulness [7,8].

With advancements in artificial intelligence (AI), platforms such as ChatGPT have become increasingly used for patient education. Studies have shown that AI platforms can improve patient knowledge regarding prostate cancer, diabetes, and hypertension by providing accessible and accurate medical information [9-11]. However, despite these promising results, there are still gaps in the clarity, reliability, and perception of AI-generated health information that need to be addressed. When assessing information on chronic kidney disease, AI-generated responses were found to have some inaccuracies when compared with clinical guidelines [12]. This gap creates challenges for patients and physicians who have different expectations regarding the disease in question and therapeutic goals.

While several studies have explored AI's role in predicting AF based on patient risk factors, fewer studies have examined the quality of AI-generated patient education for AF. One study assessed AI-generated responses to general AF-related questions, but it did not specifically evaluate patient-focused explanations or include nonmedical professional reviewers [13-16]. Another study demonstrated how AI was able to adapt responses for physicians or patients, but it did not compare responses across multiple AI platforms or analyze perceptual differences between medical professionals and laypeople [17]. In this study, we aimed to address these gaps by assessing the clarity, comprehensiveness, and accuracy of AI-generated patient education on AF. Specifically, we compared responses from ChatGPT and Google Gemini, analyzing their abilities to provide clear and accurate information to patients with AF. Additionally, 2 comparison groups, consisting of cardiologists and cardiology fellows and nonmedical professionals, were asked to review the responses. To our knowledge, this is the first study to directly compare AI-generated AF education across multiple AI platforms and distinct user groups (cardiologists vs nonmedical professionals). This study will provide valuable information about strengths and weaknesses of each AI platform, as well as the interpretation of AI-generated responses in a diverse patient population.

Methods

Study Design

A comparative analysis was conducted on September 6, 2024, using ChatGPT (GPT-4; OpenAI) and Google Gemini 2.5 Flash. Both AI platforms generated conversational responses to user queries. The prompt, "Explain atrial fibrillation and how to treat it to a patient," was entered into each AI platform. The AI-generated responses were emailed to both participant groups.

Each participant received an email with the 2 prompts, instructions on how to answer the prompts, and the survey. Both prompts were included in the same email, with ChatGPT labeled as prompt #1 and Google Gemini listed as prompt #2 (Textbox 1). Participants completed the survey twice: once for the ChatGPT response and once for the Google Gemini response. Each response to the survey was anonymous. Although ChatGPT is available as an extension on Doximity, we used the standard ChatGPT interface, as it is the most likely source for patient education. For further explanation of the study design, please refer to Figure 1. This study was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational studies.

XSL•FO
RenderX

Textbox 1. Full-text responses generated by ChatGPT and Google Gemini to the prompt, "Explain atrial fibrillation and how to treat it to a patient," on September 6, 2024. These standardized responses were sent to each participant along with a survey derived from DISCERN and the Patient Education Materials Assessment Tool for Printable Materials criteria.

ChatGPT

What is Atrial Fibrillation?

Atrial fibrillation is a type of irregular heartbeat or arrhythmia. Here's a simple way to think about it:

1. Heart's Rhythm: Your heart has four chambers: two upper ones called atria and two lower ones called ventricles. Normally, the heart beats in a steady rhythm, with the atria contracting and pushing blood into the ventricles, which then pump the blood out to the rest of your body.

2. What Happens in AFib: In atrial fibrillation, the electrical signals that control the heartbeat become chaotic. Instead of a steady, regular rhythm, the atria (the upper chambers) quiver or "fibrillate" rapidly and irregularly. This means the atria aren't contracting properly, which can cause the heart to beat irregularly or too fast.

3. Symptoms: Some people with AFib feel palpitations (a fluttering or racing heart), dizziness, shortness of breath, or fatigue. However, some people may not have any symptoms at all.

4. Risks: One of the main concerns with AFib is that it increases the risk of blood clots forming in the atria. These clots can potentially travel to the brain and cause a stroke.

How to Treat Atrial Fibrillation

The treatment for AFib depends on various factors, including the severity of symptoms, your overall health, and any underlying conditions. The main goals of treatment are to control your heart rate, restore a normal rhythm if possible, and reduce the risk of blood clots. Here are common approaches:

1. Medications:

- Rate Control: Medications like beta-blockers or calcium channel blockers can help slow down your heart rate and make it more regular.
- Rhythm Control: Some medications can help restore a normal heart rhythm.
- Blood Thinners: To reduce the risk of blood clots and stroke, you might be prescribed anticoagulants (blood thinners) like warfarin or newer options such as direct oral anticoagulants.

2. Lifestyle Changes:

- Diet and Exercise: Maintaining a healthy diet and regular exercise can improve heart health.
- Avoiding Triggers: Sometimes, factors like excessive caffeine, alcohol, or stress can trigger AFib. Identifying and managing these triggers can help.

3. Procedures:

- Electrical Cardioversion: This is a procedure where a controlled electric shock is used to try to restore a normal rhythm.
- Catheter Ablation: This procedure involves threading a catheter through your blood vessels to destroy small areas of heart tissue that are causing the irregular signals.

4. Surgery:

- In some cases, more invasive procedures like maze surgery might be considered if other treatments aren't effective.

It's important to work closely with your healthcare provider to determine the best treatment plan for you. They can help you understand the options and tailor the treatment to your specific needs. If you have any questions or concerns, don't hesitate to ask them for more information.

Google Gemini

Atrial fibrillation, often shortened to AFib, is a heart rhythm disorder where the upper chambers of your heart (the atria) beat irregularly and rapidly. This can cause blood to pool in the atria, increasing the risk of blood clots forming. These clots can travel to other parts of your body, potentially causing a stroke.

Symptoms of Atrial Fibrillation

- Rapid heartbeat:You may feel your heart racing or fluttering.
- Fatigue:Feeling tired or weak.
- Shortness of breath:Difficulty breathing, especially when active.
- Chest pain or discomfort.
- Dizziness or lightheadedness.
- Palpitations:A feeling of your heart skipping a beat or beating irregularly.

Treatment for Atrial Fibrillation

XSL•FO
RenderX

The treatment for AFib will depend on your specific situation, including the severity of your symptoms and any underlying health conditions. Here are some common treatment options:

1. Medication:

- Blood thinners: To prevent blood clots and reduce the risk of stroke.
- Rhythm control medications: To restore a normal heart rhythm.
- Beta-blockers or calcium channel blockers: To slow your heart rate and reduce blood pressure.

2. Cardioversion:

- Electrical cardioversion: Using an electric shock to restore a normal heart rhythm.
- Chemical cardioversion: Using medication to restore a normal heart rhythm.

3. Catheter ablation:

- A procedure where a thin tube with a heated tip is inserted into your heart to destroy the abnormal heart tissue causing AFib.

4. Pacemaker or implantable defibrillator:

- In some cases, a pacemaker or implantable defibrillator may be needed to help regulate your heart rhythm.

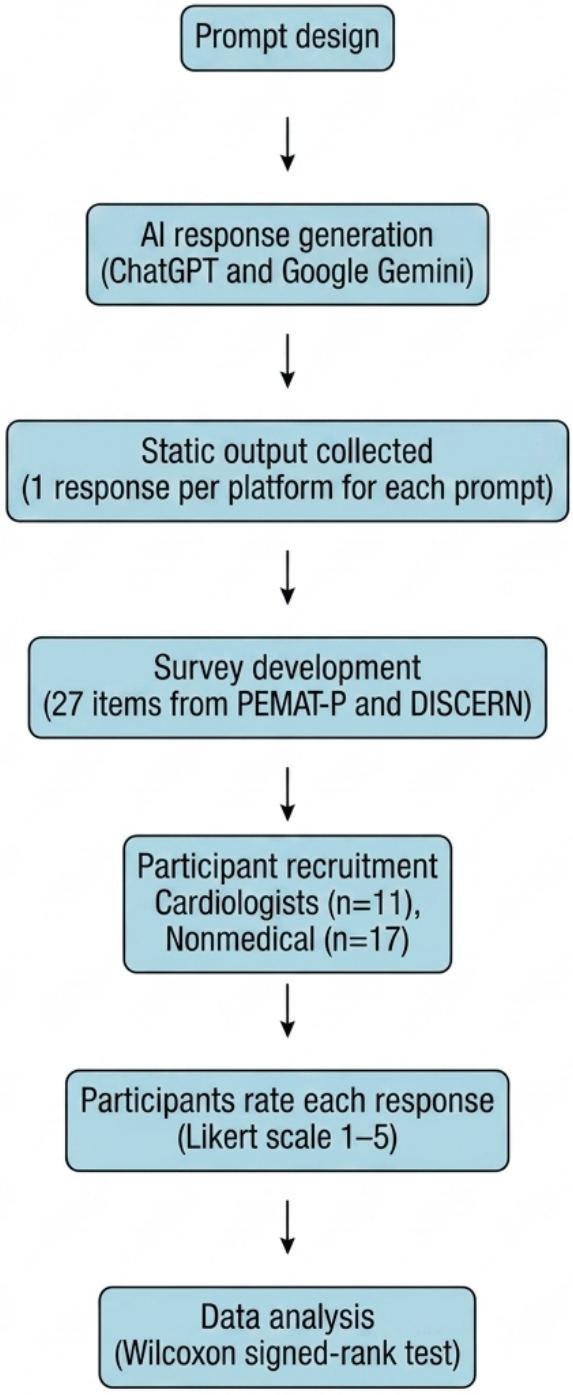
5. Lifestyle changes:

- Healthy diet: Eating a balanced diet can help manage your overall health.
- Regular exercise: Moderate exercise can help improve your heart health.
- Weight management: Maintaining a healthy weight can reduce the risk of AFib.
- Stress management: Techniques like meditation or yoga can help reduce stress.
- Limit alcohol and caffeine: Excessive consumption of these substances can trigger AFib episodes.

If you have been diagnosed with AFib, it's important to work closely with your doctor to develop a treatment plan that's right for you. By following your doctor's recommendations and making healthy lifestyle choices, you can manage your condition and improve your quality of life.

XSL•FO
RenderX

Figure 1. A flowchart explaining the study design and participant flow. Artificial intelligence (AI)–generated responses were produced by entering predefined prompts into ChatGPT and Google Gemini. The initial response from each platform was collected as a static source of patient education. A survey instrument was developed using 27 items adapted from the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) and DISCERN criteria. Participants, including 11 cardiologists and 17 nonmedical professionals, rated each AI-generated response using a 5-point Likert scale. Data were analyzed using the Wilcoxon signed-rank test to compare ratings across groups and platforms.



Assessment Tools

A survey was developed using Google Forms to evaluate responses. Clarity and readability were assessed using the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), while comprehensiveness was measured using the DISCERN criteria [18,19]. The initial survey contained 35 questions, but 6 (17.1%) duplicates between the 2 criteria were removed, resulting in a final set of 29 (82.9%) Likert scale questions.

XSL•FO
RenderX

When applying the PEMAT-P and DISCERN criteria to assess the quality of AI-generated responses, the initial response to the prompt in each AI platform was evaluated. Rather than engaging in a conversation with the AI platform, we took the initial response as a stand-alone source of patient education that a patient might receive from their cardiologist. This allowed us to assess both the content and the structure of each initial response using validated criteria.

Participant Selection

Two groups completed the survey: (1) a cardiologist group consisting of practicing cardiologists and cardiology fellows and (2) a nonmedical professional group, defined as individuals with a 4-year college degree who were not employed in the medical field. Participants from both groups were recruited via standardized emails, including reminder messages, and completed the survey between September and October 2024. The cardiologists and cardiology fellows were recruited through departmental mailing lists and professional contacts in the cardiology communities of both New Jersey and New York. The nonmedical professionals were recruited through community contacts and professional outreach in the United States and the United Kingdom. In total, 11 cardiologists across the United States and 17 nonmedical professionals from the United States and the United Kingdom filled out the survey.

Statistical Analysis

Means of survey responses were calculated for each subgroup (cardiologists, nonmedical professionals, ChatGPT, and Google Gemini). Comparisons were conducted using the Wilcoxon signed-rank test, a nonparametric test appropriate for ordinal data. A P value of <.05 was considered statistically significant.

Given the modest sample size of this study, no multivariable adjustments for potential confounders were performed. To minimize bias, all participants reviewed the same standardized AI-generated responses, and responses were analyzed in aggregate. All identifiable demographic information was removed from the data. Effect modifiers, such as reviewer expertise (medical vs nonmedical), were examined by stratified analysis of survey ratings across groups. Given that there was only 1 cardiologist who missed a response, there were no data adjustments to accommodate missing data.

Study Inclusion and Exclusion Criteria

The inclusion and exclusion criteria are presented in Textbox 2.

Textbox 2. The inclusion and exclusion criteria for each participant cohort. These criteria were used to create a comparative analysis between the cardiologist and cardiology fellow and the medical layperson groups to assess gaps in understanding of atrial fibrillation management.

Inclusion criteria

- Inpatient and outpatient cardiologists
- Individuals with a 4-year college degree in a nonmedical field
- Cardiology fellows
- English-speaking individuals

Exclusion criteria

- Participants who did not complete or attend college
- Medical students
- Graduates who obtained degrees in medical-related majors
- Physicians who did not specialize in cardiology

Ethical Considerations

This study did not meet the definition of human research as defined by Stony Brook Human Research Protection Program Standard Operating Procedure, section 1.5, as no patient information was collected. In accordance with section 3.8.5, no private information was collected, as all data were derived from publicly available AI platforms. Per section 3.4, the investigator determined that this study did not involve human subjects and did not require institutional review board approval. On the basis of the abovementioned criteria, this study falls outside the scope of research requiring institutional review board approval. This study has been reported in accordance with STROBE guidelines. Both study cohort participants were voluntary, anonymous, and were not compensated.

Results

Comparison Between Google Gemini and ChatGPT Responses

A total of 55 survey responses were collected. There were 29 (52.7%) responses to the ChatGPT-generated prompt and 26 (47.3%) responses to the Google Gemini–generated prompt (Table 1). As shown in Table 2, there were no significant differences between responses to ChatGPT and Google Gemini across all questions. From the DISCERN criteria, both platforms received low ratings for questions addressing risks of treatment, changes in quality of life, and what happens if no treatment was used (Table 2). From the PEMAT-P questions, questions assessing the use of visual aids, such as charts and graphs, had low ratings from both platforms (Table 3).

XSL•FO
RenderX

Table 1. Responses to artificial intelligence–generated prompts (n=55).

AI platform	Responses, n (%)
Google Gemini	26 (47.3)
ChatGPT	29 (52.7)

XSL•FO
RenderX

Table 2. Comparison of mean Likert scale ratings (1=strongly disagree and 5=strongly agree) for artificial intelligence–generated responses from ChatGPT and Google Gemini across 27 evaluation questions adapted from the Patient Education Materials Assessment Tool for Printable Materials and DISCERN criteria frameworks. None of the comparisons reached statistical significance.

Question	ChatGPT (n=29), mean	Google Gemini (n=26), mean	*P* value[a]
Are the aims of the prompt clear? Does it talk about the goals of the prompt?	5.000	4.500	.29
Does this prompt achieve its aims? Does it adequately describe what atrial fibrillation is and how to treat it?	4.000	4.192	.36
Does this prompt address all the questions a reader might ask?	4.000	3.615	.37
Does this prompt mention the sources of information that were used to compile the publication (other than the author or producer)?	2.000	1.500	.74
Does the prompt mention the source of the information or when that source was published?	1.000	1.500	.99
Is there any bias in this prompt?	4.000	4.213	.12
Does this prompt suggest that you read more or extract further details from other sources?	2.000	1.577	.67
Does this prompt refer to areas of uncertainty? (areas where there are disagreements among experts)	2.000	1.577	.51
Does this prompt describe how each treatment works? (description of one/some but not all is a 3)	4.000	3.654	.83
Does this prompt describe the risk of the treatments?	1.000	1.462	.77
Does this prompt describe what would happen if no treatment is used?	3.000	2.423	.83
Does this prompt describe how treatment choices will affect quality of life?	2.000	2.000	.96
Does this prompt support shared decision-making?	3.000	2.423	.49
Based on the questions above, rate the following prompt based on the comprehensiveness of the information provided.	3.000	3.3456	.48
Are medical terms used only to familiarize the audience with these terms? Does prompt define medical terms?	4.000	3.962	.42
Does the prompt feature any statements that distract from its purpose?	2.000	1.692	.67
Does the prompt use common everyday language?	4.000	4.346	.29
Is the prompt broken down into shorter sections?	4.000	4.615	.68
Does the prompt provide informative headers?	4.000	4.500	.78
Is the prompt summarized?	3.000	3.308	.72
Does this prompt use visual cues (bolding, arrows, boxes, larger font, highlighting)?	3.000	3.500	.41
Does the prompt clearly identify actions that users can take (1 action is a 3)	4.000	3.692	.41
Is the user directly addressed when describing actions that the reader can take?	3.000	3.231	.91
Is the prompt broken down into manageable and explicit steps for treatment?	3.000	3.077	.68
Does the prompt provide easy access to links or references for further reading?	2.000	2.231	.63
Does the prompt explain how to utilize any charts, graphs, tables, or diagrams to describe how to take action?	1.000	1.385	.46
Does this prompt use visual aids whenever possible to describe treatment modalities?	1.000	1.500	.95

[a]Wilcoxon signed-rank test was used to compare mean ratings.

XSL•FO
RenderX

Table 3. Comparison of mean Likert scale ratings (1=strongly disagree and 5=strongly agree) between cardiologists (n=21) and nonmedical professionals (n=34) across the same 27-item evaluation survey mentioned in Table 2.

Question	Cardiologists (n=11), mean	Nonmedical professionals (n=17), mean	*P* value[a]
Are the aims of the prompt clear? Does it talk about the goals of the prompt?	4.381	4.735	.13
Does this prompt achieve its aims? (does it adequately describe what atrial fibrillation is and how to treat it)	4.095	4.353	.10
Does this prompt address all the questions a reader might ask?	3.619	3.794	.55
Does this prompt mention the sources of information used to compile the publication (other than the author or producer)?	1.476	1.559	.92
Does the prompt mention the source of the information or when that source was published?	1.286	1.559	.38
Is there any bias in this prompt?	4.333	3.824	*.04* [b]
Does this prompt lead you to read more or extract further details from other sources?	1.524	1.765	.34
Does this prompt refer to areas of uncertainty? (areas where there are disagreements among experts)	1.619	1.529	.67
Does this prompt describe how each treatment works? (description of one/some but not all is a 3)	3.190	3.824	.06
Does this prompt describe the risk of the treatments?	1.571	1.412	.63
Does this prompt describe what would happen if no treatment is used?	1.857	2.853	*.005*
Does this prompt describe how treatment choices will affect quality of life?	2.000	1.941	.90
Does this prompt support shared decision-making?	2.143	2.824	.08
Based on the questions above, rate the following prompt based on the comprehensiveness of the information provided.	3.286	3.206	.87
Are medical terms used only to familiarize the audience with these terms? Does prompt define medical terms?	3.857	3.794	.74
Does the prompt feature any statements that distract from its purpose?	1.714	1.735	.83
Does the prompt use common everyday language?	4.429	4.176	.21
Is the prompt broken down into shorter sections?	4.286	4.706	*.04*
Does the prompt provide informative headers?	3.952	4.765	*.001*
Is the prompt summarized?	3.238	3.5	.33
Does this prompt use visual cues (bolding, arrows, boxes, larger font, highlighting)?	2.476	3.971	*<.001*
Does the prompt clearly identify actions that users can take? (1 action is a 3)	3.190	4.206	*.001*
Is the user directly addressed when describing actions that the reader can take? (you should do something)	2.714	3.471	*.04*
Is the prompt broken down into manageable and explicit steps for treatment?	3.238	2.824	.38
Does the prompt provide easy access to links or references for further reading? (1-2 tools is a 3)	1.714	2.647	*.006*
Does this prompt explain how to utilize any charts, graphs, tables, or diagrams to describe how to take actions?	1.524	1.353	.30
The prompt uses visual aids whenever possible to describe treatment modalities.	1.476	1.471	.68

[a]Wilcoxon signed-rank test was used to compare the responses between each of the groups.

[b]Italicized values denote statistical significance (*P*<.05).

XSL•FO
RenderX

Comparison Between Cardiologists and Nonmedical Professionals

Of the 55 survey responses collected, 34 (61.8%) were from noncardiologists and 21 (38.2%) were from the cardiologist group (Table 4). Table 3 shows that the cardiologist group had a significant difference in ratings of bias when compared with the nonmedical group (*P*=.03). Additionally, the nonmedical group provided a significantly higher rating when describing what would happen if there was no treatment for AF (*P*=.005; Table 3). The cardiologist group had significantly lower ratings for shorter sections of the prompt (*P*=.04), informative headers (*P*=.001), and the use of visual cues (*P*<.001; Table 3). Additionally, cardiologists rated the responses significantly lower in key areas, including the use of shorter sections (*P*=.04), informative headers (*P*=.001), and visual cues (*P*<.001).

The nonmedical group showed lower ratings when compared with the cardiology group for identifying actions (*P*=.001) and describing how those actions could be taken (*P*=.04), as well as accessibility to links for further reading (*P*=.006). Nonmedical professionals also provided significantly higher ratings for explanations of what would happen if AF remained untreated (*P*=.005). Compared with cardiologists, the nonmedical group gave significantly lower ratings for identifying actionable steps (*P*=.001), explaining how to take those actions (*P*=.04), and accessibility to links for further reading (*P*=.006).

Table 4. Distribution of responses from each cohort (n=55). The practicing cardiologist group is included in the cardiologist group.

Profession	Responses, n (%)
Cardiologist	34 (61.8)
Nonmedical professionals	21 (38.2)

Discussion

Main Findings

AI platforms are becoming increasingly accessible and are frequently used to obtain medical information and assist with shared medical decision-making. In this study, there was no significant difference in the comprehensiveness and clarity between the AI platforms, suggesting that both platforms provided a similar quality of AF information.

When comparing responses from cardiologists and nonmedical professionals, several key differences emerged. First, the cardiology group perceived significantly more bias in the AI-generated responses and significantly fewer explanations for the consequences of forgoing treatment. The nonmedical group rated the AI responses higher for identification and implementation of patient actions. Finally, the nonmedical group gave higher ratings for using shorter sections, informative headers, and visual cues.

Comparison With Other Studies

Overall, the cardiologist cohort assessed AI-generated AF information as accurate, yet incomplete and unclear for implementation. These findings align with prior research in other medical specialties. For example, studies regarding cirrhosis, hepatocellular cancer, and hand surgeries have shown similar findings where ChatGPT-generated prompts showed high accuracy and lower comprehensiveness [20,21].

Additionally, there was no difference in the quality of information provided between ChatGPT and Google Gemini. One study regarding adolescent scoliosis care showed similar applicability, comprehensiveness, and communication of AI-generated responses when reviewed by orthopedic surgeons, which aligns with the findings from the cardiologist group of this study [22]. However, there have been some papers in other specialties, particularly 1 study assessing maternal lactation information that showed a difference between multiple AI platforms, including ChatGPT and Google Gemini [23]. Further studies comparing these multiple AI platforms should be performed to determine which platforms perform best for a given medical topic.

Clinical Implications

The use of AI-generated responses could enhance patient education regarding AF. However, it is essential that these responses are comprehensive without compromising their clarity. For instance, AF education should be accessible to the average patient and include all relevant information on rate and rhythm control, anticoagulation strategies, and benefits and risks of catheter ablation. Additionally, cardiologists should review these responses for accuracy and completeness, recognizing that poor quality could adversely affect patient decision-making. Ideally, cardiologists should use high-quality AI-generated responses to supplement in-office discussions with patients.

Future Directions

One potential future direction is to evaluate whether providing AI-generated responses to patients in a cardiology clinic can improve AF management outcomes. A previous study showed that AI-assisted decision-making significantly improved decision quality and patient outcomes for osteoarthritis in an orthopedic clinic [24]. Similar research could help refine medical information provided by ChatGPT and Google Gemini. The addition of other, more medically focused AI platforms, such as OpenEvidence, could be used to test the accuracy and comprehensiveness of future responses from ChatGPT and Google Gemini.

As a continuation of this study, it could be beneficial to assess AI's adaptability by asking it to refine the responses to address potential biases in health literacy and to provide a clearer structure. The improved prompt can be compared with the original prompt from each platform to determine whether AI was able to make the necessary improvements.

Limitations

This study has several limitations that should be addressed in future research. First, the sample size in each

XSL•FO
RenderX

group—particularly among cardiologists—was modest, which may have limited the ability to detect smaller but meaningful differences [25]. Larger and more diverse cohorts could improve both the reliability and generalizability of future findings.

Additionally, each AI-generated response was labeled by platform (ChatGPT vs Google Gemini), which may have introduced preferential bias based on participants' prior familiarity or perceptions. Presenting these responses anonymously could help mitigate this bias.

Second, the PEMAT-P and DISCERN criteria were originally developed to assess static, printed patient education materials. Therefore, their application to interactive AI-generated content may have been limited. In this study, we evaluated only the initial response from each AI platform, treating it as printed educational content a patient might find in their physician's office. However, future research should consider developing or adapting evaluation criteria that reflect the dynamic and conversational nature of AI interactions.

Third, the study evaluated only 2 general-purpose AI platforms. Including additional tools, such as Microsoft Copilot or more clinically oriented systems such as OpenEvidence, could offer a more comprehensive assessment of AI-generated medical information. Our decision to limit the study to ChatGPT and Google Gemini was based on the need to reduce participant burden and optimize survey completion.

Fourth, we examined responses from 2 groups: cardiologists and nonmedical professionals. Including a broader range of medical professionals—such as internal medicine residents, general practitioners (in both inpatient and outpatient settings), cardiac nurse practitioners, and physician assistants—could reveal more nuanced insights into how AI-generated explanations are interpreted across clinical roles. Likewise, expanding the layperson cohort to include individuals without a college education may provide greater insight into gaps in health literacy and digital comprehension. Additionally, including participants who speak English as a second language could improve comprehensibility and accessibility of AI-generated responses across diverse populations.

Finally, this study focused only on the initial output provided by each AI platform. Future studies could explore the adaptability of conversational AI by prompting platforms to revise and improve their explanations in response to feedback. Comparing these revised responses to the originals could help evaluate whether AI tools are capable of meaningfully refining their content to improve clarity, reduce bias, and better meet patients' informational needs.

Conclusions

AI platforms, such as Google Gemini and ChatGPT, are increasingly used as resources for medical information due to their ability to provide personalized responses. While both platforms generate information of similar quality, their responses must be optimized to enhance patient understanding and improve the presentation of medical information. This pilot study shows the potential gaps between cardiologists' and patients' perceptions of AF education. Further collaboration between cardiologists and AI software companies could help deliver peer-reviewed, widely accepted medical standards to patients, thereby supporting cardiologists in shared decision-making.

Acknowledgments

The authors are grateful to Dr Demetrios Papakostas for his guidance and contribution to statistical analysis. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading, editing, summarizing text, and adapting and adjusting emotional tone. ChatGPT (GPT-4; OpenAI) was used. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Funding

No external financial compensation was received from any public, commercial, or not-for-profit entities for the completion of this study.

Conflicts of Interest

None declared.

References

1. Joglar JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. Jan 02, 2024;149(1):e1-156. [FREE Full text] [doi: 10.1161/CIR.0000000000001193] [Medline: 38033089]

XSL•FO
RenderX

2. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. Jul 19, 2011;155(2):97-107. [FREE Full text] [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
3. Jiang Y, Shorey S, Seah B, Chan WX, Tam WW, Wang W. The effectiveness of psychological interventions on self-care, psychological and health outcomes in patients with chronic heart failure-a systematic review and meta-analysis. Int J Nurs Stud. Feb 2018;78:16-25. [FREE Full text] [doi: 10.1016/j.ijnurstu.2017.08.006] [Medline: 28935393]
4. Habibzadeh H, Shariati A, Mohammadi F, Babayi S. The effect of educational intervention based on Pender's health promotion model on quality of life and health promotion in patients with heart failure: an experimental study. BMC Cardiovasc Disord. Oct 05, 2021;21(1):478. [FREE Full text] [doi: 10.1186/s12872-021-02294-x] [Medline: 34610821]
5. Świątoniowska-Lonc NA, Sławuta A, Dudek K, Jankowska K, Jankowska-Polańska BK. The impact of health education on treatment outcomes in heart failure patients. Adv Clin Exp Med. Apr 2020;29(4):481-492. [FREE Full text] [doi: 10.17219/acem/115079] [Medline: 32348037]
6. Timmers T, Janssen L, van der Weegen W, Das D, Marijnissen WJ, Hannink G, et al. The effect of an app for day-to-day postoperative care education on patients with total knee replacement: randomized controlled trial. JMIR Mhealth Uhealth. Oct 21, 2019;7(10):e15323. [FREE Full text] [doi: 10.2196/15323] [Medline: 31638594]
7. Kahlam JS, Sacher A, Singh C, Lo DF, Reilly JP. Assessment of YouTube videos in patient education for coronary artery disease: a DISCERN-based cross-sectional analysis. Cureus. Jun 23, 2024;16(6):e62986. [FREE Full text] [doi: 10.7759/cureus.62986] [Medline: 39044879]
8. Kahlam A, Pai S, Kahlam J, Ahlawat S. Quality of YouTube videos related to colorectal cancer screening. Cureus. Jan 12, 2023;15(1):e33684. [FREE Full text] [doi: 10.7759/cureus.33684] [Medline: 36788914]
9. Görtz M, Baumgärtner K, Schmid T, Muschko M, Woessner P, Gerlach A, et al. An artificial intelligence-based chatbot for prostate cancer education: design and patient evaluation study. Digit Health. May 02, 2023;9:20552076231173304. [FREE Full text] [doi: 10.1177/20552076231173304] [Medline: 37152238]
10. Nakhleh A, Spitzer S, Shehadeh N. ChatGPT's response to the Diabetes Knowledge Questionnaire: implications for diabetes education. Diabetes Technol Ther. Aug 2023;25(8):571-573. [FREE Full text] [doi: 10.1089/dia.2023.0134] [Medline: 37062754]
11. Almagazzachi A, Mustafa A, Eighaei Sedeh A, Vazquez Gonzalez AE, Polianovskaia A, Abood M, et al. Generative artificial intelligence in patient education: ChatGPT takes on hypertension questions. Cureus. Feb 02, 2024;16(2):e53441. [FREE Full text] [doi: 10.7759/cureus.53441] [Medline: 38435177]
12. Acharya PC, Alba R, Krisanapan P, Acharya CM, Suppadungsuk S, Csongradi E, et al. AI-driven patient education in chronic kidney disease: evaluating chatbot responses against clinical guidelines. Diseases. Aug 16, 2024;12(8):185. [FREE Full text] [doi: 10.3390/diseases12080185] [Medline: 39195184]
13. Noseworthy PA, Attia ZI, Behnken EM, Giblon RE, Bews KA, Liu S, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. Lancet. Oct 08, 2022;400(10359):1206-1212. [FREE Full text] [doi: 10.1016/S0140-6736(22)01637-3] [Medline: 36179758]
14. Sánchez de la Nava AM, Atienza F, Bermejo J, Fernández-Avilés F. Artificial intelligence for a personalized diagnosis and treatment of atrial fibrillation. Am J Physiol Heart Circ Physiol. Apr 01, 2021;320(4):H1337-H1347. [FREE Full text] [doi: 10.1152/ajpheart.00764.2020] [Medline: 33513086]
15. Lin F, Zhang P, Chen Y, Liu Y, Li D, Tan L, et al. Artificial-intelligence-based risk prediction and mechanism discovery for atrial fibrillation using heart beat-to-beat intervals. Med. May 10, 2024;5(5):414-31.e5. [FREE Full text] [doi: 10.1016/j.medj.2024.02.006] [Medline: 38492571]
16. Hillmann HA, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. Europace. Dec 28, 2023;26(1):euad369. [FREE Full text] [doi: 10.1093/europace/euad369] [Medline: 38127304]
17. Lee TJ, Campbell DJ, Rao AK, Hossain A, Elkattawy O, Radfar N, et al. Evaluating ChatGPT responses on atrial fibrillation for patient education. Cureus. Jun 4, 2024;16(6):e61680. [FREE Full text] [doi: 10.7759/cureus.61680] [Medline: 38841294]
18. Shoemaker SJ, Wolf MS, Brach C. The Patient Education Materials Assessment Tool (PEMAT) and user's guide. Agency for Healthcare Research and Quality. 2013. URL: https://www.ahrq.gov/sites/default/files/publications/files/pemat_guide.pdf [accessed 2026-04-03]
19. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. Feb 1999;53(2):105-111. [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
20. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. Jul 2023;29(3):721-732. [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
21. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of online artificial intelligence-generated information on common hand procedures. J Hand Surg Am. Nov 2023;48(11):1122-1127. [doi: 10.1016/j.jhsa.2023.08.003] [Medline: 37690015]

XSL•FO
RenderX

22. Yaş S, Yapar D, Yapar A, Özel T, Tokgöz MA, Baymurat AC, et al. Assessing the role of large language models in adolescent idiopathic scoliosis care: a comparison between ChatGPT and Google Gemini. Acta Orthop Traumatol Turc. Jul 18, 2025;59(4):222-229. [FREE Full text] [doi: 10.5152/j.aott.2025.25279] [Medline: 40728044]
23. Özer Aslan İ, Aslan MT. Benchmarking AI chatbots for maternal lactation support: a cross-platform evaluation of quality, readability, and clinical accuracy. Healthcare (Basel). Jul 20, 2025;13(14):1756. [FREE Full text] [doi: 10.3390/healthcare13141756] [Medline: 40724781]
24. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open. Feb 01, 2021;4(2):e2037107. [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
25. Noether GE. Sample size determination for some common nonparametric tests. J Am Stat Assoc. 1987;82(398):645-647. [doi: 10.1080/01621459.1987.10478478]

Abbreviations

AF: atrial fibrillation
AI: artificial intelligence
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 10.Mar.2025; peer-reviewed by A Adekola, X Ding; comments to author 21.May.2025; revised version received 05.Jan.2026; accepted 06.Jan.2026; published 05.Nov.2026

Please cite as:
Kahlam J, Sacher A, Saad K, Zhang D, Sekandalapuram R, La QD, Lo DF, Reilly JP
Assessing the Quality of AI Explanations of Atrial Fibrillation: Comparative Analysis of ChatGPT and Google Gemini
JMIR Form Res 2026;10:e73715
URL: https://formative.jmir.org/2026/1/e73715
doi: 10.2196/73715
PMID:

©Jasneel Kahlam, Alexander Sacher, Kyrollos Saad, David Zhang, Rohit Sekandalapuram, Quang Dai La, David F Lo, John P Reilly. Originally published in JMIR Formative Research (https://formative.jmir.org), 05.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX