DOI: https://dx.doi.org/10.18203/2320-6012.ijrms20244094

Original Research Article

A cross-sectional study assessing AI-generated patient information guides on common cardiovascular conditions

Mustafa Sibaa[1], Hugo Douma[2,3]*, Ireene Elsa Mathew[1], Taha Kassim Dohadwala[4], Kundaranahalli Pradeep Harshath Odeker[5], Deepa Polinati[6], Nidhi Laxminarayan Rao[7]

[1]Department of Medicine, Tbilisi State University, Tbilisi, Georgia
[2]Department of Medical Sciences, Faculty of Medicine, International European University, Ukraine
[3]Department of Computer Science, College of Natural Sciences, University of Texas at Austin, USA
[4]Department of Medicine, Faculty of Medicine, David Tvildiani Medical University, Tbilisi, Georgia
[5]Department of Medicine, Vijayanagar Institute of Medical Sciences, Ballari, Karnataka, India
[6]Fortis Hospital, Bangalore, Karnataka, India
[7]Department of Medicine, KAP Vishwanathan Government Medical College, Tiruchirapalli, Tamil Nadu, India

Received: 07 November 2024
Revised: 10 December 2024
Accepted: 19 December 2024

***Correspondence:**
Dr. Hugo Douma,
E-mail: hugodouma@utexas.edu

ABSTRACT

Background: Patient education is essential for management of CVD as it enables in earlier diagnosis, early treatment and prevention of complications. Artificial intelligence is and increasingly popular resource with applications in virtual patient counselling. Thus, the study aimed to compare the AI generated response for patient education guide on common cardiovascular diseases using ChatGPT and Google Gemini.
Methods: The study assessed the responses generated by ChatGPT 3.5 and Google Gemini for patient education brochure on angina, hypertension, and cardiac arrest. Number of words, sentences, average word count per sentence, average syllables per word, grade level, and ease level were assessed using Flesch-Kincaid Calculator, and similarity score was checked using Quillbot. Reliability was assessed using modified DISCERN score. The statistical analysis was done using R version 4.3.2.
Results: The statistical analysis exhibited that there were no statistically significant differences between the responses generated by the AI tools based on different variables except for the ease score (p=0.2043), which was statistically superior for ChatGPT. The correlation coefficient between both the two tools was negative for the ease score (r=-0.9986, p=0.0332), the reliability score (r=-0.8660, p=0.3333), but was statistically significant for ease score.
Conclusions: The study demonstrated no significant differences between the responses generated by the AI tools for patient education brochures. Further research must be done to assess the ability of the AI tools, and ensure accurate and latest information is being generated, to benefit overall public well-being.

Keywords: Angina, Artificial intelligence, Cardiac arrest, ChatGPT, Educational tool, Google Gemini, Hypertension, Patient education brochure

INTRODUCTION

Cardiac arrest, angina and hypertension represent critical cardiovascular conditions, with more than half a billion people affected by these conditions worldwide.[1] Patient education plays a crucial role in the management of CVDs by allowing early identification of symptoms, lifestyle modifications, adherence to treatment regimens, and early prevention of complications, mortality, and morbidity.

Artificial intelligence (AI) tools such as ChatGPT and Google Gemini, have the ability to mimic the human brain, which could play an essential role in the field of medicine by allowing the identification, processing, integration, and analysis of various amounts of healthcare data.[2,3] These AI platforms allow for patient engagement and compliance by providing informative data on a wide array of diseases, including CVDs, enabling patients to gather accurate medical knowledge virtually from anywhere around the globe. On the other hand, Google search, although it provides a wide array of valuable information, lacks the personalized approach offered by these AI tools. While these platforms have the inherent benefit of providing easy accessibility to medical data, it is essential to acknowledge their limitations, including the accuracy of the medical information, a lack of personal connection with healthcare professionals, and the potential for overreliance on AI-powered tools as patients with cardiac health issues might potentially turn to these AI-powered tools for suggestions and recommendations, necessitating the need for the content produced to the verified.[3]

ChatGPT (chat generative pre-trained transformer) is a language model developed by OpenAI that generates appropriate and logical replies to user input by establishing a connection between the words and sentences in natural language.[4,5] Conversely, Google's Gemini AI is a "native multimodal" model, which utilizes Google's sophisticated algorithms to simulate human-like interactions, delivering, personalized search results and informative recommendations tailored to the individual's needs.[6,7]

Both ChatGPT and Google Gemini play a crucial role in patient counselling by providing a virtual platform for healthcare assistance and mental health support, addressing patient concerns and queries, and providing the appropriate support and recommendations to improve the overall wellbeing and treatment outcome of the patient.[8] These AI-powered tools serve to complement the existing traditional counselling methods, allowing patients to better understand and make informed decisions about their cardiovascular health.

Aims and objectives

To compare ChatGPT and Google Gemini generated responses for writing patient education guide on cardiac arrest, angina and hypertension, based on readability and ease of understanding.

METHODS

This cross-sectional research study was conducted over the span of one week, from February 13 to February 16, 2024, through virtual means. All author contributed equally via virtual methods, and the data utilized in this study was sourced from ChatGPT and Google Gemini. These sources, being publicly accessible platforms, thereby exempted the study from ethical approval.

The study aimed to assess the information generated related to three common cardiac diseases: cardiac arrest, angina, and hypertension, using two AI tools: ChatGPT version 3.5 and Google Gemini version 1.0. Each AI tool was tasked with generating patient education guides for the selected cardiac conditions through three prompts: "write a patient education guide for cardiac arrest", "write a patient education guide for angina', and "write a patient education guide for hypertension". The responses generated were compiled in a Microsoft Word document for further analysis.

Subsequently, the responses were graded using multiple tools:

The Flesch-Kincaid calculator was employed to assess word count, ease of understanding, and reliability of the generated information. This tool measures readability by calculating Flesch grade level, Flesch reading ease score, reading level, average words per sentence, average syllables per word, number of sentences, and word count.[9]

Plagiarism similarity was evaluated using the Quillbot plagiarism tool to ensure the originality of the content.

The reliability of scientific text was assessed using the modified DISCERN score, which involves rating responses on a 5-point scale based on predetermined criteria related to reliability and quality.[10]

Finally, statistical analysis was conducted using R version 4.3.2. The responses generated by ChatGPT and Google Gemini were compared using the Unpaired T-test, with significance set at $p<0.05$. Additionally, the correlation between ease score and reliability score was examined using Pearson's coefficient of correlation.

RESULTS

In this study, we employed ChatGPT and Google Gemini to generate brochures aimed at patient education across a spectrum of cardiovascular diseases, namely angina, hypertension as well as cardiac arrest cases.

Table 1 provides a detailed comparison of the characteristics of responses generated by ChatGPT and Google Gemini. Google Gemini had a higher mean word count (mean=455.7, SD=78.01) compared to ChatGPT (mean=396.67, SD=50.20), as well as a higher average words per sentence (mean=25.07, SD=24.82) compared to

ChatGPT (mean=7.57, SD=2.19). Additionally, Google Gemini scored higher on both mean ease score (mean=41.33, SD=10.76) and mean reliability score (mean=4, SD=1) compared to ChatGPT (mean=29.93, SD=2.20) and (mean=3.33, SD=0.58) respectively.

Table 1: Characteristics of responses generated by ChatGPT and Google Gemini.

Variables	ChatGPT		Google Gemini		P value*
	Mean	SD	Mean	SD	
Words	396.67	50.20	455.7	78.01	0.3422
Sentences	56.0	18.74	45.0	9.85	0.4338
Average words per sentence	7.57	2.19	25.07	24.82	0.3462
Average syllables per word	2	0	1.83	0.12	0.1296
Grade level	10.93	0.84	10.03	1.72	0.4771
Ease score	29.93	2.20	41.33	10.76	0.2043
Similarity %	27.73	7.38	24.77	22.13	0.8428
Reliability score	3.33	0.58	4	1	0.3868

*Unpaired t- test. P values <0.05 are considered statistically significant.

Google Gemini exhibited a lower mean sentence count (mean=45.0, SD=9.85) compared to ChatGPT (mean=56.0, SD=18.74). Google Gemini also had fewer mean syllables per word (mean=1.83, SD=0.12) compared to ChatGPT (mean=2, SD=0), a lower mean grade level (mean=10.03, SD=1.72) than ChatGPT (mean=10.93, SD=0.84), and a lower mean. Similarity percentage (mean=24.77, SD=22.13) compared to ChatGPT (mean=27.73, SD=7.38).

Statistical analysis revealed no statistically significant differences in the aforementioned parameters between the two AI models. No significant disparities were noted in the word count (p=0.3422), sentence count (p=0.4338), average words per sentence (p=0.3462), average syllables per word (p=0.1296), grade level (p=0.4771), similarity percentage (p=0.8428), and reliability score (p=0.3868). Notably, the ease score was significantly superior for ChatGPT-generated responses compared to those from Google Gemini (p=0.2043).

Table 2: Correlation between ChatGPT and Google Gemini for (a) ease score and (b) reliability score.

Variables	Correlation coefficient (r)	P value*
ease score	-0.9986	0.0332+
reliability score	-0.8660	0.3333

*Pearson's coefficient of correlation. + Significant at 5% level of significance.

Table 2 depicts the correlation between ChatGPT and Google Gemini, specifically focusing on the ease score and reliability score. The correlation coefficient (r) measures the strength of association between these variables, with values ranging from -1 to +1. Our findings revealed a strong negative correlation between ChatGPT and Google Gemini with respect to the ease score (r=-0.9986, p=0.0332) and the reliability score (r=-0.8660, p=0.3333). This indicates that as the ease score and reliability score for ChatGPT increase, those for Google Gemini decrease, and vice-versa. The correlation coefficient was found to be statistically significant for the ease score but not for the reliability score.

DISCUSSION

This cross-sectional study investigated the potential of AI-generated patient-information guides for cardiac arrest, angina and hypertension. Our findings provide valuable insights into the variable outputs of large language models (LLM) to educate patients on specific topics.

Artificial intelligence has already revolutionized many industries, and it is an integral part of the development in many sectors, including healthcare, with regard to patient education. It has the potential to improve the patient learning experience by providing tailored content to each individual patient, availability, and ease of access.[11] Bespoke LLMs can also be trained to generate reliable and easily understandable educational content leading to improved information comprehension and retention. In this study we focused on measuring the ease and reliability score to measure the performance of currently available large language models.[10,11] In addition to the number of words, sentences, and paragraphs, as per the model outputs, the average ease score calculated was 29.93 and 41.33 for ChatGPT and Gemini indicate that the text generated by both models is readable by college students or people with even higher education. Optimally, patient education text should be written to cater to people with at least high school education. Although Gemini had performed better than ChatGPT, both models weren't generating content easily readable by high school students, nevertheless, even manually written patient education materials written by medical professionals were found to often be harder to comprehend.[12]

Using large language models trained on the existing literature may lead to unintentional plagiarism, as these models can reproduce phrasing or exact sentences from their training materials.[13] Such unintentional plagiarism can include outdated content as well, since training material can be years old, which is specifically harmful in medicine since some revised treatment protocols may remove medications from previous protocol revisions, but large language models may still include these deprecated protocols as training material.[14]

The DISCERN score is a tool specifically designed to assess the quality and reliability of written information aimed at patients regarding treatment choices.[15] However, it does not directly address general online/media content reliability. In this study, the mean Modified DISCERN score for ChatGPT-generated content was 3.33/5.00, whereas the text produced by Google Gemini had a mean score of 4.00/5.00. In another article published in July 2023, ChatGPT performed worse than the other tools.[16] Furthermore, the ease score is another methodology to assess the reading ease of a given text; it ranges from 1 to 100; the higher the score, the easier it is to comprehend the text. In our study, Google's Gemini scored higher, with an average of 41.33 compared to 29.93 scored by OpenAI's ChatGPT. However, it is important to consider the rapid development of AI products and the release of new versions of Gemini and ChatGPT. Therefore, the quality of outputs can vary accordingly, and another study found that medical content generated by ChatGPT could be systematic and precise while coming short with reference errors and the absence of academic merit.[17]

This study provides a limited comparison of the outputs from two of the most common large language models (LLM) as of the time of writing, and other LLMs or niche-specific models may theoretically have better results, which warrants further research. Furthermore, this study focused solely on LLM-generated content for angina, hypertension, and cardiac arrest. Three of the most common cardiovascular conditions worldwide need to be expanded to cover the rest of the systems. This study used the latest publicly available versions of ChatGPT and Gemini. Large language models are typically trained in publicly available data up to a certain point in time (e.g., ChatGPT 3.5 training data temporal cutoff is January 2022); consequently, they are not up-to-date with the latest medical protocol updates.

In this study, only two AI tools- ChatGPT and Google Gemini, were compared. Moreover, the study focused solely on three cardiovascular conditions, which could limit its applicability to a wider range of healthcare scenarios.

The study used ChatGPT 3.5, a free version available to everyone. However, ChatGPT 4.5 exists, offering more features. This could have potentially led to variations in the answers generated. Advancements in medical science, algorithm bias and information bias have affected the ability of AI tools to consistently provide up-to-date information which could affect the overall quality of the medical data received by the patients.

CONCLUSION

This research indicates that there is no considerable dissimilarity in the standard ease, grade, and trustworthiness scores for answers produced by the two AI models for the patient information pamphlet on cardiac arrest, angina, and hypertension. There was no relationship between the ease score and dependability score for the two programs.

Additional studies should be conducted to investigate additional AI systems in other health conditions that are more widespread in society. Whether these AI systems are capable of generating content by utilizing the latest recommendations and research needs to be evaluated. Furthermore, AI systems should be developed to deliver updated information and references to data so that they can be authenticated. If it can provide certified data, it will be embraced by the public.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation; 2023.
2. Zargarzadeh A, Javanshir E, Ghaffari A, Mosharkesh E, Anari B. Artificial intelligence in cardiovascular medicine: An updated review of the literature. J Cardiovasc Thorac Res. 2023;15(4):204.
3. Sun X, Yin Y, Yang Q, Huo T. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. Eur J Med Res. 2023;28(1):242.
4. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, Lee AG, Tavakkoli A. GPT-4: a new era of artificial intelligence in medicine. Irish Journal of Medical Science (1971-). 2023;192(6):3197-200.
5. OpenAI. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt. Accessed on 3 April 2023.
6. Pichai S. An important next step on our AI journey. Google. 2023. Available from: https://blog.google/technology/ai/bard-google-ai-search-updates/. Accessed on 3 April 2023.
7. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024:1-6.
8. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689.

9. Flesch R. Flesch-Kincaid readability test. Retrieved October. 2007;26(3):2007.
10. Khazaal Y, Chatton A, Cochand S, Coquard O, Fernandez S, Khan R, et al. Brief DISCERN, six questions for the evaluation of evidence-based content of health-related websites. Patient Educ Counsel. 2009;77(1):33-7.
11. Eapen J, Adhithyan VS. Personalization and customization of llm responses. Int J Res Publicat Rev. 2023;4(12):2617-27.
12. Kasabwala K, Agarwal N, Hansberry DR, Baredes S, Eloy JA. Readability assessment of patient education materials from the American Academy of Otolaryngology- Head and Neck Surgery Foundation. Otolaryngol Head Neck Surg. 2012;147(3):466-71.
13. Howard J, Cheung HC. Artificial intelligence in medical writing. AsiaIntervention. 2024;10(1):12-4.
14. Wu K, Wu E, Cassasola A, Zhang A, Wei K, Nguyen T, Riantawan S, Riantawan PS, Ho DE, Zou J. How well do LLMs cite relevant medical references? An evaluation framework and analyses. arXiv preprint arXiv:2402.02008. 2024.
15. Kaicker J, Borg Debono V, Dang W. Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. BMC Med. 2010;8:59.
16. Golan R, Ripps SJ, Reddy R, Loloi J, Bernstein AP, Connelly ZM, et al. ChatGPT's Ability to Assess Quality and Readability of Online Medical Information: Evidence From a Cross-Sectional Study. Cureus. 2023;15(7):e42214.
17. Kumar AH. Analysis of ChatGPT Tool to Assess the Potential of its Utility for Academic Writing in Biomedical Domain. Biology, Engineering, Medicine and Science Reports. 2023;9:4–30.

Cite this article as: Sibaa M, Douma H, Mathew IE, Dohadwala TK, Odeker KPH, Polinati D, et al. A cross-sectional study assessing AI-generated patient information guides on common cardiovascular conditions. Int J Res Med Sci 2025;13:50-4.