Title: Appropriateness of ChatGPT in answering heart failure related questions (short title: ChatGPT answering heart failure questions)

Ryan C. King, MD[a], Jamil S. Samaan, MD[b], Yee Hui Yeo, MD, MSc[b], Behram Mody, MD[a], Dawn M. Lombardo, DO[a], Roxana Ghashghaei, MD[a]

Affiliations
[a]Division of Cardiology, Department of Medicine, University of California, Irvine School of Medicine, 333 City Blvd. West, Suite 400, Orange, California, USA 92868-3298
[b]Karsh Division of Gastroenterology and Hepatology, Department of Medicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd., Los Angeles, California, USA

Abstract

Background

Heart failure requires complex management with increased patient knowledge shown to improve outcomes. The large language model (LLM), Chat Generative Pre-Trained Transformer (ChatGPT), may be a useful supplemental resource of information for patients with heart failure.

Methods

Responses produced by GPT-3.5 and GPT-4 to 107 frequently asked heart failure-related questions were graded by two reviewers board-certified in cardiology, with differences resolved by a third reviewer. The reproducibility and accuracy between GPT-3.5 and GPT-4 were compared for questions involving basic knowledge, management, prognosis, procedures, and support.

Results

GPT-4 displayed a greater proportion of comprehensive knowledge for the categories of “basic knowledge” and “management”, while GPT-3.5 performed better in the “other” category (prognosis, procedures, and support) (94.1% vs 64.7%). There were 2 total responses (1.9%) graded as “some correct and incorrect” for GPT-3.5, while no GPT-4 responses received a grade of “some correct and incorrect” or “completely incorrect”. Both models provided highly

reproducible responses, with GPT-3.5 scoring above 94% in every category and GPT-4 with 100% for all answers.

Conclusions

Both GPT-3.5 and GPT-4 answered the majority of heart failure-related questions accurately and reliably, with GPT-4 displaying superior performance overall. ChatGPT may lead to better outcomes in patients with heart failure by providing health education.

Keywords

ChatGPT, heart failure, large language models, artificial intelligence

Introduction

Heart failure is a chronic condition, and the healthcare burden in the United States (US) is forecasted to grow to approximately $70 billion annually by 2030. Hospitalizations comprise much of this cost (70%) and account for 1-2% of total hospitalizations in the US[1]. Increased patient knowledge regarding the management of their heart failure condition has been shown to lead to fewer and shorter duration of hospital admissions[2]. In search of answers, individuals often utilize online resources for information regarding their health, resulting in approximately one billion healthcare-related questions being searched on Google each day[3].

The large language model (LLM), Chat Generative Pre-Trained Transformer (ChatGPT), is an artificial intelligence (AI) model that was trained on a large dataset comprising a vast spectrum of topics and media, including medicine. It can provide text-based responses in a conversational manner to questions prompted by users[4]. Beyond its rapid growth in popularity, the model continues to improve in performance with the latest version, GPT-4, released in March of 2023,

which has shown to significantly outperform its predecessor, GPT-3.5[5]. As ChatGPT continues to grow in its capabilities, patients will be more inclined to seek information from this model and other similar LLMs regarding their healthcare. There may be a supplemental role for this technology applied to highly prevalent conditions requiring complex care, such as heart failure.

The model's utility in medicine is actively under investigation, with its basic knowledge and reasoning being tested. Prior studies have examined ChatGPT's ability to answer questions related to heart disease prevention, bariatric surgery, and cirrhosis yielding promising results[6,7,8]. We built on these studies by assessing 1) Accuracy of ChatGPT when answering questions related to heart failure 2) Reproducibility of its responses and 3) Improvement in performance between GTP-3.5 and GPT-4.

Methods

We curated a list of 125 frequently asked questions related to heart failure from medical societies, renowned medical institutions, and Facebook support groups. A total of 18 questions were excluded due to their duplicate content, nonspecific phrasing, and not being considered from a patient's perspective. The final set of 107 questions was then entered twice into each model (GPT-3.5 and GPT-4) using the "new chat" function, producing two responses per question per model. Responses were first graded independently by two board-certified cardiologists. Accuracy was graded using the following scale: 1) Comprehensive 2) Correct but inadequate 3) Some correct and some incorrect 4) Completely incorrect. This process was completed for responses from both GPT-3.5 and GPT-4. Reviewers also assessed reproducibility, defined as similar comprehensiveness and accuracy scores (1 and 2 vs. 3 and 4) between two responses per question for each model. Discrepancies in grading between the two reviewers were

resolved by a third reviewer board-certified in advanced heart failure with more than 20 years of clinical experience. Microsoft Excel (version 16.68) was used for all statistical analysis.

Results

The majority of responses from both models were graded as either "comprehensive" or "correct but inadequate" (**Table 1**). Overall, GPT-4 displayed greater comprehensive knowledge for the categories of "basic knowledge" and "management", while GPT-3.5 performed better in the "other" category (prognosis, procedures, and support) (94.1% vs 64.7%). For example, GPT-3.5 answered in general terms regarding the cardiac benefits of SGLT2 inhibitors, while GPT-4 provided a more detailed yet succinct response regarding their effects on diuresis and blood pressure. There were 2 total responses (1.9%) graded as "some correct and some incorrect" for GPT-3.5, while no GPT-4 responses received a grade of "some correct and some incorrect" or "completely incorrect". When examining reproducibility, both models provided reproducible responses for the majority of questions, with GPT-3.5 scoring above 94% in every category and GPT-4 with 100% for all answers (**Table 2**).

	GPT-3.5	**GPT-4**
Overall (N=107)		
1. Comprehensive	84 (78.5%)	89 (83.2%)
2. Correct but incomplete	21 (19.6%)	18 (16.8%)
3. Some correct and some incorrect	2 (1.9%)	0 (0.0%)
4. Completely incorrect	0 (0.0%)	0 (0.0%)
Basic Knowledge (N=49)		

1. Comprehensive	36 (73.5%)	44 (89.8%)
2. Correct but incomplete	13 (26.5%)	5 (10.2%)
3. Some correct and some incorrect	0 (0.0%)	0 (0.0%)
4. Completely incorrect	0 (0.0%)	0 (0.0%)
Management (N=41)		
1. Comprehensive	32 (78.1%)	34 (82.9%)
2. Correct but incomplete	8 (19.5%)	7 (17.1%)
3. Some correct and some incorrect	1 (2.4%)	0 (0.0%)
4. Completely incorrect	0 (0.0%)	0 (0.0%)
Other (N= 17)		
1. Comprehensive	16 (94.1%)	11 (64.7%)
2. Correct but incomplete	0 (0.0%)	6 (35.3%)
3. Some correct and some incorrect	1 (5.9%)	0 (0.0%)
4. Completely incorrect	0 (0.0%)	0 (0.0%)

Table 1. Grading of ChatGPT-3.5 and ChatGPT-4 responses for questions related to heart failure. The "basic knowledge" category included questions involving causes of heart failure, general definitions, symptoms, and diagnosis. The "management" category included questions related to general management of heart failure, medications, and lifestyle. Questions grouped in the "other" category were related to prognosis, procedures, and support.

	GPT-3.5	GPT-4
Overall (N=107)	105 (98.1%)	107 (100.0%)
Basic Knowledge (N=49)	49 (100%)	49 (100.0%)
Management (N=41)	40 (97.6%)	41 (100.0%
Other (N= 17)	16 (94.1%)	17 (100.0%)

Table 2. Reproducibility of ChatGPT-3.5 and ChatGPT-4 responses categorized by question type. Reproducibility was defined as no difference in grading categories between the two responses for each question. Grading categories were grouped by whether scores contained incorrect information. Scores of 1 and 2 meant no incorrect information was present, while 3 and 4 noted there to be incorrect information.

Discussion

We examined the accuracy and reproducibility of responses by the large language models ChatGPT-3.5 and GPT-4 to questions related to heart failure. GPT-4 outperformed GPT-3.5 by providing more comprehensive responses (83.2% vs. 78.5) as well as no incorrect responses. Both models also provided reproducible responses to the majority of questions overall (100% vs 94%). The results of our study show ChatGPT's impressive ability to provide comprehensive and reliable responses to patients' questions as well as the impressive improvement in performance by these models over such a short period of time. Our findings highlight the potential utility of LLMs as an accurate and reliable resource for patients with heart failure to potentially use under the care of their healthcare provider.

ChatGPT may potentially lead to better outcomes through empowerment with knowledge as seen in prior studies investigating the effect of health education on heart failure management[2]. We anticipate patients will continue to seek answers related to their health from ChatGPT due to its simple user interface and easy-to-understand human-like responses provided in a conversational format. GPT-4's improved performance can be attributed to its training which focused on better understanding of users' intent and processing of more complex scenarios[5]. The impressive performance of ChatGPT in this study shows that this emerging technology may be a useful tool for both patients and providers in the future. We recommend that researchers and clinicians continue investigating the capabilities and limitations of ChatGPT to maximize its impact on improving patient outcomes.

Although ChatGPT performed well with few incorrect responses in this study, there are limitations to be considered. On occasion, the model may respond with inaccurate information that is organized in a believable manner that may be deceiving to users due to its human-like text. Nonsensical responses may potentially be produced as well[4]. Furthermore, the accuracy of the model depends on the dataset that it was trained on, which has not been disclosed. The consistency of recommendations may also vary from region to region. Limitations of this study include the inability to blind the reviewers to the identity of each ChatGPT model version due to the unexpected release of GPT-4. Although a grading system involving a panel of multiple reviewers was used in this study, there may still be bias introduced through subjective review.

Conclusion

Both GPT-3.5 and GPT-4 provided accurate and reliable responses to the majority of heart failure-related questions. Notably, GPT-4 provided no incorrect responses to the questions

provided. The superior performance of GPT-4 demonstrates the impressive ability of these models to improve over a short period of time and highlights their future potential as an adjunct source of information for patients with heart failure. We recommend future investigation into the capabilities and limitations of ChatGPT to identify the impact on improving patient outcomes.

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovasc Res*. 2023 Jan 18;118(17):3272-3287. Erratum in: Cardiovasc Res. 2023 Feb 09; doi: 10.1093/cvr/cvac013.

2. Ditewig JB, Blok H, Havers J, van Veenendaal H. Effectiveness of self-management interventions on mortality, hospital readmissions, chronic heart failure hospitalization rate and quality of life in patients with chronic heart failure: a systematic review. *Patient Educ Couns*. 2010 Mar;78(3):297-315. Epub 2010 Mar 3. doi: 10.1016/j.pec.2010.01.016.

3. Murphy, M. 'Dr Google will see you now: Search giant wants to cash in on your medical queries', *The Telegraph*, 10 March 2019, https://www.telegraph.co.uk/technology/2019/03/10/google-sifting-one-billion-health-questions-day/. Accessed March 2023.

4. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2023, https://online-chatgpt.com. Accessed 18 February 2023.

5. OpenAI. GPT-4 Technical Report. 2023.

6. Sarraju A, Bruemmer D, Iterson EV, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. 2023 Mar 14;329(10):842-844. doi: 10.1001/jama.2023.1044.

7. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Srinivasan N, Park J, Burch M, Watson R, Liran O, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes Surg*. 2023 Jun;33(6):1790-1796. doi: 10.1007/s11695-023-06603-5.

8. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023 Mar 22. doi: 10.3350/cmh.2023.0089.