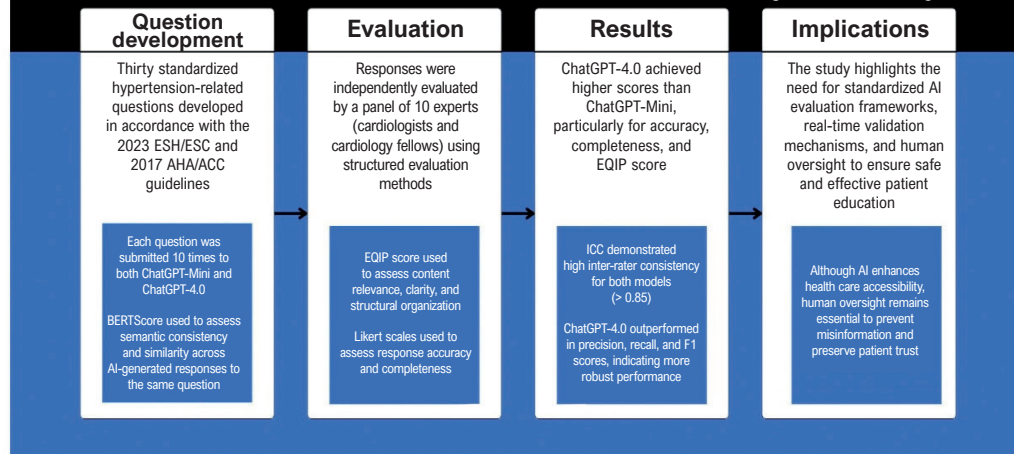


Central Illustration: Assessment of ChatGPT-4.0 versus ChatGPT-Mini in Generating Guideline-Based Hypertension Content



Can large language models accurately provide health guidance to patients with hypertension?

Aim: To compare ChatGPT-Mini and ChatGPT-4.0 in their ability to generate hypertension-related educational content aligned with clinical guidelines



Arq Bras Cardiol. 2026; 123(2):e20250498

CC: American College of Cardiology; AHA: American Heart Association; AI: artificial intelligence; EQUIP: Ensuring Quality Information for Patients; ESC: European Society of Cardiology; ESH: European Society of Hypertension; ICC: intraclass correlation coefficient.

Introduction

The use of artificial intelligence (AI) in hypertension management offers substantial potential.^{1,2} However, current evidence indicates that its use remains insufficiently systematized, particularly regarding evaluation criteria and patient-centered approaches.^{3,4} For example, AI-driven tools such as chatbots have demonstrated utility in supporting hypertension self-management, yet existing studies consistently highlight the need for more robust evaluation frameworks and clearer, more consistent guidance for patients.⁵

Almagazzachi et al. investigated the use of generative AI for patient education in hypertension and reported that, despite providing meaningful support, these tools show inconsistencies in alignment with established clinical guidelines, including those of the European Society of Hypertension (ESH).⁶ In addition, variability in chatbot performance across different languages has been documented, further emphasizing the necessity for standardized methodologies to ensure accuracy, reliability, and guideline adherence.⁷ Adopting a user-centered approach in the development of AI solutions for hypertension management may enhance systematization and improve patient outcomes.⁵ Therefore, although AI demonstrates considerable promise in this field, a more structured and patient-focused framework is required to fully realize its potential in hypertension care.⁸

Several studies have attempted to systematize the evaluation of AI-generated educational information for

patients.^{9,10} Nevertheless, these efforts have not incorporated a systematic process for developing evaluation questions based on expert consensus.¹¹

The present study aims to assess the performance of two ChatGPT model versions, ChatGPT-Mini and ChatGPT-4.0, in generating hypertension-related medical content. Validated instruments were used to evaluate the quality of patient information and to examine the strengths and limitations of AI in medical education and patient care.^{11,12} Furthermore, by applying a structured and standardized evaluation of AI-generated responses, this study seeks to identify areas in which AI-produced content may be deficient, thereby providing relevant insights to inform future AI development in health care.¹³ This systematic approach enables an assessment of how effectively AI models deliver accurate, patient-centered educational materials and supports the refinement of AI-based strategies for medical information dissemination.

Methods

Study design

This study employed a survey-based design (Figure 1). Questions addressing the diagnosis, treatment, and prognosis of hypertension were submitted to ChatGPT-Mini and ChatGPT-4.0 (Supplementary Table 1). The generated responses were compared with recommendations from the 2023 ESH guidelines and the 2017 American College of Cardiology/American Heart Association (ACC/AHA)

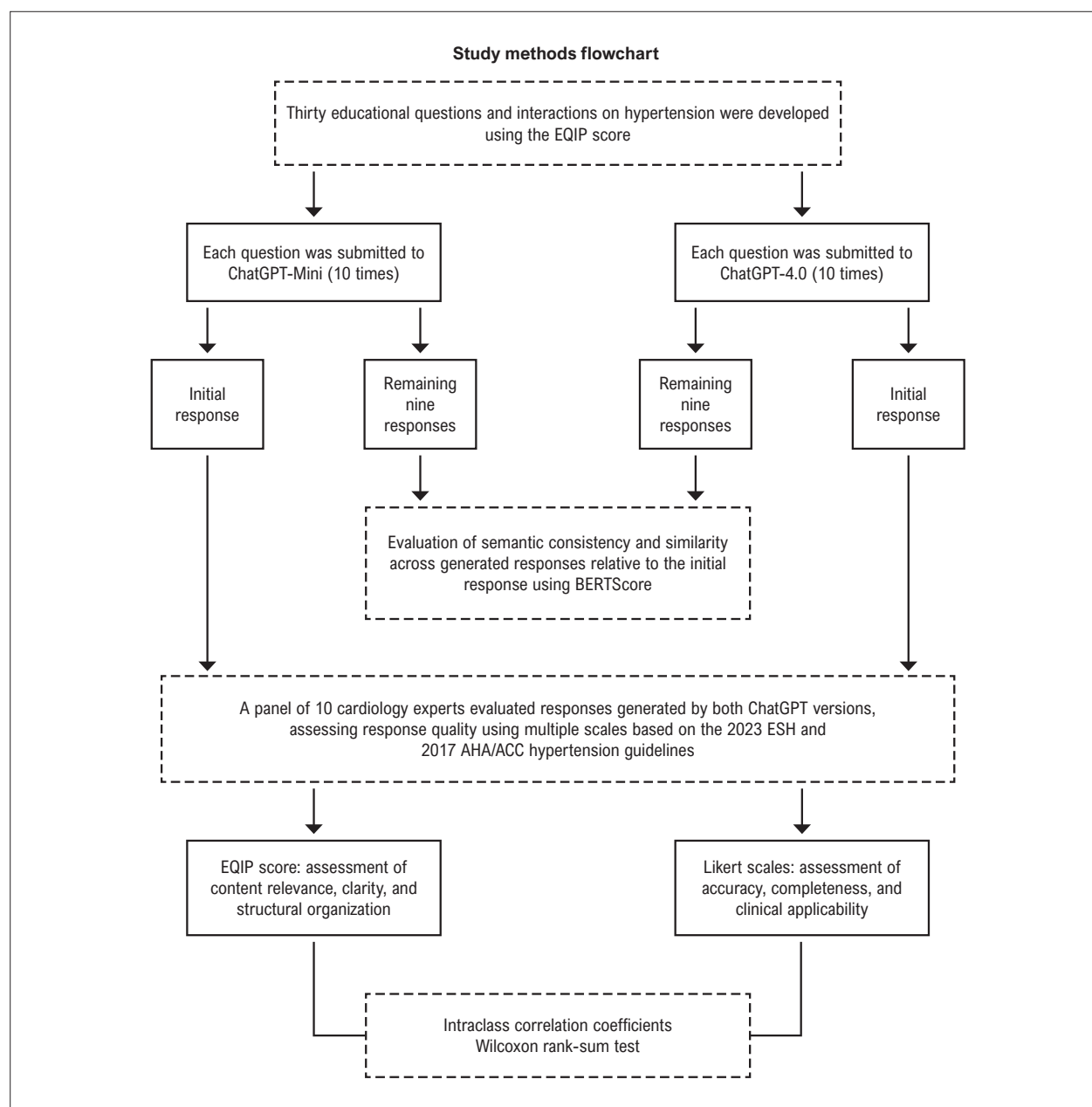


Figure 1 – Study flowchart for the evaluation of artificial intelligence-generated hypertension content. ACC: American College of Cardiology; AHA: American Heart Association; EQIP: Ensuring Quality Information for Patients; ESH: European Society of Hypertension.

hypertension guidelines. All responses were evaluated by independent cardiologists and cardiology fellows.

Data collection

The research team developed 31 questions related to hypertension (Table 1; see Supplementary Material 1 for additional details). The content and structure of each question were guided by the Ensuring Quality Information for Patients (EQIP) score.¹⁴ The EQIP score is a validated tool designed to assess the quality of information provided to patients in health

care settings, with a focus on accuracy, reliability, and usefulness in supporting informed health-related decision-making.

Each question was submitted 10 times in English to both ChatGPT versions between July 25 and July 30, 2024. All responses were systematically recorded by initiating new conversations using unconditional prompts in temporary chat mode. In cases in which the response format did not adequately address the original question, the prompt “Be specific and incorporate medical knowledge about hypertension from the guideline of hypertension from AHA/ACC 2017 and ESH 2023” was appended to the conversation as needed.

Expert rating of GPT responses

A total of 10 physicians (five cardiologists and five cardiology fellows) from multiple institutions in Brazil and the United States independently evaluated the responses generated by ChatGPT (Figure 2). Evaluators received written instructions detailing how to interpret the responses, apply the evaluation criteria, and compare the content with the 2017 ACC/AHA and 2023 ESH hypertension guidelines. Although evaluators were permitted to consult the guidelines, they were encouraged to rely on their clinical expertise when assessing response quality.

All evaluators were blinded to the source of the responses and were unaware of whether the content had been generated by ChatGPT-Mini or ChatGPT-4.0. Ratings were collected using surveys created in REDCap (Supplementary Material 1).

Measures

Based on guideline-derived information and clinical knowledge, evaluators classified each response using three components: the EQIP score to assess overall information quality and two Likert-type scales to independently assess accuracy and completeness (Supplementary Material 1).

The EQIP score was adapted for the evaluation of responses generated by ChatGPT. It consists of criteria covering three domains: content (assessing accuracy, relevance, and comprehensiveness), identification (evaluating clarity and ease of identifying information), and data structure (examining

organization, readability, and presentation of information). The modified EQIP instrument allows a maximum score of 31 points, distributed across the content (maximum 16 points), identification (maximum 5 points), and structural data (maximum 10 points) domains.

Each EQIP criterion was scored using a three-point system: 1 (correct response), 0.5 (partially correct response), or 0 (incorrect response). The total EQIP score was calculated as the sum of points across all criteria, with higher scores indicating higher-quality information. For all responses, one point was automatically awarded for the item "Can you date the above-named information?", as this element could potentially reveal the version of ChatGPT being evaluated.

Accuracy was assessed using a five-point scale: completely incorrect (no correct information and contradictory to hypertension guidelines); more incorrect than correct (predominantly incorrect and largely inconsistent with guidelines); approximately correct and incorrect (a similar proportion of correct and incorrect information, partially aligned with guidelines); more correct than incorrect (predominantly correct and generally aligned with guidelines); and completely correct (fully accurate, comprehensive, and entirely consistent with hypertension guidelines).

Completeness was evaluated using a three-point scale reflecting how thoroughly each response addressed the question: incomplete (key aspects missing or insufficiently addressed);

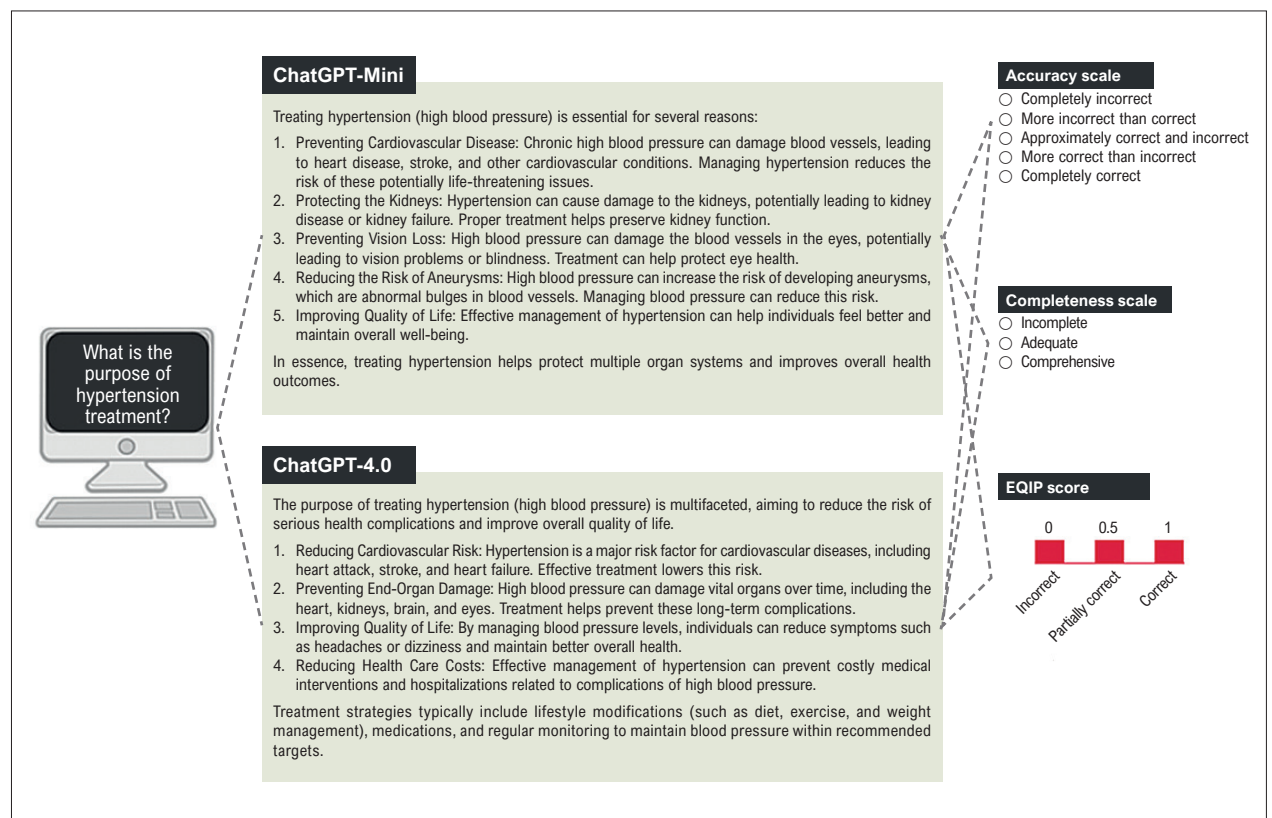


Figure 2 – Example of the evaluation of artificial intelligence-generated hypertension responses. EQIP: Ensuring Quality Information for Patients.

Original Article

adequate (all requested aspects addressed with the minimum information required); and comprehensive (all aspects addressed with additional relevant context beyond expectations).

Response consistency across the 10 repetitions of each question was assessed using BERTScore, which quantifies semantic similarity based on precision, recall, and F1 scores. Results were presented graphically to compare content similarity across responses.¹⁵

ChatGPT sampling

The minimum number of raters required to achieve an intraclass correlation coefficient (ICC) of 0.8, with a precision of ± 0.1 and a 95% CI, when evaluating the 31 items generated by each ChatGPT version, was estimated to be 10.

Statistical analysis

The distribution of continuous variables was assessed using the Shapiro–Wilk test. As all continuous variables exhibited non-normal distributions, data are presented as medians and interquartile ranges (IQRs). Comparisons between ChatGPT-Mini and ChatGPT-4.0 were performed using the Wilcoxon rank-sum test (two-sided, $p < 0.05$).

Effect sizes were reported with 95% CIs, including Hodges–Lehmann median differences to estimate location shifts and Cliff’s delta (δ) to assess stochastic dominance. 95% CIs were derived using Walsh averages or nonparametric bootstrap methods, as appropriate. Inter-rater agreement among evaluators was quantified using ICC (two-way random effects model, absolute agreement), with 95% CIs.

Comparisons between ChatGPT-4.0 and ChatGPT-Mini across accuracy, completeness, and EQIP domains were conducted using Cliff’s delta, with δ values greater than 0 indicating superiority of ChatGPT-4.0. All estimates are reported with 95% CIs, and statistical significance was defined as $p < 0.05$. Statistical analyses were performed using R (version 4.2.2). BERTScore analyses were conducted in Google Colab using the *bert-score* package.

Results

Study sample and data inclusion

All responses generated by ChatGPT-Mini and ChatGPT-4.0 were included in the analysis, with no exclusions or missing data. The standardized set of 31 questions was applied uniformly to both models. Across the 10 repetitions per question, no instances of AI hallucination were identified. In addition, no follow-up prompts were required to clarify responses, as all outputs adequately addressed the original questions.

Accuracy and completeness

As shown in Table 2, both models achieved high scores for accuracy and completeness. Median values were numerically higher for ChatGPT-4.0; however, no statistically significant differences were observed between the two models for either outcome.

Overall quality (EQIP scores)

The EQIP score, which integrates content, identification, and structural domains, indicated broadly comparable overall quality

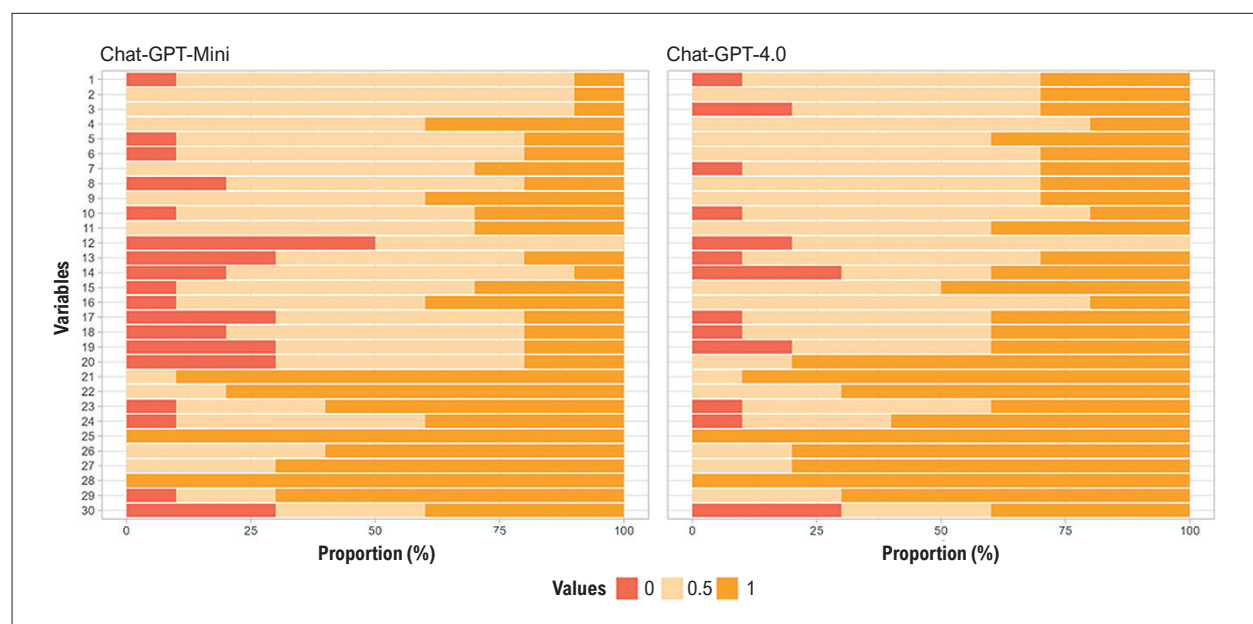


Figure 3 – Modified EQIP score–based comparison of ChatGPT responses across hypertension questions. Responses generated by ChatGPT-Mini and ChatGPT-4.0 across 30 hypertension-related questions were evaluated using a modified EQIP score. Responses were classified as 1 (correct), 0.5 (partially correct), or 0 (incorrect). The x-axis represents the proportion of responses in each category, and the y-axis lists the evaluated questions. EQIP: Ensuring Quality Information for Patients.

between the two models. As presented in Table 2, ChatGPT-4.0 tended to achieve numerically higher total EQIP scores as well as higher content and identification subscores, compared with ChatGPT-Mini. Structural scores were very similar between models, resulting in comparable overall EQIP totals.

Effect-size analyses using Cliff's δ favored ChatGPT-4.0 across all EQIP domains; however, all corresponding 95% CIs included zero (Table 3), indicating that these differences were not statistically robust. Figure 3 illustrates the distribution of EQIP scores across domains for each model, demonstrating generally favorable evaluations by raters for content, identification, and structure. In this figure, each horizontal line corresponds to an individual EQIP item: items 1-16 represent content, items 17-20 represent identification, and items 21-30 represent structure. Across all items, both models received consistently favorable and largely comparable ratings.

BERTScore evaluations

BERTScore analyses revealed differences between models in precision, recall, and response consistency. ChatGPT-Mini consistently demonstrated higher F1 scores, frequently exceeding 0.92, whereas ChatGPT-4.0 exhibited F1 scores ranging from 0.8851 to 0.9316 (see Supplementary Material 2).

For precision, ChatGPT-Mini achieved a score of 0.9356 compared with 0.8643 for ChatGPT-4.0 in response to a question regarding the definition of hypertension. Similarly, for recall, ChatGPT-Mini scored 0.9558, whereas ChatGPT-4.0 scored 0.9330 in a related follow-up question.

Overall, ChatGPT-Mini showed greater consistency in response stability, with F1 scores rarely falling below 0.90. In contrast, ChatGPT-4.0 demonstrated greater variability, including an F1 score of 0.8757 in response to a question addressing non-pharmacological treatments for hypertension.

Inter-rater reliability

Inter-rater reliability was consistently high across all evaluated metrics, with ICCs exceeding 0.80 for both models.

For EQIP scores, ICCs were 0.844 (95% CI, 0.677-0.952) for ChatGPT-Mini and 0.840 (95% CI, 0.665-0.951) for ChatGPT-4.0. For accuracy scores, ICCs were 0.742 (95% CI, 0.456-0.922) for ChatGPT-Mini and 0.688 (95% CI, 0.339-0.906) for ChatGPT-4.0. For completeness scores, ICCs were identical for both models at 0.824 (95% CI, 0.627-0.947 for ChatGPT-Mini and 0.628-0.947 for ChatGPT-4.0).

These ICC values indicate moderate to strong inter-rater agreement, supporting the reliability of the evaluation methodology used in this study. Overall, the findings suggest superior performance of ChatGPT-4.0 across key quality metrics (Central Illustration).

Discussion

As AI becomes increasingly integrated into patient education, persistent concerns remain regarding the accuracy, completeness, and consistency of AI-generated medical information. Although AI holds substantial potential to improve health care accessibility, its capacity to deliver reliable, guideline-

adherent content remains a critical challenge, particularly in the management of chronic conditions such as hypertension. In this study, we systematically evaluated the performance of ChatGPT-Mini and ChatGPT-4.0 in generating patient-centered hypertension education, with a focus on guideline alignment, informational quality, and response consistency.

Our findings indicate that ChatGPT-4.0 demonstrated incremental improvements over ChatGPT-Mini in accuracy, completeness, and response structure. However, these differences were modest and did not reach statistical significance, likely reflecting small effect sizes. ChatGPT-4.0 achieved a slightly higher mean accuracy score than ChatGPT-Mini (4.02 ± 0.35 vs 3.86 ± 0.45), along with a higher median completeness score (1.26 vs 1.10). Despite these numerical gains, both models showed notable limitations in content completeness. These observations are consistent with prior studies reporting that newer AI models improve linguistic coherence and structural organization¹¹ but may still lack sufficient depth and specificity when addressing complex clinical recommendations.^{16,17}

To evaluate the quality of AI-generated medical content systematically, we applied established assessment instruments, including the EQIP score and Likert-type scales. The EQIP score, originally developed to assess written patient education materials,¹⁸ provided a structured framework for evaluating clarity, relevance, and reliability in AI-generated responses.¹⁴ Although EQIP has not been formally validated for AI-generated content, its use in this study highlights the growing need for standardized evaluation frameworks tailored to AI-driven health care communication. Likert scales further supported the assessment of perceived accuracy and completeness, capturing professional judgments regarding the trustworthiness of AI-generated information.¹²

ChatGPT-4.0 achieved a slightly higher overall EQIP score than ChatGPT-Mini (21.4 ± 3.7 vs 19.5 ± 4.0), suggesting modest improvements in the quality of hypertension-related educational content. The most pronounced differences were observed in the content domain (9.15 ± 2.32 vs 8.35 ± 2.32) and the identification domain (3.80 ± 1.14 vs 2.85 ± 1.08), reflecting improved comprehensiveness and transparency. Nevertheless, these differences were not statistically significant, indicating that enhancements in presentation and structure may not consistently translate into more reliable or guideline-adherent content.¹⁹

When examining the proportion of fully correct responses, ChatGPT-4.0 outperformed ChatGPT-Mini, producing correct answers (score = 1) in approximately 70%-80% of evaluated questions, compared with fewer than 50% for ChatGPT-Mini. Despite this improvement, adherence to established hypertension guidelines, including the 2017 ACC/AHA and 2023 ESH recommendations, remained inconsistent. Although both models generally aligned with core guideline principles, responses frequently lacked clinical specificity or deviated from evidence-based standards. This pattern mirrors findings from broader evaluations of AI-generated patient education, where reported accuracy rates range from 71.7% to 94.3%, yet deficits in depth and clinical precision remain common.²⁰ ChatGPT-Mini demonstrated a higher proportion of incorrect or partially correct responses, particularly for questions requiring detailed clinical explanations. ChatGPT-4.0 reduced the frequency of incomplete responses to below 20% for most variables; however, 10%-30%

Original Article

of responses still fell into lower-quality categories, underscoring persistent informational gaps.

These findings are consistent with a previous study evaluating ChatGPT-4.0 across five hepato-pancreatico-biliary conditions, which reported a median EQIP score of 16 (IQR, 14.5-18), indicating substantial deficiencies in citation accuracy and structural coherence.¹⁰ While our study observed higher EQIP scores for hypertension-related content, both investigations highlight the continued limitations of AI models in fully aligning with evidence-based medical recommendations. Importantly, although newer models may improve readability and organization, their factual accuracy does not appear to have advanced proportionally.²¹ This reinforces the need for rigorous validation processes and ongoing human oversight to ensure that improvements in clarity do not compromise adherence to clinical guidelines.^{20,22}

Despite these limitations, ChatGPT-4.0 demonstrates potential as a supplementary tool for patient education in hypertension management. Its capacity to rapidly generate guideline-informed explanations may reduce the informational burden on health professionals, particularly for initial discussions of treatment strategies and lifestyle interventions.^{19,20} Prior research suggests that better-informed patients are more likely to adhere to prescribed therapies, potentially improving outcomes in chronic conditions such as hypertension.^{23,24} Nevertheless, AI-generated content alone may not yet provide sufficiently comprehensive or actionable guidance, especially given ongoing public skepticism toward AI applications in health care.²⁵⁻²⁷

A critical concern in AI-generated medical content is the phenomenon of "hallucinations," in which models produce plausible but incorrect information. Notably, no hallucinations were identified in either model across the 10 repeated interactions in this study, which suggests reliable performance within this specific clinical domain. However, previous studies indicate that hallucinations remain a broader risk in health care AI, particularly for complex or less well-defined topics.^{28,29} This risk, combined with known sensitivity to prompt design, underscores the continued importance of clinician oversight in clinical and educational settings.³⁰

AI is also increasingly influencing the development of clinical guidelines and patient communication by facilitating the generation of clinical questions and supporting real-time interactions. AI-driven systems have shown promise in prioritizing clinical recommendations and streamlining information delivery.^{31,32} Nonetheless, inconsistent adherence to clinical standards remains a major limitation, necessitating ongoing refinement of these technologies.³³

Response consistency was further evaluated using BERTScore metrics, which quantify semantic similarity across repeated responses. ChatGPT-Mini demonstrated greater response stability, achieving higher precision (0.9356 vs 0.8643) and recall (0.9558 vs 0.9330) than ChatGPT-4.0. In contrast, ChatGPT-4.0 exhibited greater variability, with F1 scores ranging from 0.8851 to 0.9316. These findings suggest that ChatGPT-Mini may be better suited for structured patient education materials requiring high consistency, whereas ChatGPT-4.0's variability may enhance engagement in more dynamic, conversational contexts. Similar patterns have been reported in prior studies, where newer models prioritize flexibility at the expense of strict consistency.^{34,35}

Finally, inter-rater reliability was high across all evaluated domains, with ICC values exceeding 0.80 for EQIP, accuracy, and completeness scores. This strong agreement among evaluators supports the robustness of the assessment methodology and indicates that observed differences reflect genuine variations in model performance rather than rater subjectivity.

In summary, this study contributes to the growing literature on the role of AI in health care communication and patient education. While advances in AI models have improved clarity and structural organization, ensuring accuracy, consistency, and guideline adherence remains essential for their safe integration into clinical practice.^{26,36} Our findings highlight the urgent need for standardized, rigorous evaluation frameworks for AI-generated medical content. Emerging approaches, such as structured assessment models including the Conversational Reasoning Assessment Framework for Testing in Medicine, may offer valuable tools for systematically evaluating AI performance in health care contexts.³⁷ Future research should focus on real-time validation mechanisms that dynamically link AI-generated responses to up-to-date clinical guidelines and promote interdisciplinary collaboration among clinicians, AI developers, and regulatory authorities to establish robust ethical, safety, and quality standards for AI use in health care.²⁵

Study limitations

Although this study provides relevant insights into the performance of ChatGPT-Mini and ChatGPT-4.0 in patient education, several limitations should be acknowledged. First, the analysis was restricted to a single clinical domain, hypertension, which limits the generalizability of the findings to other medical conditions. Because AI performance may vary across specialties, future investigations should evaluate AI-generated content across a broader range of clinical topics.

Second, despite the use of structured evaluation tools such as EQIP and BERTScore, the assessment relied on expert human judgment, which is inherently subjective, even with standardized rating criteria. Variability in expert interpretation may have influenced scoring, underscoring the need for complementary automated and consensus-based validation approaches in future research.

Third, the evaluation was limited to English-language outputs, thereby restricting applicability to non-English-speaking populations. As prior studies have demonstrated variability in AI performance across languages, further research is warranted to assess linguistic differences in AI-generated medical content and to ensure equitable access to accurate patient education globally.

In addition, the AI models were assessed in a controlled, static testing environment and were not exposed to real-world patient interactions or iterative feedback. AI-generated medical content is inherently dynamic, influenced by model updates and evolving clinical guidelines. Future studies should incorporate real-time validation strategies that continuously benchmark AI outputs against current clinical standards to maintain accuracy and relevance.

Finally, this study focused exclusively on ChatGPT-Mini and ChatGPT-4.0. Other large language models with distinct architectures and training paradigms (e.g., Claude, Gemini,

and Med-PaLM) may demonstrate different performance characteristics. Comparative evaluations involving multiple AI systems are therefore necessary to more comprehensively define the role of AI in patient education.

Conclusion

This study demonstrates that although ChatGPT-4.0 offers improved structural coherence compared with ChatGPT-Mini, it does not produce statistically significant gains in factual completeness or response consistency. These findings highlight the importance of implementing standardized evaluation frameworks, real-time validation mechanisms, and continued human oversight to support the safe and effective use of AI in patient education.

Future research should emphasize dynamic validation strategies and the development of universally accepted evaluation standards to ensure that medical AI systems remain clinically reliable and ethically responsible. While AI technologies have the potential to enhance health care accessibility, human supervision remains essential to prevent misinformation and preserve patient trust. Ultimately, the responsible integration of AI into medical practice will depend on sustained interdisciplinary collaboration to uphold rigorous standards of accuracy, safety, and ethical integrity.

Author Contributions

Conception and design of the research: Ataídes RJC, Campos MAG, Vissoci JRN, Renato Delascio Lopes RD; Acquisition of data: Ataídes RJC, Campos MAG; Analysis and interpretation of the data and Statistical analysis: Ataídes RJC, Campos MAG, Souza JVP; Writing of the manuscript: Ataídes RJC, Campos MAG, Rocha RC, Lacalle AA, Vieira CB, Artioli T, Medeiros TC, Souza Filho EM, Gismondi R, Campana EMG, Romeo FJ, Razuk C, Vissoci JRN, Renato Delascio Lopes

RD; Critical revision of the manuscript for content: Ataídes RJC, Campos MAG, Souza JVP, Rocha RC, Lacalle AA, Vieira CB, Artioli T, Medeiros TC, Souza Filho EM, Gismondi R, Campana EMG, Romeo FJ, Razuk C, Vissoci JRN, Renato Delascio Lopes RD.

Potential conflict of interest

No potential conflict of interest relevant to this article was reported.

Sources of funding

There were no external funding sources for this study.

Study association

This study is not associated with any thesis or dissertation work.

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Use of Artificial Intelligence

During the preparation of this work, the author(s) used ChatGPT and Grammarly for refining commands and correcting grammar, punctuation. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability Statement

The underlying content of the research text is contained within the manuscript.

References

1. Gala D, Behl H, Shah M, Makaryus AN. The Role of Artificial Intelligence in Improving Patient Outcomes and Future of Healthcare Delivery in Cardiology: A Narrative Review of the Literature. *Healthcare*. 2024;12(4):481. doi: 10.3390/healthcare12040481.
2. Visco V, Izzo C, Mancusi C, Rispoli A, Tedeschi M, Virtuoso N, et al. Artificial Intelligence in Hypertension Management: An Ace up Your Sleeve. *J Cardiovasc Dev Dis*. 2023;10(2):74. doi: 10.3390/jcdd10020074.
3. Vinufrancis A, Al Hussein H, Patel HV, Nizami A, Singh A, Nunez B, et al. Assessing the Quality and Reliability of AI-Generated Responses to Common Hypertension Queries. *Cureus*. 2024;16(8):e66041. doi: 10.7759/cureus.66041.
4. Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, et al. Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis with Human Experts. *JMIR Med Educ*. 2024;10:e51282. doi: 10.2196/51282.
5. Griffin AC, Khairat S, Bailey SC, Chung AE. A Chatbot for Hypertension Self-Management Support: User-Centered Design, Development, and Usability Testing. *JAMIA Open*. 2023;6(3):o0ad073. doi: 10.1093/jamiaopen/o0ad073.
6. Almagazzachi A, Mustafa A, Sedeh AE, Gonzalez AEV, Polianovskaia A, Abood M, et al. Generative Artificial Intelligence in Patient Education: ChatGPT Takes on Hypertension Questions. *Cureus*. 2024;16(2):e53441. doi: 10.7759/cureus.53441.
7. Ioannidis A, Tsounis D, Bouras G, Komninou D, Pechlevanis A, Zalokosta E, et al. Accuracy of Artificial Intelligence Chatbots' Replies in Greek vs. English and in Accordance with the 2023 ESH Guidelines for the Management of Arterial Hypertension. *J Hypertens*. 2024;42(Suppl 1):e3. doi: 10.1093/eurjpc/zwae175.101.
8. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. *Cureus*. 2023;15(11):e49252. doi: 10.7759/cureus.49252.
9. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and Accuracy of Artificial Intelligence ChatGPT in Providing Information on Ophthalmic Diseases and Management to Patients. *Eye*. 2024;38(7):1368-1373. doi: 10.1038/s41433-023-02906-0.
10. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res*. 2023;25:e47479. doi: 10.2196/47479.
11. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw Open*. 2023;6(10):e2336483. doi: 10.1001/jamanetworkopen.2023.36483.

Original Article

12. Xu X, Chen Y, Miao J. Opportunities, Challenges, and Future Directions of Large Language Models, Including ChatGPT in Medical Education: A Systematic Scoping Review. *J Educ Eval Health Prof.* 2024;21:6. doi: 10.3352/jeehp.2024.21.6.
13. Gritti MN, Alturki H, Farid P, Morgan CT. Progression of an Artificial Intelligence Chatbot (ChatGPT) for Pediatric Cardiology Educational Knowledge Assessment. *Pediatr Cardiol.* 2024;45(2):309-13. doi: 10.1007/s00246-023-03385-6.
14. Moul B, Franck LS, Brady H. Ensuring Quality Information for Patients: Development and Preliminary Validation of a New Instrument to Improve the Quality of Written Health Care Information. *Health Expect.* 2004;7(2):165-75. doi: 10.1111/j.1369-7625.2004.00273.x.
15. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT [Internet]. New York: arXiv; 2020 [cited 2025 Feb 08]. Available from: <http://arxiv.org/abs/1904.09675>.
16. Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, et al. Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. *J Med Internet Res.* 2024;26:e55939. doi: 10.2196/55939.
17. Yang R, Nair SV, Ke Y, D'Agostino D, Liu M, Ning Y, et al. Disparities in Clinical Studies of AI Enabled Applications from a Global Perspective. *NPJ Digit Med.* 2024;7(1):209. doi: 10.1038/s41746-024-01212-7.
18. Wei K, Fritz C, Rajasekaran K. Answering Head and Neck Cancer Questions: An Assessment of ChatGPT Responses. *Am J Otolaryngol.* 2024;45(1):104085. doi: 10.1016/j.amjoto.2023.104085.
19. Lewin S, Chetty R, Ithdayhid AR, Dwivedi G. Ethical Challenges and Opportunities in Applying Artificial Intelligence to Cardiovascular Medicine. *Can J Cardiol.* 2024;40(10):1897-906. doi: 10.1016/j.cjca.2024.06.029.
20. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med.* 2023;183(6):589-96. doi: 10.1001/jamainternmed.2023.1838.
21. Stanceski K, Zhong S, Zhang X, Khadra S, Tracy M, Korla L, et al. The Quality and Safety of Using Generative AI to Produce Patient-Centred Discharge Instructions. *NPJ Digit Med.* 2024;7(1):329. doi: 10.1038/s41746-024-01336-w.
22. Armondas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, et al. Use of Artificial Intelligence in Improving Outcomes in Heart Disease: A Scientific Statement from the American Heart Association. *Circulation.* 2024;149(14):e1028-50. doi: 10.1161/CIR.0000000000001201.
23. Regmi K, Jones L. A Systematic Review of the Factors - Enablers and Barriers - Affecting E-Learning in Health Sciences Education. *BMC Med Educ.* 2020;20(1):91. doi: 10.1186/s12909-020-02007-6.
24. Brown MT, Bussell JK. Medication Adherence: WHO Cares? *Mayo Clin Proc.* 2011;86(4):304-14. doi: 10.4065/mcp.2010.0575.
25. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review. *NPJ Digit Med.* 2024;7(1):258. doi: 10.1038/s41746-024-01258-7.
26. Jasińska-Stroschein M, Dymek J, Drozd M, Sierpniowska O, Jędra A, Stankiewicz A, et al. An Evaluation of Written Materials for Supporting Hypertensive Patient Education and Counselling When Performing a New Medicine Service in Poland. *BMC Med Educ.* 2024;24(1):521. doi: 10.1186/s12909-024-05523-x.
27. Akinrinmade AO, Adebile TM, Ezuma-Ebong C, Bolaji K, Ajufu A, Adigun AO, et al. Artificial Intelligence in Healthcare: Perception and Reality. *Cureus.* 2023;15(9):e45594. doi: 10.7759/cureus.45594.
28. Colasacco CJ, Born HL. A Case of Artificial Intelligence Chatbot Hallucination. *JAMA Otolaryngol Head Neck Surg.* 2024;150(6):457-8. doi: 10.1001/jamaoto.2024.0428.
29. Gondode P, Duggal S, Mahor V. Artificial Intelligence Hallucinations in Anaesthesia: Causes, Consequences and Countermeasures. *Indian J Anaesth.* 2024;68(7):658-61. doi: 10.4103/ija.ija_203_24.
30. Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. *Cureus.* 2023;15(9):e44720. doi: 10.7759/cureus.44720.
31. Sousa-Pinto B, Vieira RJ, Marques-Cruz M, Bognanni A, Gil-Mata S, Jankin S, et al. Artificial Intelligence-Supported Development of Health Guideline Questions. *Ann Intern Med.* 2024;177(11):1518-29. doi: 10.7326/ANNALS-24-00363.
32. Sun G, Zhou YH. AI in Healthcare: Navigating Opportunities and Challenges in Digital Communication. *Front Digit Health.* 2023;5:1291132. doi: 10.3389/fdgh.2023.1291132.
33. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The Impact of Artificial Intelligence on the Person-Centred, Doctor-Patient Relationship: Some Problems and Solutions. *BMC Med Inform Decis Mak.* 2023;23(1):73. doi: 10.1186/s12911-023-02162-y.
34. Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, et al. Potential and Limitations of ChatGPT 3.5 and 4.0 as a Source of COVID-19 Information: Comprehensive Comparative Analysis of Generative and Authoritative Information. *J Med Internet Res.* 2023;25:e49771. doi: 10.2196/49771.
35. Tang Y, Xiao Z, Li X, Zhang Q, Chan EW, Wong IC, et al. Large Language Model in Medical Information Extraction from Titles and Abstracts with Prompt Engineering Strategies: A Comparative Study of GPT-3.5 and GPT-4 [Internet]. *medRxiv*; 2024 [cited 2025 Feb 28]. p. 2024.03.20.24304572. Available from: <https://www.medrxiv.org/content/10.1101/2024.03.20.24304572v1>
36. Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics.* 2021;10(5):593. doi: 10.3390/electronics10050593.
37. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An Evaluation Framework for Clinical Use of Large Language Models in Patient Interaction Tasks. *Nat Med.* 2025;31(1):77-86. doi: 10.1038/s41591-024-03328-5.

*Supplemental Materials

For additional information Supplemental Material 1, please click here.
For additional information Supplemental Material 2, please click here.



This is an open-access article distributed under the terms of the Creative Commons Attribution License