ORIGINAL RESEARCH

Bagcilar Med Bull
DOI: 10.4274/BMB.galenos.2026.96720

Reliability and Consistency of ChatGPT-4o's Responses to Frequently Asked Questions About Hypertension

ChatGPT-4o'nun Hipertansiyon Hakkında Sıkça Sorulan Sorulara Verdiği Yanıtların Güvenilirliği ve Tutarlılığı

Ekrem Bilal Karaayvaz, Berk Batuhan Bayraktar, Göksel Güz

İstanbul University, İstanbul Faculty of Medicine, Department of Cardiology, İstanbul, Turkey

Abstract

Objective: To evaluate the quality, validity and reliability of the answers given by ChatGPT-4o to frequently asked questions about systemic arterial hypertension (HT), which directly concerns public health.

Method: In this study, 30 frequently asked questions about HT from health forums and hospital websites were compiled and divided into four categories: General HT questions (n=10), treatment-related questions (n=10), specific questions (n=5), and questions based on the 2024 European Society of Cardiology (ESC) guidelines (n=5). Questions were asked in Turkish to ChatGPT-4o and the answers were evaluated by three cardiologists using the global quality scale (GQS). The consistency of the answers was tested by repeating the same questions on different days.

Results: 56.7% of the responses were rated as high quality and comprehensive answers (GQS 5), 40% as largely comprehensive and good quality answers (GQS 4) and 3.3% as moderate quality (GQS 3). 80% of the questions based on the ESC guidelines received a GQS 5 score and demonstrated 100% reproducibility.

Conclusion: ChatGPT-4o can provide reliable, accurate, and repeatable answers to frequently asked questions about HT. However, it should be noted that users should support this information with professional medical advice.

Keywords: Artificial intelligence, ChatGPT-4o, systemic hypertension

Öz

Amaç: Toplum sağlığını doğrudan ilgilendiren sistemik arteriyel hipertansiyon (HT) hakkında sıkça sorulan sorulara ChatGPT-4o'nun verdiği yanıtların kalite, geçerlilik ve güvenilirliğini değerlendirmektir.

Yöntem: Bu çalışmada sağlık forumları ve hastane web sitelerinde yer alan HT ile ilgili 30 sık sorulan soru derlenmiş ve dört kategoriye ayrılmıştır: Genel HT soruları (n=10), tedaviyle ilgili sorular (n=10), spesifik sorular (n=5) ve 2024 Avrupa Kardiyoloji Derneği (ESC) kılavuzuna dayalı sorular (n=5). Sorular ChatGPT-4o'ya Türkçe yöneltilmiş ve verilen yanıtlar üç kardiyoloji uzmanı tarafından global kalite ölçeği (GQS) ile değerlendirilmiştir. Yanıtların tutarlılığı, aynı soruların farklı günlerde tekrarlanarak sorulmasıyla test edilmiştir.

Bulgular: Yanıtların %56,7'si yüksek kaliteli ve kapsamlı yanıtlar (GQS 5), %40'ı büyük oranda kapsamlı ve iyi kalitede yanıtlar (GQS 4) ve %3,3'ü ise orta düzeyde kalite (GQS 3) olarak değerlendirilmiştir. ESC kılavuzuna dayalı soruların %80'i GQS 5 puanı almış ve %100 tekrarlanabilirlik göstermiştir.

Sonuç: ChatGPT-4o, HT ile ilgili sık sorulan sorulara güvenilir, doğru ve tekrarlanabilir yanıtlar verebilmektedir. Ancak kullanıcıların bu bilgileri profesyonel tıbbi danışmanlıkla desteklemesi gerektiği unutulmamalıdır.

Anahtar kelimeler: ChatGPT-4o, sistemik arteriyel hipertansiyon, yapay zeka

Address for Correspondence: Lec, Ekrem Bilal Karaayvaz, MD, İstanbul University, İstanbul Faculty of Medicine, Department of Cardiology, İstanbul, Turkey
E-mail: ekrembilal@gmail.com **ORCID:** orcid.org/0000-0002-0105-6167
Received: 26.09.2025 **Accepted:** 04.04.2026 **Epub:** 15.04.2026

Cite this article as: Karaayvaz EB, Bayraktar BB, Güz G. Reliability and consistency of ChatGPT-4o's responses to frequently asked questions about hypertension. Bagcilar Med Bull. [Epub Ahead of Print]

Introduction

ChatGPT is an artificial intelligence (AI) language model developed by OpenAI and designed to produce texts that can mimic human conversations (1). This technology has recently become a platform for people seeking information on a wide range of topics. As with other curiosities, people seek to use ChatGPT as a source of health information. Having information across a wide spectrum, from symptoms and diagnosis to treatment options, is particularly appealing to people. It is not surprising that ChatGPT is preferred by people due to its proactive nature and its potential to provide them with additional information based on the information it obtains by offering them a wide range of options. The popularity of ChatGPT and people's desire to use this AI to obtain information on any subject, receive suggestions, and even make decisions raise the question of how reliable the answers given by this AI language model are (2-4). People who want to use this platform should be able to obtain accurate health information and assess how accurate, evidence-based, and up-to-date the information is. Some studies have examined ChatGPT's validity and reliability across diverse topics within various health domains (5-8).

Systemic arterial hypertension (HT) is a major health problem affecting more than one billion adults worldwide (9). It has been shown that uncontrolled HT causes 9.4 million deaths and 212 million healthy life years lost each year (10). HT, a public health problem, is a health term frequently searched by people on internet search engines to obtain information about both diagnosis and treatment and follow-up (11). In this study, we aimed to investigate the quality, validity, and reliability of ChatGPT-4o's responses to HT-related questions, which are of particular concern to public health.

Materials and Methods

This study was conducted between March 18, 2025, and April 15, 2025. Since no patient data were used, informed consent was not required and therefore was not obtained.

Ethical approval for the study was obtained from the Institutional Ethics Committee of Medicana International İstanbul Hospital (decision number: 2025/2031; date: 24.09.2025).

The questions used in the study were compiled from frequently asked queries about HT on health forums and hospital websites. Questions for advertising purposes, those requiring personal responses, and repetitive questions were excluded from the study. Only questions written in Turkish were considered; a total of 30 were evaluated. The questions were divided into four groups: General questions about HT (10); questions about HT treatment (10); specific questions about HT (5); and questions related to the 2024 European Society of Cardiology (ESC) HT guidelines (12). The guidelines were reviewed, and 5 questions about HT were prepared. All questions are presented in Table 1.

Questions were directed to ChatGPT-4o in Turkish. The responses provided by ChatGPT-4o were evaluated by two cardiologists, each with at least 10 years of experience, and they were unaware of the score prior to the study. If the evaluations of the two physicians were the same, this score was recorded directly. If there was a difference between the evaluations, a third experienced cardiologist also performed an evaluation, and the final score was calculated as the arithmetic mean of the three scores. The consistency of ChatGPT's responses was tested by asking each question twice, on different days. Each question was submitted to ChatGPT-4o twice, with an interval of at least 7 days between queries. This interval was chosen to minimize potential short-term contextual or memory-related effects and to better assess the reproducibility of responses over time. The same ChatGPT account was used for all queries, ensuring consistency in model access. Each question was submitted in a new independent session, with chat history cleared prior to each query, so no prior prompts or responses were visible to the model. This approach was deliberately adopted to reduce contextual carryover effects and to ensure that each response was generated independently.

The quality and reliability of ChatGPT-4o responses were assessed using the global quality scale (GQS), developed to evaluate the accuracy and adequacy of medical content. GQS is a 1-to-5 scoring system used to assess the quality and reliability of written medical content.

According to scoring:

- GQS 1: Low quality, poorly organized, missing much of the essential information, and useless content for the patient.
- GQS 2: Overall poor structure and content, some information available but missing important topics, of little benefit to the patient.
- GQS 3: Moderate quality, some important information is adequately covered but others are inadequate, providing moderate benefit to the patient.

Table 1. Questions asked to ChatGPT-4o and answer scores

General questions about hypertension	**GQS**
1. What is hypertension?	4
2. What are the symptoms of hypertension?	5
3. How is hypertension diagnosed?	4
4. What are the causes of hypertension?	4
5. Are home blood pressure measurements from the wrist reliable?	5
6. Hypertension is more common in which age groups?	5
7. Is hypertension hereditary? Does my risk increase if it runs in my family?	4
8. Which organs are damaged by hypertension?	5
9. What is the relationship between hypertension and stress?	5
10. What kind of health problems can hypertension cause if left untreated?	5
Questions about hypertension treatment	
1. How is hypertension treated?	4
2. Does garlic and lemon juice lower blood pressure?	5
3. Are hypertension medications used for life?	5
4. How can blood pressure be lowered without medication?	4
5. How should hypertension patients eat?	5
6. How does salt consumption affect hypertension?	5
7. What lifestyle changes are recommended for hypertension?	4
8. How does exercise affect blood pressure?	4
9. Are alternative medicine methods effective in the treatment of hypertension?	3
10. Can hypertension be treated with surgery?	5
Specific questions about hypertension	
1. How is hypertension treated during pregnancy?	4
2. Does hypertension increase the risk of heart attack and stroke?	4
3. Does hypertension affect sexual life?	5
4. How alcohol and smoking affect hypertension?	4
5. Does blood pressure rise when visiting the doctor?	5
Hypertension questions based on the ESC guidelines	
1. How is hypertension diagnosed? What blood pressure levels are considered hypertension according to ESC guidelines?	5
2. What are the criteria for initiating antihypertensive treatment according to ESC guidelines?	5
3. How are treatment initiation thresholds determined based on blood pressure level, age and comorbidities?	5
4. According to ESC guidelines, which agents should be included in first-line drug therapy for hypertension? In what situations is combination therapy recommended?	4
5. According to ESC guidelines, what is the role of lifestyle changes in hypertension management? In which cases are measures such as diet, exercise, and salt restriction considered sufficient?	5

GQS: Global quality scale, ESC: European Society of Cardiology

• GQS 4: Good quality and structure, most important information is included, some omissions but useful for the patient.

• GQS 5: High quality and well organized, all important topics are covered comprehensively, extremely useful for the patient.

Statistical Analysis

All analyses were performed using the Statistical Package for the Social Sciences (SPSS) version 22.0 (SPSS Inc., Chicago, USA). Subcategory GQS scores are presented as percentages, and mean scores and repeatability rates as figures.

Interobserver reliability among the cardiologists evaluating the ChatGPT-4o responses was assessed using the intraclass correlation coefficient (ICC). Given that the GQS scores are ordinal and that more than two raters were involved, a two-way random-effects model for absolute agreement was applied. ICC values were interpreted according to commonly accepted criteria, with values below 0.50 indicating poor agreement, values of 0.50-0.75 indicating moderate agreement, values of 0.75-

0.90 indicating good agreement, and values above 0.90 indicating excellent agreement. Statistical analyses were performed using SPSS version 22.0 (SPSS Inc., Chicago, USA).

Results

In total, ChatGPT-4o answered 17 out of 30 questions (56.7%) comprehensively, providing high-quality, well-structured answers and earning a GQS score of 5. Twelve questions (40%) received a GQS score of 4, with answers that were largely comprehensive and of good quality, whereas only one question (3.3%) received a GQS score of 3, with answers of moderate quality and usefulness. The mean GQS score for the answers is shown in Figure 1.

6 (60%) of the responses to general HT questions were rated as high quality and comprehensive, receiving a GQS score of 5. Four (40%) responses received GQS scores of 4.

While 5 (50%) of the treatment-related questions received a GQS score of 5, 4 (40%) received a GQS score of 4. Only 1 (10%) question — "Are alternative medicine methods effective in the treatment of HT?"— received a GQS score of 3 because its content was insufficiently comprehensive and provided limited benefit.

In the specific questions category, 2 questions (40%) received a GQS 5 score, while the remaining 3 questions (60%) were rated as GQS 4.

Of ChatGPT-4o's responses to questions derived from the ESC guidelines, 4 (80%) were scored as GQS 5, indicating the highest quality and comprehensiveness, while 1 response was scored as GQS 4.

The reproducibility rates of ChatGPT-4o responses are presented in Figure 2. A reproducibility rate of 80% was observed for general and specific questions on HT. This rate was highest for questions based on ESC guidelines at 100%, while the lowest reproducibility rate was observed for questions on HT treatment at 70%.

Interobserver agreement analysis demonstrated good-to-excellent consistency among the cardiologists evaluating ChatGPT-4o responses. The calculated ICC was 0.82 (95% confidence interval: 0.76-0.87), indicating a high degree of agreement between expert raters and supporting the reliability of the scoring methodology employed in this study.

Discussion

Since its launch in November 2022, ChatGPT has quickly become a popular AI platform that both patients and clinicians use to obtain information and facilitate decision-making in diagnosis and treatment processes (13). In this study, in which we investigated the quality and validity of ChatGPT's answers to questions about HT, we observed that ChatGPT provided high-quality, structured answers to more than half of the questions (56.7%) and good-quality answers to 40% of the questions. Half of the answers to treatment-related questions were high-quality, comprehensive, and extremely useful for patients, and nearly half (40%) were determined to be good-quality and useful; these answers had a GQS score of 4. According to the results of our study, ChatGPT's answers to general HT and treatment questions were high-quality, but it is not sufficient to prove that these have clinical accuracy.

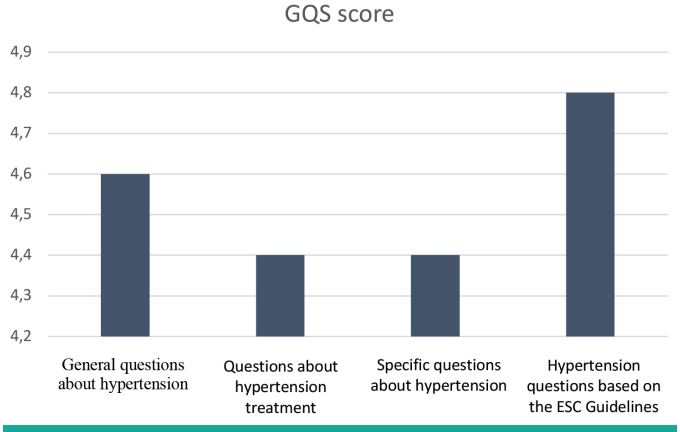

Figure 1. Average GQS scores of question answers
GQS: Global quality scale, ESC: European Society of Cardiology

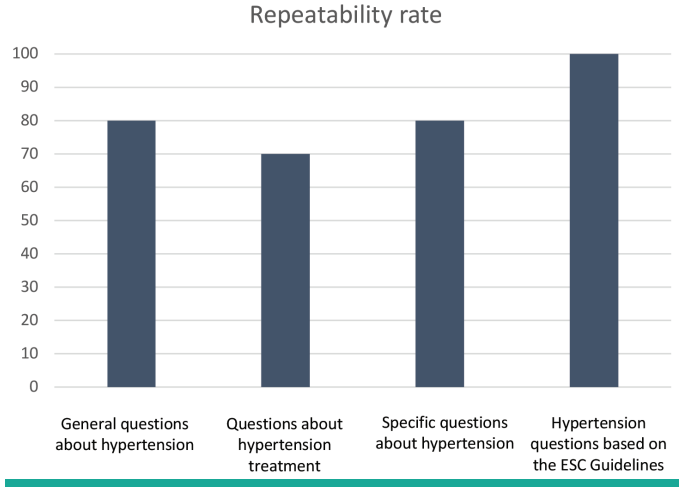

Figure 2. Repeatability rates of question answers
ESC: European Society of Cardiology

HT and/or some cardiological diseases have been addressed in different studies in the literature in terms of the functionality of AI (14-16). In an evaluation of the 100 question HT, Almagazzachi et al. (17) reported that ChatGPT responses were appropriate in 92.5% of cases and inappropriate in 7.5%, with a reproducibility rate of 93% for ChatGPT. Kerkütlüoğlu et al. (18) evaluated ChatGPT in terms of knowledge and disease management regarding pulmonary HT, and 10 experts evaluated the performance of ChatGPT. Accordingly, the responses generated by ChatGPT were found to be reliable, with an average score of 8.4 (7.7-9.2), and valuable, with an average score of 7.9 (7.4-8.2).

In a study of laboratory and demographic data of 40 HT patients treated in a rural clinic in Georgia, Al Tibi et al. (19) compared the medical recommendations made by the cardiologist to the patients with the recommendations made by ChatGPT for the same patients and laboratory data. In contrast to the other studies mentioned above, discrepancies between the cardiologist and GPT-4 regarding general recommendations occurred in 95% of the 40 patients, while only 10.2% of specific medication recommendations were consistent between the cardiologist and GPT-4. Furthermore, the cardiologist and GPT-4 did not agree on medication changes. The authors highlighted the existence of different optimal laboratory value ranges among patients, the lack of a holistic analysis of GPT-4, and the need to provide complementary information to the model as the reasons for this discrepancy.

Our study also evaluated specific questions about HT, the ESC guidelines. Accordingly, 4 (80%) of the answers provided the highest quality and comprehensiveness. The answers to the ESC-based questions were also 100% reproducible. In their study conducted in 2023, Kusunose et al. (20) examined the ability of ChatGPT to accurately answer clinical questions about the Japanese Society of Hypertension's Hypertension Management Guidelines (JSH 2019). Similar to our study, ChatGPT had an 80% accuracy rate in responses to clinical questions. ChatGPT's performance in terms of HT responses was similar across two different guidelines. That ChatGPT complied with the guidelines shows that clinicians can use this model for information in their daily practice.

We emphasize that our work differs from prior studies by evaluating ChatGPT-4o, a recently released, more advanced model, whereas most prior studies assessed earlier versions of ChatGPT or other AI systems. using a patient-centered question set written entirely in Turkish, addressing a non-English language context that remains underrepresented in the current literature; including guideline-based questions derived from the 2024 ESC HT Guidelines, allowing for a structured assessment of guideline concordance; systematically assessing reproducibility by repeating all questions on different days under controlled conditions.

Study Limitations

Our study has some limitations. First, our patient-centered approach in determining the questions made those questions inherently subjective, and a relatively small number of questions and reviewers were included. Second, there may be differences in approach among the medical professionals and cardiologists who served as raters assessing the accuracy and consistency of the answers and information provided by ChatGPT. Furthermore, questions submitted to ChatGPT were limited to those from the healthcare forums and hospital websites examined for HT. It remains unclear how many questions are optimal for evaluating ChatGPT. Asking questions only in Turkish may have introduced a language bias. ChatGPT is rapidly evolving, and reproducibility may change in future versions.

Conclusion

We demonstrated that ChatGPT produces significantly accurate and reproducible answers to a variety of medical questions related to HT. ChatGPT may provide reliable information about HT, but it is important to seek professional medical advice before making any decisions about HT. Despite the limitations of the study, ChatGPT may serve as a useful source of information for both patients and healthcare professionals when used carefully. ChatGPT can provide only general information and support; a healthcare professional can make specific recommendations based on a holistic assessment of the patient's individual characteristics and diagnostic findings. Studies including a larger number of questions and medical professionals will shed more light on the use of ChatGPT as a source of information about HT.

Ethics

Ethics Committee Approval: Ethical approval for the study was obtained from the Institutional Ethics Committee of Medicana International İstanbul Hospital (decision number: 2025/2031; date: 24.09.2025).

Informed Consent: Since no patient data were used, informed consent was not required and therefore was not obtained.

Footnotes

Authorship Contributions

Surgical and Medical Practices: B.B.B., Concept: E.B.K., Design: E.B.K., G.G., Data Collection or Processing: E.B.K., G.G., Analysis or Interpretation: E.B.K., B.B.B., Literature Search: B.B.B., G.G., Writing: E.B.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. OpenAI. ChatGPT: optimizing language models for dialogue, 2023. Available from:https://openai.com/blog/chatgpt/.
2. Pugliese N, Wai-Sun Wong V, Schattenberg JM, Romero-Gomez M, Sebastiani G; NAFLD Expert Chatbot Working Group; et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2024;22(4):886-889.e5.
3. 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:e2336483.
4. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024;66(1):73-79.
5. Alabdulmohsen DM, Almahmudi MA, Alhashim JN, Almahdi MH, Alkishy EF, Almossabeh MJ, et al. Is ChatGPT a reliable source of patient information on asthma? Cureus. 2024;16(7):e64114.
6. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023;13:1256459.
7. Beilby K, Hammarberg K. ChatGPT: a reliable fertility decision-making tool? Hum Reprod. 2024;39(3):443-447.
8. Tunçer G, Güçlü KG. How reliable is ChatGPT as a novel consultant in infectious diseases and clinical microbiology? Infect Dis Clin Microbiol. 2024;6(1):55-59.
9. Gaziano TA, Bitton A, Anand S, Weinstein MC; International Society of Hypertension. The global cost of nonoptimal blood pressure. J Hypertens. 2009;27(7):1472-1427.
10. GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the global burden of disease study 2015. Lancet. 2016;388(10053):1659-1724. Erratum in: Lancet. 2017;389(10064):e1.
11. Hypertension. GoogleTrends. [Internet]. Available from:https://trends.google.com/trends/explore?date=today%205-y&q=%2Fm%2F0k95h&hl=tr.
12. McEvoy JW, McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, et al. 2024 ESC Guidelines for the management of elevated blood pressure and hypertension. Eur Heart J. 2024;45(3):3912-4018. Erratum in: Eur Heart J. 2025;46(14):1300. Erratum in: Eur Heart J. 2025;46(45):4949
13. Layton AT. AI, machine learning, and ChatGPT in hypertension. Hypertension. 2024;81(4):709-716.
14. Andreadis K, Rodriguez DV, Zakreuskaya A, Chen J, Gonzalez J, Mann D. Bridging gaps with generative AI: enhancing hypertension monitoring through patient and provider insights. Stud Health Technol Inform. 2024;316:939-943.
15. Saeed A, AlShafea A, A F, Bin Saeed A. Pacemaker malfunction in a patient with congestive heart failure and hypertension. Cureus. 2023;15(2):e34574.
16. Alam SF, Thongprayoon C, Miao J, Pham JH, Sheikh MS, Garcia Valencia OA, et al. Advancing personalized medicine in digital health: the role of artificial intelligence in enhancing clinical interpretation of 24-h ambulatory blood pressure monitoring. Digit Health. 2025;11:20552076251326014.
17. Almagazzachi A, Mustafa A, Eighaei Sedeh A, Vazquez Gonzalez AE, Polianovskaia A, Abood M, et al. Generative artificial intelligence in patient education: ChatGPT takes on hypertension questions. Cureus. 2024;16(2):e53441.
18. Kerkütlüoğlu M, Kaya E, Gökmen R. Trustworthiness, value, danger, and readability of ChatGPT-generated responses to health questions related to pulmonary arterial hypertension. Cureus. 2024;16(10):e71472.
19. Al Tibi G, Alexander M, Miller S, Chronos N. A retrospective comparison of medication recommendations between a cardiologist and ChatGPT-4 for hypertension patients in a rural clinic. Cureus. 2024;16(3):e55789.
20. Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension Guidelines. Circ J. 2023;87(7):1030-1033.