A multidisciplinary assessment of ChatGPT's knowledge of amyloidosis

Ryan C. King, MD[a*], Jamil S. Samaan, MD[b], Yee Hui Yeo, MD, MSc[b], David C. Kunkel, MD[c], Ali A. Habib, MD[d], Roxana Ghashghaei, MD[a]

*Corresponding author: kingrc@hs.uci.edu

Affiliations

[a]Division of Cardiology, Department of Medicine, University of California, Irvine Medical Center, 101 The City Dr. S, Orange, California, USA

[b]Karsh Division of Gastroenterology and Hepatology, Department of Medicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd., Los Angeles, California, USA

[c]Division of Gastroenterology, Department of Medicine, University of California, San Diego Medical Center, 200 W Arbor Dr, San Diego, California, USA

[d]Division of Neurology, University of California, Irvine Medical Center, 101 The City Dr. S, Orange, California, USA

Abstract

Amyloidosis is a rare, multisystem disease with several subtypes including AA (secondary), AL (amyloid light chain), and ATTR (transthyretin amyloidosis). In addition to variable symptoms and multidisciplinary management, amyloidosis being a rare disease further contributes to patients being at risk for decreased health literacy regarding their condition. Increased access to education materials containing simple, plain language may bridge literacy gaps and improve outcomes for patients with rare diseases such as amyloidosis. The large language model (LLM), Chat Generative Pre-Trained Transformer (ChatGPT), may be a powerful tool for improving the availability of accurate and easy to understand education materials. Amyloidosis-related questions from cardiology, gastroenterology, and neurology were sourced from esteemed medical

societies and institutions along with amyloidosis Facebook support groups and inputted into ChatGPT-3.5 and GPT-4. Answers were graded on 4-point scale with both models responding to the majority of questions with either “comprehensive” or “correct but inadequate” answers with only 1 (1.2%) answer by GPT-3.5 graded as “completely inaccurate”. When assessing reproducibility, GPT-3.5 scored reliably on more than 83.3% of responses, while GPT-4 produced above 98.2% consistent answers. Our findings show that ChatGPT can potentially serve as a supplemental tool in disseminating vital health education to patients living with amyloidosis.

Keywords: amyloidosis, ChatGPT, multidisciplinary, large language model, rare disease

Introduction

Amyloidosis is a chronic multisystem disease that comprises several subtypes including AA (secondary), AL (amyloid light chain), and ATTR (transthyretin amyloidosis), with the latter two being the most common but often underdiagnosed [1]. AL amyloidosis is diagnosed in roughly 2,500 to 5,000 individuals annually in the United States (US), while the exact incidence of ATTR and AA remain unknown due to challenges and delays in diagnosis. This uncertainty is attributed to the diverse range of symptoms affecting this patient population who often have varying presentations affecting multiple organ systems [2, 3]. Diagnosing and caring for patients living with amyloidosis relies on effective multidisciplinary collaboration between specialists in fields including but not limited to cardiology, gastroenterology, and neurology [4].

In addition to the complex nature of symptoms and management, Amyloidosis is also considered a rare disease, a designation further contributing to this patient population

being at risk for decreased health literacy regarding their condition. A notable scarcity of patient education materials (PEMs) exists for rare diseases compared to common ones, with nearly a tenfold difference which has previously been shown to adversely affect health outcomes [5]. The Centers for Disease Control and Prevention (CDC) posits that improved health literacy could prevent up to one million hospitalizations annually and save $25 billion in total healthcare costs [6]. Increased access to education materials containing simple, plain language is a promising strategy to help bridge literacy gaps and improve outcomes especially for patients with rare diseases such as amyloidosis.

Artificial intelligence (AI), an emerging technology, may be a powerful tool for improving the availability of accurate and easy to understand information for rare and complex diseases like amyloidosis. Chat Generative Pre-Trained Transformer (ChatGPT), an AI-driven large language model (LLM) released in late 2022, has gained widespread adoption, attracting 1.8 billion users per month. [7]. Unlike traditional search engines, which return web page listings, ChatGPT generates human-like text in a structured, conversational format through an intuitive user interface. This is achieved via reinforcement learning from human feedback (RLHF), wherein the model's responses are refined through feedback loops to optimize performance [8]. With ongoing improvement and training on an extensive dataset spanning diverse topics including medicine, ChatGPT's accuracy and reliability in answering questions are expected to increase. In March of this year GPT-4.0, the predecessor to the original GPT-3.5, was released and has demonstrated its superior performance in answering clinical questions [9, 10]. This rapid improvement in performance over a short period of time makes large language models a potential asset for both patients and healthcare providers seeking information on diseases like amyloidosis.

As with any emerging technology, rigorous evaluation is essential to ensure its efficacy and safety. It is imperative to evaluate the capabilities and limitations of these models during their nascent stages to detect knowledge gaps before their broad adoption by patients and providers. Earlier studies have demonstrated ChatGPT's impressive accuracy and reliability in answering clinical questions related to coronary artery disease, cirrhosis, and bariatric surgery [11, 12, 13]. This study aims to build upon previous literature by employing a multidisciplinary approach to assessing ChatGPT's 1) accuracy in answering questions related to amyloidosis, particularly concerning cardiology, gastroenterology, and neurology; 2) reproducibility of responses; and 3) performance improvement of GPT-4 compared to GPT-3.5.

Materials and Methods

A total of 98 amyloidosis-related questions were sourced from esteemed medical societies and institutions and inputted into ChatGPT. Questions from amyloidosis Facebook support groups were incorporated for a more comprehensive patient perspective. Of these, 56 addressed general amyloidosis topics, while 42 were specific to cardiology (12), gastroenterology (15), and neurology (15). Each question was inputted twice into both GPT-3.5 and GPT-4, yielding two distinct responses per question for each model. Neurology-related questions were only inputted into GPT-4. Responses were assessed on a scale: 1) Comprehensive, 2) Correct but inadequate, 3) some correct and some incorrect 4) Completely incorrect. Reproducibility was evaluated by categorizing responses into those containing either no incorrect information (grades 1 and 2) or those with some or completely incorrect information (grades 3 and 4). Two independent reviewers, board-certified in cardiology and

gastroenterology with expertise in amyloidosis, assessed general questions and questions in their respective specialties. Discrepancies in general question grading were resolved through discussion to reach a consensus. An additional reviewer, board-certified in neurology and specializing in amyloidosis, graded the neurology-specific responses for GPT-4. Microsoft Excel (version 16.68) was used to conduct the statistical analysis.

Results

Both ChatGPT models responded to the majority of questions with either "comprehensive" or "correct but inadequate" answers (**Table 1**). GPT-4 demonstrated comprehensive responses more frequently than GPT-3.5 for general questions (94.6% vs 85.7%) and gastroenterology (60.0% vs 53.3%). For specialty-specific responses, Cardiology was graded the highest for both models with both receiving 83.3% comprehensive scores. There was a total of 8 (9.6%) responses for GPT-3.5 containing "some correct and some incorrect" information compared to 5 (5.1%) for GPT-4. One gastroenterology question answered by GPT-3.5 received the only (1.2%) "completely incorrect" grade in response to the evidence for using supplements like probiotics and digestive enzymes to enhance digestion. When assessing reproducibility of specialty-specific responses, GPT-3.5 scored reliably 96.4% on general questions, 83.3% for cardiology, and 93.3% for gastroenterology (**Table 2**). GPT-4 produced reproducible responses for 98.2% of general responses and 100% of responses for all specialties. A distinction between models was observed in responses to the prevalence and presentation of systemic multiorgan amyloidosis. GPT-3.5 gave a broad overview

omitting the ATTR subtype, whereas GPT-4 included details on systemic involvement for each subtype.

	GPT-3.5	GPT-4
General, Total (N=56)		
1	48 (85.7%)	53 (94.6%)
2	4 (7.1%)	3 (5.4%)
3	4(7.1%)	0 (0.0%)
4	0 (0.0%)	0 (0.0%)
Differences in grading	15 (25.9%)	17 (29.3%)
Cardiology (N=12)		
1	10 (83.3%)	10 (83.3%)
2	0 (0.0%)	2 (16.7%)
3	2 (16.7%)	0 (0.0%)
4	0 (0.0%)	0 (0.0%)
GI (N=15)		
1	8 (53.3%)	9 (60%)
2	4 (26.7%)	3 (20%)
3	2 (13.3%)	3 (20%)
4	1 (6.7%)	0 (0.0%)
Neurology (N=15)		
1	-	10 (66.7%)
2	-	3 (20%)
3	-	2 (13.3%)
4	-	0

Table 1. Accuracy of ChatGPT for answers to amyloidosis related questions.

Accuracy grading was based on a scale of 1 = comprehensive, 2 = correct but inadequate, 3 = some correct and some incorrect, and 4 = completely incorrect. Differences in grading between reviewers were resolved through discussion to arrive at a final score for a given answer.

	GPT-3.5	GPT-4
General (N=56)	54 (96.4%)	55 (98.2%)
Cardiology, (N=12)	10 (83.3%)	12 (100%)
GI (N=15)	14 (93.3%)	15 (100%)
Neurology (N=15)	-	15 (100%)

Table 2. Reproducibility of ChatGPT answers for amyloidosis related answers.

Reproducibility was graded based on responses containing correct or incorrect information with grouping of scores of 1 and 2 (comprehensive; correct but inadequate) vs 3 and 4 (some correct and some incorrect; completely incorrect) together.

Discussion

Large language models are an emerging technology

There is a growing body of literature examining ChatGPT's knowledge related to common and prevalent health conditions, but studies evaluating its performance for rare diseases are limited. In this study, we used an interdisciplinary panel of experts in amyloidosis from cardiology, gastroenterology, and neurology to evaluate the accuracy and reliability of GPT-4 and GPT-3.5 in answering amyloidosis related questions. Both models produced comprehensive responses to over 85% of general questions, with GPT-4 outperforming. GPT-3.5 produced responses containing inaccurate information

slightly more often than GPT-4 (10.8% vs 5.1%) and provided the only "completely inaccurate" response of the study for one gastroenterology question. For cardiology questions, both models surpassed their performance in gastroenterology and neurology. This higher proficiency may stem from the prevalence of cardiac manifestations in amyloidosis and the models' possible enhanced exposure to relevant data during training. With over 83.3% reproducibility for GPT-3.5 and over 98.2% for GPT-4, GPT's high reliability and accuracy in this study further bolsters its prospective utility in aiding patients and providers to improve amyloidosis outcomes through enhanced patient health education. While its performance is impressive, we stress the role of these large language models as adjunct rather than replacement of care provided by a team of licensed healthcare professionals.

Previous studies have also shown ChatGPT's commendable performance in accuracy and reliability concerning cardiovascular disease prevention queries, with the majority of responses deemed appropriate and dependable [11]. In more intricate scenarios encompassing clinical vignettes describing atrial fibrillation, congenital heart disease, heart failure, and cholesterol levels, ChatGPT's responses were assessed as predominantly reliable, valuable for patients, and crucially, not hazardous. Impressively, many of these responses were favored over those generated by a standard Google search [14].

Yeo et al. (2023) demonstrated that GPT-3.5 accurately answered over 75% of questions related to basic knowledge and 66% of diagnostic questions concerning cirrhosis and hepatocellular carcinoma [12]. Notably, in a follow up study on cirrhosis the authors demonstrated that GPT-4 significantly improved over GPT-3.5, providing

superior performance and rectifying errors made by its predecessor [9]. Another improvement by GPT-4 over GPT-3.5 was seen in a prior study examining GPT's performance on the United Kingdom neurology licensing exam revealed that GPT-3.5 did not achieve a passing score, while GPT-4 passed comfortably [15]. Additionally, this study highlighted GPT's generalizability in providing accurate information based on medical guidelines from regions outside the US. These findings indicate that both ChatGPT models can reliably provide accurate information on a wide range of clinical queries and that their capabilities are continually evolving. The discrepancy in accuracy and minimal improvement between models seen in our study compared to prior work may be due to the burgeoning data regarding amyloidosis of the GI tract and its rare nature.

Given the relatively recent release of ChatGPT, data testing the Large Language Model's (LLM) clinical accuracy concerning rare diseases such as amyloidosis is scarce. Mehnan et al. have demonstrated remarkable diagnostic precision for both common and rare diseases, with GPT-4 notably outperforming GPT-3.5. Interestingly, the authors suggested that GPT's responses weren't merely a reiteration of existing online content, attributing its capacity to suggest a comprehensive differential diagnosis to an understanding of its rationale [10]. This assessment of GPT displaying a near human-like understanding may be due to its increased training on human dialogue through RLHF and increased parameters included in the GPT-4 dataset [16].

Although the use of ChatGPT should always complement a healthcare provider's guidance, this emergent technology could prove beneficial for both patients and providers when applied to rare diseases like amyloidosis in the future but in its current

state the model requires further testing of its limitations. The LLM has the potential to simplify and increase accessibility of patient education materials (PEMs), thereby fostering health education-driven empowerment through conversational interactions. With the continuous evolution of ChatGPT's capabilities and the easy-to-use interface, it is expected that its user base will correspondingly expand. The prospect of expedited diagnoses and establishing care with a multidisciplinary medical team in a timelier manner could be additional impacts of ChatGPT and ultimately improve outcomes for patients living with amyloidosis.

Strengths and Limitations

This study is among the first in employing a multidisciplinary approach to evaluate ChatGPT's knowledge of amyloidosis, leveraging expertise from physicians across various specialties. This holistic approach enabled a thorough assessment of ChatGPT's abilities in addressing clinical queries related to amyloidosis, a rare disease necessitating advancements in health education, diagnostics, and management for improved patient outcomes. Moreover, this is the inaugural study to scrutinize GPT's performance in the context of amyloidosis.

ChatGPT's limitations encompass the undisclosed nature of its primary training dataset and the absence of citations in its responses to medical queries. Incorporating references to reputable sources, such as esteemed medical society websites or peer-reviewed studies, would enhance clinical relevance and reliability. Additionally, ChatGPT occasionally exhibits a phenomenon termed 'hallucinations,' wherein it generates confident but entirely spurious responses [8].

While this study's multidisciplinary approach was comprehensive, it relied on a single physician reviewer from each specialty. Future research could bolster validity by

engaging multiple reviewers within each specialty to minimize the potential for subjective bias. It would also be beneficial to include physicians specializing in hematology, oncology, and nephrology as reviewers due to the integral involvement of those specialties in caring for patients with amyloidosis.

Conclusion

ChatGPT delivered accurate and reliable responses to amyloidosis related questions across general and specialty-specific queries. ChatGPT can potentially serve as a supplemental tool in disseminating vital health education to patients and assisting providers in addressing diagnostic challenges of this rare disease. However, the presence of some incorrect responses underscores the necessity of utilizing this technology alongside a team of licensed healthcare professionals.

Acknowledgements

ChatGPT-4 was used in the editing process of this manuscript.

Funding Statement

There was no funding obtained for this study.

Declaration of interest statement

The authors have no disclosures or conflicts of interest to declare.

References

[1] Papingiotis G, Basmpana L, Farmakis D. Cardiac amyloidosis: epidemiology, diagnosis and therapy. *E-Journal of Cardiology Practice*. 2021 Apr, 19;19-21.

[2] Bajwa F, O'Connor R, Ananthasubramaniam K. Epidemiology and clinical manifestations of cardiac amyloidosis. *Heart Fail Rev*. 2022 Sep;27(5):1471-1484. doi: 10.1007/s10741-021-10162-1.

[3] Kharoubi M, Bézard M, Galat A, et. al. History of extracardiac/cardiac events in cardiac amyloidosis: prevalence and time from initial onset to diagnosis. *ESC Heart Fail*. 2021 Dec;8(6):5501-5512. doi: 10.1002/ehf2.13652.

[4] Kittleson M, Ruberg F, et al. 2023 ACC Expert Consensus Decision Pathway on Comprehensive Multidisciplinary Care for the Patient with Cardiac Amyloidosis. *J Am Coll Cardiol*. 2023 Mar, 81 (11) 1076–1126. doi.org/10.1016/j.jacc.2022.11.022.

[5] Falcão M, Allocca M, Rodrigues AS, et. al. A Community-Based Participatory Framework to Co-Develop Patient Education Materials (PEMs) for Rare Diseases: A Model Transferable across Diseases. *Int J Environ Res Public Health*. 2023 Jan 5;20(2):968. doi: 10.3390/ijerph20020968.

[6] Talking Points About Health Literacy. *CDC- Centers for Disease Control and Prevention*. Last updated May 21. 2021, https://www.cdc.gov/healthliteracy/shareinteract/TellOthers.html#.

[7] Duarte F. Number of ChatGPT Users (2023). *Exploding Topics*, 13 July 2023, https://explodingtopics.com/blog/chatgpt-users#:~:text=users%20are%20American-,How%20Many%20ChatGPT%20Users%20Are%20There%3F,1.8%20billion%20visitors%20per%20month.

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

[9] Yeo YH, Samaan JS, Ng WH, et. al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. medrxiv [Preprint]. May 5, 2023. doi: https://doi.org/10.1101/2023.05.04.23289482

[10] Mehnan L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. medrxiv [Preprint]. April 27, 2023. doi: https://doi.org/10.1101/2023.04.20.23288859

[11] 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.

[12] Yeo YH, Samaan JS, Ng WH, et, al. 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.

[13] 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.

[14] Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value and danger of ChatGPT-generated responses to health questions. Eur J Cardiovasc Nurs. 2023 Apr 24:zvad038. doi: 10.1093/eurjcn/zvad038.

[15] Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open. 2023 Jun 15;5(1):e000451. doi: 10.1136/bmjno-2023-000451.

[16] OpenAI. GPT-4 Technical Report. 2023.