Circulation Journal
Circ J 2023; 87: 1030–1033
doi:10.1253/circj.CJ-23-0308

RAPID COMMUNICATION

Evaluation of the Accuracy of ChatGPT in Answering Clinical Questions on the Japanese Society of Hypertension Guidelines

Kenya Kusunose, MD, PhD; Shuichiro Kashima; Masataka Sata, MD, PhD

Background: To assist healthcare providers in interpreting guidelines, clinical questions (CQ) are often included, but not always, which can make interpretation difficult for non-expert clinicians. We evaluated the ability of ChatGPT to accurately answer CQs on the Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019).

Methods and Results: We conducted an observational study using data from JSH 2019. The accuracy rate for CQs and limited evidence-based questions of the guidelines (Qs) were evaluated. ChatGPT demonstrated a higher accuracy rate for CQs than for Qs (80% vs. 36%, P value: 0.005).

Conclusions: ChatGPT has the potential to be a valuable tool for clinicians in the management of hypertension.

Key Words: ChatGPT; Guidelines; Hypertension; Large language models

Clinical decision-making for the management of patients relies heavily on accurate interpretation of clinical guidelines,[1] this can sometimes be challenging, especially for healthcare professionals who lack expertise in a particular clinical domain. To assist interpretation, evidence-based questions and answers, termed "clinical questions" (CQs), are often, but not always, included in guidelines,[2] and without CQs it can be difficult for non-expert clinicians to effectively use the guideline.

Recent significant advances in artificial intelligence (AI),[3–5] especially the large language models (LLMs), and the self-regressive LMs in particular, have attracted considerable attention.[6] On November 30, 2022, OpenAI launched ChatGPT, which is a refined and accessible LLM that is becoming more widely used as a new level of service for retrieving information, answers, or solutions. Limited studies have evaluated the effectiveness of non-domain specific natural language models, such as ChatGPT, in assisting healthcare professionals in interpreting clinical guidelines.[7]

Hypertension affects approximately 1 billion people worldwide and is a significant risk factor for cardiovascular diseases such as stroke and coronary artery disease.[8] Japan has one of the highest prevalence rates of hypertension in the world, and hypertension-related diseases are a major public health issue.[9] The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019) is a key document for clinical management in Japan and includes CQs.[10] In this study, we aimed to evaluate the accuracy of ChatGPT's responses to CQs on JSH 2019 to determine whether AI can aid clinicians in the interpretation of the guidelines.

Methods

Study Design

We used an observational cohort design and the data source for this study was the JSH 2019.[10]

ChatGPT

ChatGPT (OpenAI, San Francisco, CA, USA) is a natural language processing tool based on the Generative Pre-trained Transformer 3.5 architecture. It is pre-trained on a large corpus of text data, enabling it to generate responses to a wide range of text-based inputs. In this study, ChatGPT was used to generate responses to CQs related to JSH 2019, and to evaluate its accuracy these answers were compared with the correct responses as outlined in the guidelines.

Outcome Measures

On April 11, 2023, the questions in Japanese were manually entered into the ChatGPT interface. In instances where the answer format was uncertain from the original question, the statement "Provide one answer" was appended as nec-

Received May 7, 2023; revised manuscript received May 18, 2023; accepted May 28, 2023; J-STAGE Advance Publication released online June 7, 2023 Time for primary review: 11 days

Department of Cardiovascular Medicine, Tokushima University Hospital, Tokushima (K.K., S.K., M.S.); Department of Cardiovascular Medicine, Nephrology, and Neurology, Graduate School of Medicine, University of the Ryukyus, Okinawa (K.K.), Japan
Mailing address: Kenya Kusunose, MD, PhD, Department of Cardiovascular Medicine, Tokushima University Hospital, 2-50-1 Kuramoto, Tokushima 770-8503, Japan. email: echo.cardio@gmail.com

ISSN-1346-9843

essary. For certain questions that were linked, the second question was entered after the first question's text, to enable ChatGPT to provide answers. The primary outcome measure of this study was the accuracy of ChatGPT's responses to CQs and to limited evidence-based questions (Qs) on JSH 2019. Accuracy was defined as the proportion of correct responses generated by ChatGPT out of the total number of questions asked. The accuracy was judged by a certified cardiology doctor (K.K.).

The following standards were utilized to ascertain the accuracy of the responses.

(1) Queries that could be resolved with a binary answer (YES or NO) were judged to be accurate if they corresponded precisely.
(2) For inquiries that elicited a numerical response, the answer was deemed accurate if the number provided corresponded exactly.
(3) Questions that were responded to in writing were deemed accurate if there was no disparity between the answer and the contents of the response given by ChatGPT. Even if not all elements were included, the query was considered correct if it did not comprise any errors.

Regarding the variability of ChatGPT's answers, we used the default value of 0.7 for the temperature parameter. The temperature parameter in AI language models is a control parameter for the model's output randomness. It affects the probability distribution used by the model when generating text, effectively modifying how conservative or adventurous the model is when it creates new output.

Statistical Analysis

Descriptive analyses were performed to evaluate the accuracy of ChatGPT's responses. We compared the accuracy rate of responses to the 2 types of questions: CQs and Qs related to limited evidence-based questions, recommendation level and evidence level, and in Japanese and English. We used a chi-square test to compare the proportions of correct responses between groups. We used Shannon Entropy to measure the degree of uncertainty or randomness in the responses generated by ChatGPT. Specifically, we calculated the entropy of the set of responses generated by ChatGPT for each CQ. Shannon Entropy is a statistical measure that quantifies the amount of information or uncertainty in a probability distribution. It ranges from 0 (no uncertainty) to 1 (maximum uncertainty). To calculate the entropy, we first constructed a set of responses generated by ChatGPT for each CQ by repeating the same question 10 times. We then calculated the frequency of each unique response in the set and used these frequencies to compute the entropy of the set. The statistical analysis was performed using a standard statistical software package (MedCalc Software 18; Mariakerke, Belgium). The threshold for statistical significance was set to $P<0.05$.

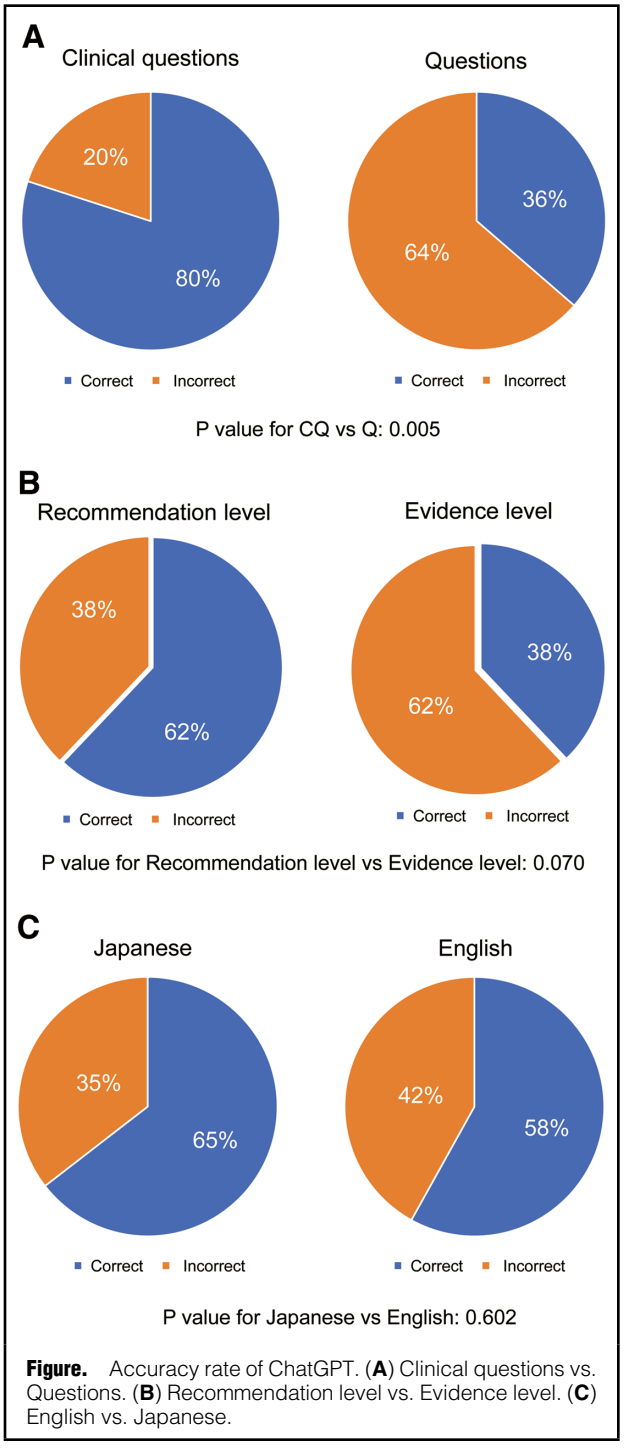

Figure. Accuracy rate of ChatGPT. (**A**) Clinical questions vs. Questions. (**B**) Recommendation level vs. Evidence level. (**C**) English vs. Japanese.

Results

Descriptive analyses were performed to evaluate the accuracy of ChatGPT in providing responses to the CQs and Qs on JSH 2019. There were 17 CQs with 3 sub-questions and 9 Qs in JSH2019. Because the 9th Q had separate answers for 3 measurements, it was split into 3 separate questions for evaluation. A total of 31 questions were put to ChatGPT, and it correctly answered 20, resulting in an overall accuracy rate of 64.5%.

The accuracy rate for CQs and Qs of the guidelines was also evaluated. ChatGPT demonstrated a higher accuracy rate for questions related to CQs (80%) compared with Qs related to limited evidence-based questions (36%, P value: 0.005) (**Figure A**). Additionally, ChatGPT showed a trend of higher accuracy rate for recommendation level (62%) than for evidence level (38%, P value: 0.070) (**Figure B**). No significant differences in accuracy were observed between questions that were originally written in Japanese (65%) and those that were translated into Japanese from English (58%, P value: 0.602) (**Figure C**).

Table. Accuracy and Entropy When ChatGPT Was Asked the Same Clinical Question 10 Times

	Accuracy	Entropy
CQ1	0.9	0.468996
CQ2	0.9	0.468996
CQ3	1	0
CQ4	1	0
CQ5	1	0
CQ6	1	0
CQ7	0.8	0.721928
CQ8	0.4	0.970951
CQ9	0	0
CQ10-1	0.9	0.468996
CQ10-2	0.9	0.468996
CQ11	1	0
CQ12	0	0
CQ13-1	0.4	0.970951
CQ13-2	1	0
CQ14	0.8	0.721928
CQ15	0.2	0.721928
CQ16	0.4	0.970951
CQ17-1	0.4	0.970951
CQ17-2	1	0

We used the Shannon Entropy to test whether ChatGPT could ask the same question 10 times and get the same answer (**Table**). Of the 21 CQs, 9 had zero entropy (i.e., the answers were all identical). Of the remaining 12 questions, 7 had an entropy >0.5 (i.e., there was an unacceptable blurring of answers). The tendency of questions with high entropy was examined and it was not related to the length of the text, the strength of the evidence or the recommendations.

Discussion

ChatGPT provided accurate responses to CQs related to the JSH 2019 guidelines for the management of hypertension. However, the overall accuracy rate of 64.5% may not be sufficient to characterize ChatGPT as "accurate" and possessing "the potential to be a valuable tool". ChatGPT showed a higher accuracy rate for CQs than for Qs related to limited evidence-based questions, and a trend of higher accuracy rate for recommendation level than for evidence level. No significant differences in accuracy were observed between questions originally written in Japanese and those translated from English.

Comparison With Prior Studies

Although there are no prior studies that have evaluated the accuracy of the ChatGPT's responses to CQs specifically about guidelines, there are several studies that have looked at the accuracy of ChatGPT on CQs. A recent study in the USA found that ChatGPT adequately answered 21 of 25 questions in response to Cardiovascular Disease Prevention Recommendations.[7] The reasons for wrong answers were related to information not commercially available, such as incorrect low-density lipoprotein cholesterol cutoff values in the most recent guidelines or failure to include genetic considerations regarding familial high cholesterol. Our study results suggest that ChatGPT can be a useful tool for healthcare professionals in answering CQs on hypertension, especially evidence-based questions. However, caution should be exercised when using ChatGPT for more complex and nuanced questions. Furthermore, although we demonstrated the accuracy of ChatGPT in answering CQs, our study did not address the effect of the ChatGPT on clinical decision making or patient outcomes.

Concerns About LLMs

In the rapidly evolving field of healthcare, AI, particularly natural language processing models such as ChatGPT, has shown promise in interpreting and answering CQs. However, it is important to note that these models might produce "hallucinations", or plausible-sounding but unverified or incorrect information, and have limitations due to the training data.[11,12] The ChatGPT model, trained on a range of internet text available up until September 2021, might reflect more recent medical knowledge not included in the guidelines at the time of their publication. In the future, the advent of extended LMs that can perform real-time searches to include up-to-date information in their responses could potentially address these concerns. However, we must recognize and work within the limitations of existing models such as ChatGPT at this time. It can be a valuable tool when used with a clear understanding of its capabilities and limitations, including the potential for hallucinations and the lack of real-time information updates.

Clinical Implications

To our knowledge, this is the first study to evaluate the accuracy of ChatGPT in providing responses to CQs related to the JSH guidelines for the management of hypertension. Our study showed that ChatGPT has the potential to be a valuable tool for clinicians who need quick access to accurate information on hypertension management. Nonetheless, the limitations of ChatGPT in answering certain types of questions should be considered when utilizing it. Furthermore, our study provided important insights into the limitations of ChatGPT in responding to certain types of questions, which can guide the development of future natural language processing tools.

Although AI presents an innovative tool for answering clinical queries, users must remain aware that responses are not always backed by sufficient evidence. Mitigating risks and advocating for adequate safeguards and regulations by governments of various countries becomes crucial in harnessing the potential of AI in healthcare.[12]

Study Limitations

Our study has several limitations that should be considered when interpreting the results. First, the sample size was relatively small, which may limit generalization of our findings. Second, our study focused only on the JSH guidelines for the management of hypertension in Japan, and further research is needed to evaluate the accuracy of ChatGPT in responding to CQs related to other medical specialties and guidelines. Third, it is possible that with more carefully crafted prompts (e.g., "according to the recent evidence"), the percentage of correct answers may be higher. Future research can be planned to explore optimal prompting strategies to increase the model's accuracy. Finally, our study did not evaluate the effect of using ChatGPT on clinical outcomes, and further research is needed to determine whether it improves patient outcomes.

Conclusions

The results of our study indicate that ChatGPT has the potential to be a valuable tool for clinicians in the management of hypertension. Overall, the results demonstrated that ChatGPT had a high accuracy rate in providing responses to CQs and Qs on JSH 2019. The accuracy rate varied by available evidence, suggesting that further improvements to ChatGPT may be necessary for several types of questions.

We suggest AI can be used as a supplementary tool in healthcare, providing preliminary guidance and quick reference to established guidelines. However, there are several limitations in the current AI model, and thus its role should be to complement expert medical knowledge.

Disclosures

None.

Contributors

K.K. conceived the idea for this study. K.K. and S.K. conducted the data analyses. The initial draft of the manuscript was produced by K.K. All authors were involved in interpreting the results and writing the manuscript. All authors read and approved the final manuscript.

Sources of Funding

This work was partially supported by grants from JSPS Kakenhi Grants (no. 23K07509 to K.K.), the Japan Agency for Medical Research and Development (AMED, JP22uk1024007 to K.K.) and Takeda Science Foundation (K.K.).

No Patient and Public Involvement

This research was done without patient involvement.

References

1. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005; **165:** 1493–1499.
2. Chakraborty S, Brijnath B, Dermentzis J, Mazza D. Defining key questions for clinical practice guidelines: A novel approach for developing clinically relevant questions. *Health Res Policy Syst* 2020; **18:** 113.
3. Kusunose K. Steps to use artificial intelligence in echocardiography. *J Echocardiogr* 2021; **19:** 21–27.
4. Kusunose K, Hirata Y, Yamaguchi N, Kosaka Y, Tsuji T, Kotoku J, et al. Deep learning for detection of exercise-induced pulmonary hypertension using chest X-ray images. *Front Cardiovasc Med* 2022; **9:** 891703.
5. Omori H, Kawase Y, Mizukami T, Tanigaki T, Hirata T, Okubo M, et al. Diagnostic accuracy of artificial intelligence-based angiography-derived fractional flow reserve using pressure wire-based fractional flow reserve as a reference. *Circ J* 2023; **87:** 783–790.
6. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. *Lancet Digital Health* 2023; **5:** e105–e106.
7. 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.
8. Münzel T, Hahad O, Sørensen M, Lelieveld J, Duerr GD, Nieuwenhuijsen M, et al. Environmental risk factors and cardiovascular diseases: A comprehensive expert review. *Cardiovasc Res* 2022; **118:** 2880–2902.
9. Hisamatsu T, Miura K. Epidemiology and control of hypertension in Japan: A comparison with Western countries. *J Hum Hypertens* 2021, doi:10.1038/s41371-021-00534-3.
10. Umemura S, Arima H, Arima S, Asayama K, Dohi Y, Hirooka Y, et al. The Japanese Society of Hypertension guidelines for the management of hypertension (JSH 2019). *Hypertens Res* 2019; **42:** 1235–1481.
11. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology* 2023; **307:** e230163.
12. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D. Threats by artificial intelligence to human health and human existence. *BMJ Glob Health* 2023; **8:** e010435.