Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation

Ivan Zeljkovic, MD, PhD,[1,2] Matea Novak, JD,[2,3,5] Ana Jordan, MD,[1] Ante Lisicic, MD,[1] Tatjana Nemeth-Blažić, MD, PhD,[4] Nikola Pavlovic, MD, PhD,[1] Šime Manola, MD, PhD[1]

From the [1]Department of Cardiovascular Diseases, Dubrava University Hospital, Avenija Gojka Šuška, Zagreb, Croatia, [2]Catholic University of Croatia, Zagreb, Croatia, [3]RIT Croatia, Rochester Institute of Technology, Zagreb, Croatia, and [4]Croatian Institute of Public Health, Zagreb, Croatia.

BACKGROUND As artificial intelligence and large language models continue to evolve, their application in health care is expanding. OpenAI's Chat Generative Pre-trained Transformer 4 (ChatGPT-4) represents the latest advancement in this technology, capable of engaging in complex dialogues and providing information.

OBJECTIVE This study explores the correctness of ChatGPT-4 in informing patients about atrial fibrillation.

METHODS This cross-sectional observational study involved ChatGPT-4 in responding to a structured set of 108 questions across 10 categories related to atrial fibrillation. These categories included basic information, treatment options, lifestyle adjustments, and more, reflecting common patient inquiries. The model's responses were evaluated by a panel of 3 cardiologists on the basis of accuracy, comprehensiveness, clarity, relevance to clinical practice, and patient safety. The total correctness of ChatGPT-4 was quantitatively assessed through scores assigned in each category, and statistical analysis was performed to identify significant differences in performance across categories.

RESULTS ChatGPT-4 provided correct and relevant answers with considerable variability across categories. It excelled in "Lifestyle Adjustments" and "Daily Life and Management" with perfect and near-perfect scores but struggled with "Miscellaneous Concerns" scoring lower. Statistical analysis confirmed significant differences in total scores across categories ($P = .020$).

CONCLUSION Our results suggest that while ChatGPT-4 is reliable in categories with structured and direct queries, it shows limitations when handling complex medical queries that require in-depth explanations or clinical judgment. ChatGPT-4 demonstrates promising potential as a tool for patient-focused informing in atrial fibrillation, particularly in straightforward informing content.

KEYWORDS GPT-4; Patient; Informing; Atrial fibrillation; Large language models; AI in health care

(Heart Rhythm O^2 2025;6:58–63) © 2024 Heart Rhythm Society. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Atrial fibrillation (AF) is the most common arrhythmia worldwide, affecting millions and contributing significantly to morbidity and health care burden.[1,2] As health care increasingly uses technology, large language models (LLMs) are emerging as valuable tools in this domain.[3–7] These artificial intelligence (AI) systems, developed through advanced machine learning techniques such as deep learning and natural language processing, are proficient at parsing and generating human-like text.[8] They show great promise in interpreting complex medical data and improving clinical decision making.[9] For instance, Chat Generative Pre-trained Transformer (ChatGPT) has demonstrated notable success on examinations such as the United States Medical Licensing Examination, achieving scores at or above passing thresholds.[10] Moreover, in studies evaluating ChatGPT-4's clinical reasoning and soft skills, its performance often surpassed that of previous users of AMBOSS, a leading medical learning platform.[11] ChatGPT's understanding of standards such as the American Heart Association's Basic Life Support and Advanced Cardiovascular Life Support shows its potential for guideline-aligned medical responses.[12] Despite these advancements, ChatGPT's performance varies. For example, while it excelled in the Ophthalmic Knowledge Assessment Program, results differed across subspecialties.[13,14] A recent assessment of GPT-3.5 and ChatGPT-4 on the Polish Medical Final Examination showed that ChatGPT-4 achieved a mean accuracy of 79.7%, surpassing the passing criteria for all Medical Final Examination versions, yet generally scoring below medical students.[15] This variability highlights the ongoing challenges and the critical need to evaluate the accuracy and applicability of LLMs in complex medical contexts, which partly motivates this study. In addition, ChatGPT-4,

[5]Present address: Student at School of Medicine, Catholic University of Croatia, Zagreb, Croatia. **Address reprint requests and correspondence:** Dr Ivan Zeljkovic, Department of Cardiovascular Diseases, Dubrava University Hospital, Avenija Gojka Šuška 6, 10 000 Zagreb, Croatia. E-mail address: ivanzeljkov@gmail.com.

https://doi.org/10.1016/j.hroo.2024.10.005

KEY FINDINGS

- Within the comprehensive evaluation of Generative Pre-trained Transformer 4's (ChatGPT-4's) correctness in patient-focused informing for atrial fibrillation, ChatGPT-4 can provide informative and relevant answers across different categories.
- ChatGPT-4's performance tends to vary across different sets of questions.
- ChatGPT-4 is reliable in categories with structured and direct queries.
- ChatGPT-4 shows limitations when handling complex medical queries that require in-depth explanations or clinical judgment.

capable of complex dialogues, now processes graphical inputs, enhancing its clinical utility.[16,17]

Another relevant study evaluated patient information on AF and cardiac implantable electronic devices provided by 3 different LLMs: Google Bard, Bing Chat, and ChatGPT Plus.[18] The findings indicated that while the responses generated by these LLMs were generally easy to understand, their reliability and completeness varied, highlighting the importance of cautious use in clinical settings. Building on this, our study aimed to thoroughly assess ChatGPT-4's ability to inform patients about AF, covering a wide range of questions from understanding the condition to optimizing care.

Methods

Study design and objectives

This study used ChatGPT-4 (OpenAI, San Francisco, CA) to answer a series of patient-centered questions about AF, using the text-based interfaces provided by OpenAI as of February 2024. The study aimed to assess ChatGPT-4's performance in conveying complex medical information to patients diagnosed with or at risk of AF. The questions were sourced from real-life patient interactions during medical consultations and hospitalizations, capturing a wide range of concerns from basic information about the condition to detailed inquiries about symptoms, diagnosis, treatment, and living with the disease. ChatGPT-4 was selected for this study because of its recognition as the most advanced and sophisticated LLM available at the time of the experiment, known for its superior performance in generating accurate and coherent responses.[11,14,15]

Question compilation and categorization

We compiled a comprehensive list of 108 questions related to AF, categorized into 10 distinct categories/themes:

Category 1 - Basics of Atrial Fibrillation (11 questions)
Category 2 - Signs and Symptoms (9 questions)
Category 3 - Diagnostic Tests (10 questions)
Category 4 - Treatment Modalities (17 questions)
Category 5 - Pulmonary Vein Isolation (PVI) Procedural and Post-Procedural Care (11 questions)
Category 6 - Comorbidities and Complications 15 questions)
Category 7 - Lifestyle Adjustments (13 questions)
Category 8 - Daily Life and Management (10 questions)
Category 9 - Stroke Prevention (7 questions)
Category 10 - Miscellaneous Concerns (5 questions)

All questions posed to ChatGPT-4 during the study are given in Online Supplemental Appendix 1.

Data collection protocol

Each question was presented to ChatGPT-4 through the OpenAI text-based chat interface in a new thread, with the model's temperature parameter set at the default level of 1.0, ensuring a balance between response variability and reliability. This setting was intended to simulate a realistic interactive session between a patient and an AI-based informing tool.

Evaluation metrics

The evaluation framework consisted of 5 key metrics (Table 1):

1. Accuracy and Evidence Base: Assessment of the factual accuracy of the responses and their adherence to the latest guidelines and research.
2. Comprehensiveness: Evaluation of the depth and breadth of the information provided.
3. Clarity and Understandability: Determination of how well the responses could be understood by patients with varying levels of health literacy.
4. Relevance to Clinical Practice: Relevance of the AI-generated advice to actual clinical scenarios.

Table 1 Evaluation metrics and explanations

Criterion	Explanation
Accuracy and Evidence Base	Is the answer supported by the latest research and practice, clinical guidelines, and consensus statements?
Comprehensiveness	Do the answers cover the necessary breadth and depth?
Clarity and Understandability	Are the answers clearly written with information communicated in a way that is accessible to readers with varying levels of medical knowledge?
Relevance to Clinical Practice	How applicable are the answers to real-world clinical settings?
Patient Safety and Non-directiveness	Do the answers prioritize patient safety by avoiding any advice or information that could potentially lead to harm, and are they provided as options and information without suggesting a specific course of action (excluding emergency situations)?

5. Patient Safety and Non-directiveness: Focus on the safety of the information provided and ensuring that the responses did not direct or prescribe specific medical actions, except in scenarios requiring emergency care.

Repeatability protocol

To assess the consistency of ChatGPT-4's responses, a repeatability test was conducted. Two questions from each category were randomly selected and re-posed to ChatGPT-4 after an interval of 14 days, using a different thread for each question to ensure the independence of responses. This approach was designed to mimic real-world conditions where patients may ask similar questions over time. The responses were independently evaluated by the same panel of cardiologists using the previously defined metrics. This step was important to assess the stability of the AI's performance over time and to validate the reliability of its informing outputs.

Expert review and scoring

The ChatGPT-4 responses were reviewed by a panel of 3 senior cardiologists, all subspecialists in electrophysiology, each with over 10 years of experience. Given the high level of expertise among the reviewers and the objective nature of the evaluation criteria, the panel reviewed the responses collectively. Therefore, individual kappa scores for the inter-reviewer agreement were not calculated. Each response was rated on a scale of 0 (not adequate), 1 (partially adequate), and 2 (adequate) for each criterion given in Table 1. To assess the readability of ChatGPT-4's responses, the Flesch Reading Ease Score (FRES) was calculated for each answer provided by the model. This metric evaluates the ease with which a given text can be understood, based on sentence length and syllable count per word.[19] The FRES was calculated using the standard formula, where higher scores indicate easier readability, with a general scale ranging from 0 (very difficult) to 100 (very easy).

The study protocol was reviewed and approved by the hospital's ethics committee, ensuring compliance with the ethical standards stipulated in the Declaration of Helsinki. This study did not involve human participants directly; instead, it used anonymized patient queries to assess the AI's performance. Therefore, no patient's consent was required. The study's design, conduct, and dissemination were managed by the research team without patient or public involvement.

Statistical analysis

The distribution of total scores was assessed using the Shapiro-Wilk test to determine normality. Continuous variables are reported as mean ± SD when the distribution conformed to normality and as median with interquartile range when the distribution was non-normal. The Kruskal-Wallis test was used to compare the total scores across the 10 different categories, suitable for nonparametric analysis of >2 independent samples. Pairwise comparisons between categories were performed using the Mann-Whitney *U* test to identify specific pairs with significant differences. Categorical variables are presented as count and percentage. A predefined *P*-value threshold of <.05 was established to denote statistical significance. All statistical analyses were performed using Python (version 3.9) and its scientific libraries, including NumPy (version 1.21) was utilized for storage and manipulation of data (np.array), SciPy (version 1.9) for statistical testing, Pandas (version 1.4.4) for data manipulation, and Matplotlib (version 3.5; https://matplotlib.org/stable/#about-us) for data visualization. By structuring the study in this manner, we aimed to rigorously evaluate the potential of ChatGPT-4 as a tool for patient-focused informing, focusing on its ability to provide accurate, comprehensive, and understandable information tailored to the needs of individuals with AF.

Results

This observational cross-sectional study assessed the performance of ChatGPT-4 in answering 108 questions across 10 categories related to AF.

Category performance

The study revealed that responses categorized under "Lifestyle Adjustments" (category 7) scored the highest overall, achieving perfect scores of 2.0 in Comprehensiveness, Clarity, Relevance to Practice, and Patient Safety, leading to an overall perfect category score of 10.0. Close behind were the categories "Daily Life and Management" (category 8) and "Comorbidities and Complications" (category 6), with total scores of 9.9 and 9.6, respectively, indicating high reliability and usefulness of the responses in these areas. Evaluation scores across all criteria for 10 categories are given in Table 2.

Overall scores

The mean scores across all categories highlighted a relatively consistent performance with average values close to 2 for most criteria. The overall averages were as follows: Accuracy and Evidence Base: 1.87; Comprehensiveness: 1.79; Clarity and Understandability: 1.78; Relevance to Clinical Practice: 1.87; and Patient Safety and Non-directiveness: 1.94. The Kruskal-Wallis test revealed significant differences in total scores across the 10 categories (Kruskal-Wallis H = 19.644; *P* = .020), indicating variability in the correctness of ChatGPT-4's responses to different types of AF-related questions. Subsequent pairwise Mann-Whitney *U* tests were performed to identify which specific category comparisons contributed to this variability. Notably, category 1 (Basics of Atrial Fibrillation) significantly differed from category 10 (Miscellaneous Concerns) (*P* = .016). Categories 2 (Signs and Symptoms) and 7 (Lifestyle Adjustments) also showed significant differences (*P* = .004), as did categories 2 and 8 (Daily Life and Management) (*P* = .033). Another significant finding was between categories 7 (Lifestyle Adjustments) and 10 (Miscellaneous Concerns) (*P* < .001), indicating a substantial difference in the quality of responses provided.

Table 2 Evaluation scores of ChatGPT-4's responses

Category	Accuracy	Comprehensiveness	Clarity	Relevance	Patient safety	*Total*
1	1.91	1.91	1.91	2.00	2.00	*9.7*
2	1.67	1.56	1.56	1.78	1.78	*8.3*
3	1.90	1.80	1.80	1.90	2.00	*9.4*
4	1.76	1.71	1.71	1.65	1.82	*8.6*
5	1.73	1.82	1.64	1.91	1.90	*9.0*
6	2.00	1.87	1.80	1.93	2.0	*9.6*
7	2.00	2.00	2.00	2.00	2.00	*10.0*
8	2.00	1.90	2.00	2.00	2.00	*9.9*
9	2.00	1.71	1.71	1.86	2.00	*9.3*
10	1.60	1.20	1.40	1.60	1.80	*7.6*

Category 1 - Basics of Atrial Fibrillation; Category 2 - Signs and Symptoms; Category 3 -Diagnostic Tests; Category 4 - Treatment Modalities; Category 5 - Pulmonary Vein Isolation (PVI) Procedural and Post-Procedural Care; Category 6 - Comorbidities and Complications; Category 7 - Lifestyle Adjustments; Category 8 - Daily Life and Management; Category 9 - Stroke Prevention; Category 10 - Miscellaneous Concerns.

ChatGPT-4 = Generative Pre-trained Transformer 4.

One representative question from each category along with the generated responses and corresponding reviewer ratings is given in Online Supplemental Appendix 2.

In performing the repeatability protocol, the results consistently contained the same core information (following the criteria stated earlier) and medical terminology. However, the formulation of the responses varied, not affecting the correctness of the responses.

Readability assessment of ChatGPT-4's responses

The FRES for ChatGPT-4's responses ranged from 34.06 to 58.90, with a mean score of 41.85 ± 8.62. "Lifestyle Adjustments" and "Daily Life and Management" categories had the highest mean FRES (52.34 ± 6.21), while "Comorbidities and Complications" and "Miscellaneous Concerns" had the lowest (36.42 ± 4.95; $P < .05$). The Kruskal-Wallis test revealed significant differences across categories (statistic = 21.845; $P = .012$). Pairwise Mann-Whitney U tests confirmed higher readability in "Lifestyle Adjustments" than in "Comorbidities and Complications" and "Miscellaneous Concerns" ($P = .004$ and $P = .008$, respectively).

Discussion

This study offers one of the most comprehensive evaluations of ChatGPT-4's correctness in patient-focused informing and awareness for AF, covering a broad spectrum of queries from basic information to complex clinical questions. The study's main findings are as follows: (1) ChatGPT-4's responses varied significantly across different categories, with the highest scores in "Lifestyle Adjustments" and the lowest in "Miscellaneous Concerns." (2) The AI demonstrated stronger performance in categories with structured lifestyle-related queries than in those requiring complex medical knowledge. and (3) Statistical analysis confirmed significant differences in total scores across categories, indicating variability in the model's performance.

Categories that involve straightforward advice or guidelines (such as lifestyle adjustments) received higher scores, suggesting that ChatGPT-4 answers more correctly and reliably when dealing with direct, actionable information. In contrast, categories that likely required deeper understanding or integration of complex medical data (such as "Miscellaneous Concerns") scored lower, highlighting potential areas for improvement in the AI's architecture or training. Accuracy and patient safety were highest in "Comorbidities and Complications" (category 6), showcasing the model's correctness with complex intersecting health conditions. "Lifestyle Adjustments" (category 7) excelled in comprehensiveness and clarity, likely because of the structured nature of the questions. Surprisingly, "Basics of Atrial Fibrillation" (category 1) demonstrated strong relevance to clinical practice, indicating well-covered foundational knowledge.

Statistical analysis revealed significant differences across categories, similar to results from previous studies,[18,20–22] where AI's appropriateness in preventive cardiology and cardiac electrophysiology, as well as patient-centered informing and education, varied markedly. Our results suggest that while the model performs well in safety and relevance, there is room for improvement in the clarity and comprehensiveness of the information provided. For instance, in questions related to AF and sudden cardiac death, ChatGPT-4 did not make the crucial connection between AF in the setting of ventricular preexcitation, which is a significant risk factor for sudden cardiac death, especially in younger patients.

Furthermore, the study done by Azizi et al[23] and Kassar et al[24] documented variability in response quality between patient-oriented and clinician-oriented prompts in managing AF, especially those regarding AF phenotype. Their study underscored that while patient-directed responses often met appropriateness criteria, with ChatGPT-4 scoring 100% similar to that of our study, clinician-directed responses frequently lacked either comprehensiveness or accuracy.[23,24] Similarly, our research found that responses to complex medical questions, particularly those requiring detailed clinical judgment, were inconsistent, echoing the challenges noted by Azizi et al[23] in eliciting high-level scientific responses from AI. In our research, questions related to lifestyle achieved a better overall score, with almost perfect scores, compared with Azizi et al, where a smaller overall set of questions achieved only 71.4%. Moreover, recent studies suggest that while AI can generate generally reliable and understandable content for patient-focused informing, significant gaps remain in covering all necessary clinical aspects, with a notable portion of responses missing key information, especially earlier forms of AI.[18,19,24] This aligns with our observation that while ChatGPT-4 can effectively communicate direct, actionable advice, it struggles with the nuances of more complex medical queries, such as those involving AF triggers or detailed treatment mechanisms.

This study revealed significant variability in the readability of ChatGPT-4's responses, with FRES highlighting that simpler lifestyle-related queries ("Lifestyle Adjustments" and "Daily Life and Management") had the highest readability while more complex topics ("Comorbidities and Complications" and "Miscellaneous Concerns") were harder to understand, similar to the results of Kassar et al[24] and Siddiqui et al.[25] The overall mean FRES of 41.85 ± 8.62 suggests that much of the content is difficult to read, potentially limiting accessibility for patients with lower health literacy. Statistical analysis confirmed significant differences in readability across categories ($P = .012$), with pairwise comparisons showing that responses to "Lifestyle Adjustments" were significantly easier to understand than those related to "Comorbidities and Complications" ($P = .004$) and "Miscellaneous Concerns" ($P = .008$). This reflects ChatGPT-4's strengths in delivering straightforward information but also underscores its limitations in presenting complex medical knowledge in a patient-friendly manner. The findings suggest that while ChatGPT-4 shows promise as a tool for patient-focused informing in AF, further refinement is needed to ensure accessibility across literacy levels, with clinician oversight remaining essential for managing complex conditions. Comparatively, the use of ChatGPT-4 in patient-focused education for AF must be approached with caution. As seen in other studies, AI performance in medical diagnostic tasks, such as interpreting ECGs or imaging, significantly benefits from the addition of contextual information, such as clinical symptoms.[11,16] This suggests that integrating comprehensive patient data and contextual information could enhance the AI's performance in patient-focused informing. Implementation could be facilitated by specialized software that incorporates elements of the patient's medical history and serves as an additional layer between the patient and ChatGPT-4.[26] While integrating patient-specific medical history into LLMs such as ChatGPT-4 may enhance performance, it raises concerns regarding patient confidentiality. To mitigate risks, such implementations could be restricted to secure local hospital networks, ensuring data protection. In addition, it should be noted that there is a pressing need for regulatory frameworks to govern the use of LLMs in health care, ensuring compliance with privacy laws and ethical standards. Finally, in this context, the concept of patient-centered care, introduced by the American Institute of Medicine to improve the quality of health care services in the 21st century, could potentially be approached with ChatGPT-4 and primarily gravitates around understanding, engagement, and empowerment in managing AF.[25]

In light of these observations, future research should focus on refining the model's training processes to handle complex medical inquiries more uniformly and ensure that the information provided is accurate, comprehensive, and clear. Further studies could explore integrating AI tools such as ChatGPT-4 with other patient interaction platforms to provide a holistic approach to patient-focused informing and care.

Limitations

The interpretation of the results from the present study is subject to several limitations that must be considered. First, the questions addressed to ChatGPT-4 were predefined and may not encompass the full spectrum of queries typical to patients with AF, potentially limiting the generalizability of the findings to real-world settings. The selection of questions could influence the model's demonstrated correctness by focusing on areas where AI might inherently perform better because of the structured nature of the information.

Second, the study used a relatively small sample size for each category, which may not sufficiently capture the variability in patient questions or the complexity of real-life patient interactions. Even larger data sets might provide a more robust analysis of the AI's capabilities and allow a more nuanced understanding of its performance across a diverse array of patients' inquiries. Third, the analysis was based on numerical scoring of responses without qualitative feedback from patients who might interact with this technology in a practical setting. The lack of empirical data on user satisfaction or the informational impact of the AI's responses limits our understanding of its practical utility and acceptance in clinical or educational environments. Fourth, the study did not account in full for the dynamic nature of AI learning capabilities. ChatGPT-4, like many AI models, periodically updates its training and algorithms on the basis of new data, which could alter its performance over time. The results presented here reflect the model's capabilities at a specific point in time and may not represent its future or past performance accurately. Last, while the study provides valuable insights into the potential of AI in patient-focused informing, it does not address significant ethical considerations such as the implications of AI errors in patient understanding or decision making.

Conclusion

This study represents one of the most comprehensive evaluations of ChatGPT-4's correctness in patient-focused informing for AF. The findings suggest that while ChatGPT-4 can provide informative and relevant answers across different categories, its performance tends to vary. The use of AI for patient-focused informing is a promising development in health care, providing an accessible and scalable means to supplement traditional patient-provider interactions. While ChatGPT-4 demonstrates considerable potential in delivering patient-focused information for AF, strategic enhancements and rigorous validation are essential to optimize its utility and ensure it can reliably support patient needs across a broader spectrum of medical inquiries.

Funding Sources: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosures: The authors have no conflicts of interest to disclose.

Authorship: All authors attest they meet the current ICMJE criteria for authorship.

Patient Consent: This study did not involve human participants directly; instead, it used anonymized patient queries to assess the AI's performance. Therefore, no patient's consent was required.

Ethics Statement: The study protocol was reviewed and approved by the hospital's ethics committee, ensuring compliance with the ethical standards stipulated in the Declaration of Helsinki.

References

1. Kornej J, Börschel CS, Benjamin EJ, Schnabel RB. Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights. Circ Res 2020; 127:4–20.
2. Lippi G, Sanchis-Gomar F, Cervellin G. Global epidemiology of atrial fibrillation: an increasing epidemic and public health challenge. Int J Stroke 2021;16:217–221.
3. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172–180.
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31–38.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930–1940.
6. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns (N Y) 2024;5:100943.
7. Pavlov M, Barić D, Novak A, Manola Š, Jurin I. From statistical inference to machine learning: a paradigm shift in contemporary cardiovascular pharmacotherapy. Br J Clin Pharmacol 2024;90:691–699.
8. Nov O, Singh N, Mann D. Putting ChatGPT's Medical Advice to the (Turing) Test: survey study. JMIR Med Educ 2023;9:e46939.
9. Moons P, Van Bulck L. ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. Eur J Cardiovasc Nurs 2023;22:e55–e59.
10. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198.
11. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13:16492.
12. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? Resuscitation 2023;185:109732.
13. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3:100324.
14. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770.
15. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023;13:20512.
16. Lisicic A, Serman A, Jordan A, et al. Does ChatGPT-4 succeed in the ECG interpretation: friend or foe to cardiologists? Europace 2024;26:euae102.655.
17. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023;20:998–1003.
18. 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 2024;26:euad369.
19. Flesch R. A new readability yardstick. J Appl Psychol 1948;32:221–233.
20. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023; 329:842–844.
21. Kassab J, Hadi El Hajjar A, Wardrop RM III, Brateanu A. Accuracy of online artificial intelligence models in primary care settings. Am J Prev Med 2024; 66:1054–1059.
22. Kerbage A, Kassab J, El Dahdah J, Burke CA, Achkar JP, Rouphael C. Accuracy of ChatGPT in common gastrointestinal diseases: impact for patients and providers. Clin Gastroenterol Hepatol 2024;22:1323–1325.
23. Azizi Z, Alipour P, Gomez S, et al. Evaluating recommendations about atrial fibrillation for patients and clinicians obtained from chat-based artificial intelligence algorithms. Circ Arrhythm Electrophysiol 2023;16:415–417.
24. Kassar A, Macheret F, Kondamudi N, et al. Performance of large language models as a resource for patients and healthcare professionals on atrial fibrillation. Heart Rhythm 2024;21:2048–2050.
25. Siddiqui E, Shah AM, Sambol J, Waller AH. Readability assessment of online patient education materials on atrial fibrillation. Cureus 2020;12:e10397.
26. Novak A, Rode F, Lisičić A, et al. The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology. medRxiv published online ahead of print January 30, 2024; https://doi.org/10.1101/2023.08.08.23293689.