

ChatGPT Responses to Clinical Questions in the Japan Atherosclerosis Society Guidelines for Prevention of Atherosclerotic Cardiovascular Disease 2022

Takashi Hisamatsu, Mari Fukuda, Minako Kinuta and Hideyuki Kanda

Department of Public Health, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

Aims: Artificial intelligence is increasingly used in the medical field. We assessed the accuracy and reproducibility of responses by ChatGPT to clinical questions (CQs) in the Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022 (JAS Guidelines 2022).

Methods: In June 2024, we assessed responses by ChatGPT (version 3.5) to CQs, including background questions (BQs) and foreground questions (FQs). Accuracy was assessed independently by three researchers using six-point Likert scales ranging from 1 (“completely incorrect”) to 6 (“completely correct”) by evaluating responses to CQs in Japanese or translated into English. For reproducibility assessment, responses to each CQ asked five times separately in a new chat were scored using six-point Likert scales, and Fleiss kappa coefficients were calculated.

Results: The median (25th–75th percentile) score for ChatGPT’s responses to BQs and FQs was 4 (3–5) and 5 (5–6) for Japanese CQs and 5 (3–6) and 6 (5–6) for English CQs, respectively. Response scores were higher for FQs than those for BQs (P values <0.001 for Japanese and English). Similar response accuracy levels were observed between Japanese and English CQs (P value 0.139 for BQs and 0.586 for FQs). Kappa coefficients for reproducibility were 0.76 for BQs and 0.90 for FQs.

Conclusions: ChatGPT showed high accuracy and reproducibility in responding to JAS Guidelines 2022 CQs, especially FQs. While ChatGPT primarily reflects existing guidelines, its strength could lie in rapidly organizing and presenting relevant information, thus supporting instant and more efficient guideline interpretation and aiding in medical decision-making.

See editorial vol. 32: 560-562

Key words: Autonomic intelligence, ChatGPT, Accuracy, Reproducibility, Guidelines

Introduction

Accurate interpretation of clinical guidelines by health care professionals is crucial in decision-making in the clinical setting for the management of patients¹⁾. However, this can be challenging for health care professionals who lack specialized expertise in a specific area²⁾. To facilitate interpretation of the guidelines, evidence-based questions and answers, called clinical questions (CQs), are useful for providing clear and specific guidance for health care

professionals³⁾.

The use of artificial intelligence (AI) in the medical field has developed rapidly in recent years with the emergence of applications and technologies to improve patient management and evidence-based medicine⁴⁾. In particular, large language models (LLMs), which use advanced deep learning techniques, emulate human-like interactions by effectively responding to natural language queries^{5, 6)}. The ChatGPT bot, which is a refined and accessible LLM, was launched for public use on November 2022 by

Address for correspondence: Takashi Hisamatsu, Department of Public Health, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, 2-5-1 Shikata-cho, Kita-ku, Okayama 700-8558 Japan E-mail: hisamatsu@okayama-u.ac.jp

Received: June 30, 2024 Accepted for publication: September 12, 2024

Copyright©2025 Japan Atherosclerosis Society

This article is distributed under the terms of the latest version of CC BY-NC-SA defined by the Creative Commons Attribution License.

Open AI (San Francisco, CA, USA). ChatGPT is based on a diverse set of texts available on the Internet, including books, articles, and websites. ChatGPT can provide comprehensive, fluent, multilingual, and knowledge-based responses or solutions in various fields, including medicine^{6, 7)}. However, few studies have examined whether LLMs such as ChatGPT can support health care professionals in interpreting clinical guidelines²⁾.

Atherosclerotic cardiovascular disease (ASCVD) is a major cause of mortality and disability in Japan, and its prevention is of great clinical and public health importance. The Japan Atherosclerosis Society (JAS) Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022 (JAS Guidelines 2022) aim to guide health care professionals in the comprehensive management and appropriate treatment of risk factors for the prevention of ASCVD⁸⁾.

Aims

In the present study, we examined the accuracy and reproducibility of responses given by ChatGPT to CQs in the JAS Guidelines 2022 to determine whether AI can aid health care professionals in the interpretation of the guidelines.

Methods

Study Design and Questions

In June 2024, we assessed the responses by ChatGPT-3.5 (GPT version 3.5, OpenAI) to the CQs, designated as background questions (BQs) and foreground questions (FQs), in the JAS Guidelines 2022⁸⁾. The list of BQs and FQs in the JAS Guidelines 2022 is shown in [Supplementary Table 1](#). The BQs comprise 28 questions regarding background information on topics such as the incidence, symptoms, and natural course of diseases and are mainly based on epidemiological observational studies. The FQs comprise 27 questions related to decision-making in clinical practice regarding the choice of medical treatment. Among FQs, clinical issues for which multiple options exist in the current medical process and the balance of benefits and harm is unclear, and for which a recommendation is expected to improve patient outcomes, are considered important clinical issues. Basically, using the FQs, PICO (Patient, Intervention, Comparison, Outcome) can be established, and recommendations are mainly based on randomized controlled trials.

Ethical review was not required for the present study because our research did not use humans,

animals, or any related information.

Assessment of Responses by ChatGPT

To generate responses, CQs (BQs and FQs) originally published in Japanese were entered as queries in GPT-3.5. Assessment of the accuracy of ChatGPT's responses to the CQs was conducted independently by three expert researchers: two physicians (T.H. and H.K.) and one epidemiologist (M.F.) According to an assessment of ChatGPT's responses to medical questionnaires in a prior study⁷⁾, the accuracy of the provided answers was measured using a six-point Likert scale, as follows: (1) completely incorrect; (2) more incorrect than correct; (3) nearly equally correct and incorrect; (4) more correct than incorrect; (5) nearly all correct; and (6) completely correct. No prior training was conducted among the researchers to standardize their evaluations. Next, we repeated the assessment of ChatGPT's response accuracy to the CQs translated into English. In the case of responses limited to Japanese individuals where the CQs did not specify this limitation (i.e., BQs 8–13) in the JAS guidelines 2022, we entered CQs that included the phrase “for Japanese individuals” or “in Japanese individuals” into ChatGPT-3.5. For the assessment of reproducibility, each CQ in Japanese was asked five times separately in a new chat, and the responses were assessed using a six-point Likert scale. At the end of the study period, scores on the six-point Likert scale were collected and statistically analyzed.

Statistical Analysis

For the analysis of ChatGPT's response accuracy based on six-point Likert scale scores ranging from 1 to 6 for answers to CQs (BQs or FQs) originally in Japanese or those translated into English, we calculated the median (25th–75th percentile) of scores according to assessments by the three researchers. The difference in response scores between BQs and FQs was statistically analyzed using the Mann–Whitney *U* test. Similarly, the difference in scores for responses to CQs between Japanese and English was assessed using the Mann–Whitney *U* test. For reproducibility assessment, the Fleiss kappa coefficient was calculated to test the agreement between scores for responses to CQs asked five times separately as follows: poor (Fleiss kappa coefficient <0.40), good (0.40 to 0.75), and excellent agreement (>0.75)⁹⁾. We also calculated Shannon entropy¹⁰⁾ to quantitatively assess the uncertainty in the distribution of the responses generated by ChatGPT. Shannon entropy measures the degree of uncertainty, with higher entropy values representing a greater range in the response

Table 1. Six-point Likert scale scores for responses to clinical questions in Japanese in the JAS Guidelines 2022 as assessed by three researchers

	Researchers				Researchers		
	A	B	C		A	B	C
BQ1	3	2	4	FQ1	6	6	5
BQ2	2	1	2	FQ2	4	4	5
BQ3	5	5	3	FQ3	6	5	5
BQ4	2	2	2	FQ4	6	5	5
BQ5	5	4	5	FQ5	1	1	1
BQ6	6	6	4	FQ6	5	4	6
BQ7	6	6	5	FQ7	6	6	5
BQ8	5	6	4	FQ8	6	6	6
BQ9	4	3	3	FQ9	6	5	6
BQ10	1	1	2	FQ10	6	5	6
BQ11	2	2	2	FQ11	6	6	6
BQ12	4	5	3	FQ12	6	6	4
BQ13	2	1	1	FQ13	6	5	6
BQ14	4	3	4	FQ14	6	5	5
BQ15	5	4	5	FQ15	6	5	5
BQ16	5	5	3	FQ16	4	3	4
BQ17	6	6	6	FQ17	4	4	5
BQ18	5	5	6	FQ18	6	6	5
BQ19	4	4	5	FQ19	6	6	6
BQ20	5	5	4	FQ20	5	6	5
BQ21	6	6	4	FQ21	6	5	5
BQ22	6	6	5	FQ22	5	5	5
BQ23	6	5	4	FQ23	6	6	6
BQ24	6	6	6	FQ24	5	6	5
BQ25	2	1	2	FQ25	6	6	5
BQ26	6	6	6	FQ26	4	4	3
BQ27	4	2	4	FQ27	6	6	6
BQ28	4	4	4				

The six-point Likert scale was defined as follows: (1) completely incorrect; (2) more incorrect than correct; (3) nearly equally correct and incorrect; (4) more correct than incorrect; (5) nearly all correct; and (6) completely correct.

Abbreviations: BQ, background question; FQ, foreground question; JAS Guidelines 2022, Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022.

distribution. The analyses were performed using a Stata, version 16.0 (StataCorp LLC, College Station, TX, USA). Two-tailed P values <0.05 were considered statistically significant.

Results

Table 1 presents six-point Likert scale scores for responses by ChatGPT to each BQ and FQ originally in Japanese. Examples of correct and incorrect responses by ChatGPT to CQs are presented in **Supplementary Table 2**. Overall, the median (25th–75th percentile) score according to assessment by the three researchers was 4 (3–5) and 5 (5–6), for BQs and FQs, respectively. The scores were higher for

responses to FQs than those for BQs in Japanese (P value <0.001). The six-point Likert scale scores for responses to each BQ and FQ translated into English are shown in **Table 2**. Overall, the median (25th–75th percentile) score from assessment by the three researchers was 5 (3–6) and 6 (5–6) for BQs and FQs, respectively. The scores were higher for responses to FQs than those for BQs in English (P value <0.001). Similar scores for responses to CQs were observed between Japanese and English (P value 0.139 for BQs and 0.586 for FQs). Six-point Likert scale scores for the responses to each BQ and FQ asked five times separately in a new chat are shown in **Table 3**. The kappa coefficient to test the agreement between scores for responses to CQs asked five times separately was

Table 2. Six-point Likert scale scores for responses to clinical questions translated into English in the JAS Guidelines 2022 as assessed by three researchers

	Researchers				Researchers		
	A	B	C		A	B	C
BQ1	4	5	3	FQ1	2	2	3
BQ2	4	3	3	FQ2	4	4	5
BQ3	5	5	5	FQ3	6	6	5
BQ4	5	4	3	FQ4	5	4	4
BQ5	5	5	5	FQ5	1	1	1
BQ6	6	6	5	FQ6	6	5	6
BQ7	6	6	5	FQ7	6	6	5
BQ8	3	2	3	FQ8	6	4	5
BQ9	2	1	2	FQ9	6	5	4
BQ10	1	1	2	FQ10	6	6	6
BQ11	4	3	3	FQ11	6	6	5
BQ12	4	5	3	FQ12	6	6	6
BQ13	2	3	2	FQ13	6	6	6
BQ14	5	4	4	FQ14	6	6	6
BQ15	6	5	5	FQ15	5	5	5
BQ16	5	4	4	FQ16	6	5	5
BQ17	6	6	6	FQ17	6	4	5
BQ18	6	6	6	FQ18	6	6	4
BQ19	6	6	4	FQ19	6	6	4
BQ20	5	3	4	FQ20	6	5	5
BQ21	6	6	6	FQ21	6	6	6
BQ22	6	6	5	FQ22	3	2	3
BQ23	5	5	5	FQ23	6	6	6
BQ24	6	6	6	FQ24	6	6	6
BQ25	5	6	5	FQ25	6	6	6
BQ26	6	6	5	FQ26	6	5	6
BQ27	4	2	4	FQ27	6	6	6
BQ28	4	4	4				

The six-point Likert scale was defined as follows: (1) completely incorrect; (2) more incorrect than correct; (3) nearly equally correct and incorrect; (4) more correct than incorrect; (5) nearly all correct; and (6) completely correct.

Abbreviations: BQ, background question; FQ, foreground question; JAS Guidelines 2022, Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022.

0.76 for BQs and 0.90 for FQs. The Shannon entropy values for BQs ranged from 0.0 to 1.371 and were predominantly 0.0 for FQs, except for a few FQs ([Supplementary Table 3](#)).

Discussion

In the responses by ChatGPT-3.5 to CQs in the JAS Guidelines 2022, the accuracy was relatively high for FQs compared with that for BQs. No significant differences in accuracy were observed between CQs originally published in Japanese and those translated into English. The reproducibility was also high between ChatGPT's responses to CQs asked five times separately in new chats (i.e., excellent agreement for

both BQs and FQs); however, the Fleiss kappa coefficient was greater for FQs than that for BQs. The higher Shannon entropy values observed for BQs indicate greater uncertainty and variability in the responses, whereas the predominantly 0.0 values for FQs suggest a higher consistency and lower uncertainty in the generated responses. One possible reason for the differences in accuracy and reproducibility of responses by ChatGPT-3.5 between FQs and BQs may be owing to the characteristics of the two types of question. FQs are related to decision-making in clinical practice regarding the choice of medical treatment, and the answers are generally based on evidence from randomized controlled trials⁸). BQs involve background information on the topic of

Table 3. Six-point Likert scale scores for responses to each clinical question in the JAS Guidelines 2022 asked five times separately in a new chat

	Times						Times				
	1	2	3	4	5		1	2	3	4	5
BQ1	3	3	3	4	3	FQ1	6	6	6	6	6
BQ2	2	2	1	2	2	FQ2	4	4	4	4	4
BQ3	5	5	5	5	5	FQ3	6	6	6	6	6
BQ4	2	2	3	2	2	FQ4	6	6	6	6	6
BQ5	5	5	5	5	5	FQ5	1	1	1	1	1
BQ6	6	6	6	4	6	FQ6	5	5	5	5	5
BQ7	6	6	6	6	6	FQ7	6	6	6	6	6
BQ8	5	5	5	5	4	FQ8	6	6	6	6	6
BQ9	4	4	3	4	4	FQ9	6	6	6	6	6
BQ10	1	1	2	2	1	FQ10	6	6	6	6	6
BQ11	2	3	2	2	2	FQ11	6	6	6	6	6
BQ12	4	4	4	4	4	FQ12	6	6	6	6	6
BQ13	2	2	2	2	2	FQ13	6	5	6	5	6
BQ14	4	4	4	4	4	FQ14	6	6	6	6	6
BQ15	5	5	5	5	5	FQ15	6	6	6	6	6
BQ16	5	5	5	5	5	FQ16	4	3	4	4	4
BQ17	6	6	6	6	6	FQ17	4	4	4	4	4
BQ18	5	5	5	4	5	FQ18	6	6	6	6	6
BQ19	4	4	4	4	5	FQ19	6	6	6	6	6
BQ20	5	5	5	5	5	FQ20	5	5	5	5	5
BQ21	6	6	6	6	6	FQ21	6	6	5	6	6
BQ22	6	6	6	6	6	FQ22	5	5	5	5	5
BQ23	6	5	6	6	6	FQ23	6	6	6	6	6
BQ24	6	6	6	6	6	FQ24	5	5	5	5	5
BQ25	2	2	2	2	2	FQ25	6	6	6	6	6
BQ26	6	6	6	6	6	FQ26	4	4	4	4	4
BQ27	4	4	2	3	4	FQ27	6	6	6	6	6
BQ28	4	4	4	4	4						

The six-point Likert scale was defined as follows: (1) completely incorrect; (2) more incorrect than correct; (3) nearly equally correct and incorrect; (4) more correct than incorrect; (5) nearly all correct; and (6) completely correct.

Abbreviations: BQ, background question; FQ, foreground question; JAS Guidelines 2022, Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022.

diseases (e.g., symptoms, prevalence/incidence, natural course), and the answers are mainly based on epidemiological observational findings⁸⁾.

Few studies have investigated the accuracy of responses by LLMs such as ChatGPT to CQs in the clinical guidelines²⁾. To our knowledge, only one study assessed the accuracy of responses by ChatGPT-3.5 to CQs and other questions from the Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019). That study found accuracy rates of 80% and 36% for the CQs and other questions, respectively²⁾. Similar to our results, the accuracy of responses by ChatGPT-3.5 in that study was higher for CQs related to clinical decision-making for the treatment of hypertension based on the

evidence from randomized controlled trials than for other questions. Additionally, some prior studies have reported the accuracy or reproducibility of ChatGPT's responses to medical questions but not related to clinical guidelines^{7, 11, 12)}.

When we defined an "incorrect answer" as ones where two or more researchers assigned a Likert scale score of 1 or 2, the following CQs were identified as incorrect: BQ2, BQ4, BQ10, BQ11, BQ13, BQ25, and FQ5 in Japanese, and BQ9, BQ10, BQ13, FQ1, and FQ5 in English. For questions related to blood cholesterol levels (BQ2, BQ4), medication adherence (BQ25), total energy intake (FQ1), and n-6 polyunsaturated fatty acids intake (FQ5), ChatGPT's inaccuracies may stem from difficulties in fully

grasping the complexity and context of these issues¹³). Specifically, when addressing certain diseases or medical conditions, ChatGPT tends to simplify health information and reduce the complexity of specialized content¹⁴), leading to responses that may not fully align with the intended context. For questions related to subclinical atherosclerotic markers (BQ9, BQ10, BQ11, BQ13) in Japanese individuals, the inaccuracies may be due to limitations in ChatGPT's training data regarding regional and racial specificity¹⁵). While extensive, the training data may lack detailed epidemiological information relevant to specific populations, such as the Japanese, resulting in responses that do not fully capture the nuances required for precise guidance per the JAS Guidelines 2022. The differences in the accuracy of ChatGPT's responses between Japanese and English, though not statistically significant, may be largely due to language biases and disparities in the information available in the training data for each language¹²).

When we defined an “uncertainty” as cases where cases where different answers were received in two or more of the five different chat sessions for the same question, uncertain answers were identified for the following CQs: BQ10, BQ27, and FQ13. A possible reason for these uncertain answers may be the diverse range of data used in ChatGPT's training^{16, 17}). Its vast dataset, drawn from various sources, can lead to different data points or perspectives being referenced, resulting in response variability^{16, 17}). For example, in the CQ about the prevalence of familial hypercholesterolemia in the Japanese population (BQ27), three different answers were observed across the five sessions, suggesting data inconsistencies in ChatGPT's sources. Overall, our findings underscore the need to carefully consider ChatGPT's potential uncertainty and inaccuracies.

For health care professionals, medical training and education are lifelong processes, and keeping up with the latest guidelines, techniques, and research can be challenging. In this respect, ChatGPT can provide instant access to pertinent and updated medical information and resources for health care professionals, resulting in improvement in their continuous learning and development and enhancement of their skills and knowledge^{18, 19}).

Our study focused on assessing the accuracy and reproducibility of ChatGPT's responses to CQs in existing guidelines (i.e., JAS Guidelines 2022). ChatGPT could potentially assist guideline developers by generating preliminary recommendations or synthesizing evidence from available research data, thus helping streamline the guideline creation process. Moreover, by incorporating input from multiple

experts, ChatGPT could aid in the interpretation of evidence and the formulation of consensus-based recommendations. However, it is important to acknowledge that clinical guidelines development requires expert judgment and clinical insight, and AI tools like ChatGPT should be seen as complementary rather than replacing human expertise²⁰). Therefore, we suggest that ChatGPT be considered as an adjunct to, rather than a replacement for, expert-driven processes in clinical guidelines development.

However, the limitations of ChatGPT should be considered when using this LLM. The accuracy of ChatGPT depends on the quality of its training data, which are not publicly disclosed and likely have a limited focus on medicine¹⁸). The undisclosed data may include inaccuracies, resulting in the assimilation of misinformation, unbalanced content, and biases during the LLM training process²¹). In fact, ChatGPT-3.5 was trained on data up to September 2021 and is unable to integrate data published after model training⁶). We did not use ChatGPT-4.0 in the present study owing to its high cost. Another concern regarding ChatGPT is that it has the tendency to produce “hallucinations” or plausible but factually incorrect information⁶). For example, in our study, ChatGPT generated inaccurate information (“hallucinations”) regarding the predictive ability of the ankle-brachial index for the Japanese population in the response to the CQ (BQ13) ([Supplementary Table 2](#)). This issue likely arises from limitations in ChatGPT's training data, which may lack specific epidemiological details for certain populations, such as the Japanese¹⁵), leading it to rely on general knowledge or data from other populations. At this time, it is essential to recognize and operate within the limitations of existing models. When used with a clear understanding of its capabilities and limitations, ChatGPT and other AI applications can serve as valuable tools.

The present study had several limitations. First, we focused on the JAS Guidelines 2022 for the comprehensive management and appropriate treatment of risk factors in the prevention of ASCVD. Therefore, further studies are needed to investigate the accuracy and reproducibility of ChatGPT in responding to CQs in other medical fields and clinical guidelines. Second, no prior training was conducted among the researchers to standardize their evaluations of ChatGPT's responses. This could have introduced variability in the assessment process, as differences in individual interpretation may have affected the consistency of the accuracy scores assigned to the responses. Third, the version of ChatGPT used will be trained on data in the future, which may result in

changing the accuracy of its responses.

In conclusion, we found that ChatGPT had relatively high accuracy and high reproducibility in its responses to CQs, especially FQs related to decision-making in clinical practice about the choice of medical treatment, per the JAS Guidelines 2022. While we observed some discrepancies in the accuracy and reproducibility of responses across different types of CQs, our findings highlight ChatGPT's potential to efficiently organize and present relevant guideline information, thereby potentially reducing the cognitive burden on healthcare professionals. Rather than introducing new information, ChatGPT could serve as a tool that facilitates quicker access to and interpretation of existing guidelines, ultimately supporting informed decision-making in clinical practice.

However, significant challenges remain, including reliance on undisclosed training data²¹, the risk of generating factually incorrect information ("hallucinations")⁶, and limitations in addressing regional, racial, and language-specific nuances^{12, 15}. These issues suggest that while ChatGPT can be a valuable tool, it should complement rather than replace human expertise in clinical settings²⁰. We also emphasize the need for more advanced models trained on high-quality medical data² and the establishment of standards to evaluate AI-generated information. Our findings indicate that with further research and improvements, AI technologies like ChatGPT could have the potential to play a significant role in enhancing clinical practice.

Acknowledgement

We thank Analisa Avila, MPH, ELS, of Edanz (<https://jp.edanz.com/ac>) for editing a draft of this manuscript.

Conflict of Interest:

None.

Funding

None.

References

- 1) Graber ML, Franklin N and Gordon R: Diagnostic error in internal medicine. *Arch Intern Med*, 2005; 165: 1493-1499
- 2) Kusunose K, Kashima S and Sata M: Evaluation of the Accuracy of ChatGPT in Answering Clinical Questions on the Japanese Society of Hypertension Guidelines. *Circ J*, 2023; 87: 1030-1033
- 3) Chakraborty S, Brijnath B, Dermentzis J and Mazza D: Defining key questions for clinical practice guidelines: a novel approach for developing clinically relevant questions. *Health Res Policy Syst*, 2020; 18: 113
- 4) Bohr A, Memarzadeh K: The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare*. 2020; 25-60. <https://doi.org/10.1016/b978-0-12-818438-7.00002-2>
- 5) Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, Baabdullah AM, Koohang A, Raghavan V, Ahuja M, Albanna H, Albashrawi MA, Al-Busaidi AS, Balakrishnan J, Barlette Y, Basu S, Bose I, Brooks L, Buhalis D, Carter L, Chowdhury S, Crick T, Cunningham SW, Davies GH, Davison RM, De RH, Dennehy D, Duan YQ, Dubey R, Dwivedi R, Edwards JS, Flavian C, Gauld R, Grover V, Hu MC, Janssen M, Jones P, Junglas I, Khorana S, Kraus S, Larsen KR, Latreille P, Laumer S, Malik FT, Mardani A, Mariani M, Mithas S, Mogaji E, Nord JH, O'Connor S, Okumus F, Pagani M, Pandey N, Papagiannidis S, Pappas IO, Pathak N, Pries-Heje J, Raman R, Rana NP, Rehm SV, Ribeiro-Navarrete S, Richter A, Rowe F, Sarker S, Stahl BC, Tiwari MK, van der Aalst W, Venkatesh V, Viglia G, Wade M, Walton P, Wirtz J and Wright R: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inform Manage*, 2023; 71: 102642
- 6) Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G and Moy L: ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*, 2023; 307: e230163
- 7) Niko MM, Karbasi Z, Kazemi M and Zahmatkeshan M: Comparing ChatGPT and Bing, in response to the Home Blood Pressure Monitoring (HBPM) knowledge checklist. *Hypertens Res*, 2024; 47: 1401-1409
- 8) Okamura T, Tsukamoto K, Arai H, Fujioka Y, Ishigaki Y, Koba S, Ohmura H, Shoji T, Yokote K, Yoshida H, Yoshida M, Deguchi J, Dobashi K, Fujiyoshi A, Hamaguchi H, Hara M, Harada-Shiba M, Hirata T, Iida M, Ikeda Y, Ishibashi S, Kanda H, Kihara S, Kitagawa K, Kodama S, Koseki M, Maezawa Y, Masuda D, Miida T, Miyamoto Y, Nishimura R, Node K, Noguchi M, Ohishi M, Saito I, Sawada S, Sone H, Takemoto M, Wakatsuki A and Yanai H: Japan Atherosclerosis Society (JAS) Guidelines for Prevention of Atherosclerotic Cardiovascular Diseases 2022. *J Atheroscler Thromb*, 2024; 31: 641-853
- 9) Fleiss JL: Measuring nominal scale agreement among many raters. *Psychol Bull*, 1971; 76: 378-382
- 10) Shannon CE: A mathematical theory of communication. *Bell Syst Tech J*, 1948; 27: 379-423
- 11) Sarraju A, Bruemmer D, Van Itermen E, Cho L, Rodriguez F and Laffin L: Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*, 2023; 329: 842-844
- 12) Yano Y, Nishiyama A, Suzuki Y, Morimoto S, Morikawa T, Gohda T, Kanegae H and Nakashima N: Relevance of

- ChatGPT's Responses to Common Hypertension-Related Patient Inquiries. *Hypertension*, 2024; 81: e1-e4
- 13) Nazir A and Wang Z: A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta Radiol*, 2023; 1: 100022
- 14) Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, Rezwani A and Dunn AG: New Frontiers in Health Literacy: Using ChatGPT to Simplify Health Information for People in the Community. *J Gen Intern Med*, 2024; 39: 573-577
- 15) Peters U: Algorithmic Political Bias in Artificial Intelligence Systems. *Philos Technol*, 2022; 35: 25
- 16) Kochanek K, Skarzynski H and Jedrzejczak WW: Accuracy and Repeatability of ChatGPT Based on a Set of Multiple-Choice Questions on Objective Tests of Hearing. *Cureus*, 2024; 16: e59857
- 17) Jedrzejczak WW, Skarzynski PH, Raj-Koziak D, Sanfins MD, Hatzopoulos S and Kochanek K: ChatGPT for Tinnitus Information and Support: Response Accuracy and Retest after Three and Six Months. *Brain Sci*, 2024; 14: 465
- 18) Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, Qiu Y, Tang S, Car J, Wang YX, Wong TY, Tham YC and Chung KC: ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac*, 2023; 41: 100905
- 19) Moons P and 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
- 20) Rajpurkar P, Chen E, Banerjee O and Topol EJ: AI in health and medicine. *Nat Med*, 2022; 28: 31-38
- 21) Harrer S: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 2023; 90: 104512

Supplementary Table 1. List of background questions (BQs) and foreground questions (FQs) in the JAS Guidelines 2022

BQ1	Does LDL cholesterol predict the incidence and mortality of ASCVD in the Japanese population?
BQ2	Does total cholesterol (TC) predict the incidence and mortality of ASCVD in the Japanese population?
BQ3	Does non-HDL cholesterol predict the incidence and mortality of ASCVD in the Japanese population?
BQ4	Does HDL cholesterol predict the incidence and mortality of ASCVD in the Japanese population?
BQ5	Does triglyceride (TG) predict the incidence and mortality of ASCVD in the Japanese population?
BQ6	Are blood glucose and HbA1c associated with the CVD incidence and/or death from CVD in non-diabetic adults?
BQ7	Do familial hypercholesterolemia, non-cardiogenic cerebral infarction (cardiogenic cerebral embolism), PAD, microvascular complications, smoking, and persistent poor glycemic control increase the risk of CAD in patients with diabetes without a history of CAD?
BQ8	Do asymptomatic vascular lesions on brain MRI have a predictive power for the incidence of ASCVD beyond the clustering of classical risk factors (or a prediction model)?
BQ9	Do intima-media thickness or plaque (IMT/plaque) findings by carotid artery ultrasonography have a predictive power for the incidence of ASCVD beyond the clustering of classical risk factors (or a prediction model)?
BQ10	Do coronary stenosis and CAC score using coronary CT have predictive ability for the incidence of ASCVD beyond the clustering of classical risk factors (or a prediction model)?
BQ11	Does PWV have predictive ability for the incidence of ASCVD beyond the clustering of classical risk factors (or a prediction model)?
BQ12	Does CAVI have the predictive ability for the incidence of ASCVD beyond the clustering of classical risk factors (or a prediction model)?
BQ13	Does ABI have predictive ability for the incidence of ASCVD beyond the clustering of classical risk factors (or a predictive model)?
BQ14	What dyslipidemia is associated with MASLD/MASH*? *Originally published as NAFLD/NASH: now described as MASLD (metabolic dysfunction-associated steatotic liver disease) / MASH (metabolic dysfunction-associated steatohepatitis)
BQ15	Is MASLD/MASH a high-risk condition for ASCVD?
BQ16	Are there any evaluation methods to predict the onset and death of ASCVD in Japanese?
BQ17	Do aerobic exercise and physical activity reduce the incidence of ASCVD in adults?
BQ18	Does resistance exercise prevent the incidence of ASCVD in adults?
BQ19	Does reducing sedentary time prevent the incidence of ASCVD in adults?
BQ20	Have indications, efficacy, and safety of drugs for dyslipidemia been established?
BQ21	Is it necessary to regularly perform clinical examinations after the initiation of drug therapy?
BQ22	Does the concomitant use of statins with drugs metabolized by CYPs increase the incidence of adverse effects?
BQ23	Can the use of fixed-dose combination drugs in the treatment of dyslipidemia be recommended for the prevention of serum lipids and the incidence of ASCVD?
BQ24	Is medication adherence related to serum lipid levels and the incidence of ASCVD?
BQ25	What factors influence drug adherence?
BQ26	Is lipid management by setting a control target (or a goal for cholesterol-lowering rate) effective in preventing ASCVD?
BQ27	What is the prevalence of Familial Hypercholesterolemia (FH) in Japan?
BQ28	What are the prognosis and main complications of patients with FH?
FQ1	Is limiting total energy intake and maintaining an appropriate body weight effective in preventing ASCVD?
FQ2	For the prevention of ASCVD incidence, is it recommended to maintain an adequate fat energy ratio for the Japanese under an appropriate total energy intake?
FQ3	For the prevention of ASCVD incidence, is it recommended to reduce saturated fatty acids or replace saturated fatty acid intake with other unsaturated fatty acids (monounsaturated and polyunsaturated fatty acids) with an appropriate total energy intake?
FQ4	For the prevention of ASCVD incidence, is it recommended to increase the intake of n-3 polyunsaturated fatty acids?
FQ5	For the prevention of ASCVD incidence, is it recommended to increase the intake of n-6 polyunsaturated fatty acids?
FQ6	For the prevention of ASCVD incidence, is it recommended to increase the intake of monounsaturated fatty acids?

(Cont. Supplementary Table 1)

FQ7	For the prevention of ASCVD incidence, is it recommended to limit trans fatty acids?
FQ8	For the prevention of ASCVD incidence, is it recommended to limit cholesterol intake?
FQ9	For the prevention of ASCVD incidence, is it recommended to increase dietary fiber intake?
FQ10	For the prevention of the incidence of ASCVD, is it recommended to reduce the intake of processed foods containing fructose?
FQ11	Is Japanese dietary pattern recommended for the prevention of ASCVD?
FQ12	Is aerobic exercise recommended for adults to improve serum lipids?
FQ13	Is resistance exercise recommended for adults to improve serum lipids?
FQ14	Is exercise therapy recommended in addition to the diet for adults to improve serum lipids?
FQ15	For the purpose of improving obesity and dyslipidemia, should health counseling based on health behavior theory be recommended over general health counseling?
FQ16	Can LDL cholesterol-lowering therapy aimed at control targets be recommended for the prevention of ASCVD in Japanese patients?
FQ17	Is drug therapy for hypertriglyceridemia recommended for the prevention of ASCVD?
FQ18	Maximal tolerated dose strong statins are recommended as first choice in drug therapy for the secondary prevention of CAD?
FQ19	Is drug therapy targeting LDL cholesterol below 70 mg / dL for the secondary prevention of CAD associated with high-risk conditions recommended?
FQ20	Is LDL cholesterol lowering therapy with drugs other than statins recommended for the prevention of ASCVD?
FQ21	Is the addition of cholesterol-lowering non-statin drugs (ezetimibe, anion exchangers, probucol, PCSK9 inhibitors) to statins recommended for the prevention of recurrent ASCVD?
FQ22	In patients with hypertriglyceridemia or hypo-HDL cholesterolemia, is the co-administration of fibrates, SPPARM α , nicotinic acid derivatives, or n-3 polyunsaturated fatty acids with statins recommended for the prevention of the incidence of ASCVD?
FQ23	Is comprehensive strict control of blood glucose, lipid, and blood pressure recommended for patients with diabetes from the early stage?
FQ24	Is strict LDL cholesterol management recommended in patients with diabetes complicated by PAD, microangiopathy (retinopathy, nephropathy, neuropathy), or in the presence of smoking, along with management of other risk factors?
FQ25	Can statins be recommended as the first choice in drug therapy for FH heterozygotes (HeFH)?
FQ26	Is lipoprotein apheresis therapy recommended for homozygous FH (HoFH) and severe HeFH with drug resistance?
FQ27	Is it recommended to start treatment early in pediatric patients with FH?

Abbreviations: ABI, ankle-brachial index; ASCVD, atherosclerotic cardiovascular disease; CAC, coronary artery calcification; CAD, coronary artery disease; CAVI, cardio-ankle vascular index; CT, computed tomography; CVD, cardiovascular disease; CYP, Cytochrome P450; HbA1c, hemoglobin A1c; HDL, high-density lipoprotein; LDL, low-density lipoprotein; MRI, magnetic resonance imaging; NAFLD, Nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis; PAD, peripheral arterial disease; PCSK9, Proprotein convertase subtilisin/kexin type 9; SPPARM α , Selective Peroxisome Proliferator-Activated Receptor Alpha Modulators; PWV, pulse wave velocity

Supplementary Table 2. Examples of correct and incorrect responses by ChatGPT to CQs in the JAS Guidelines 2022

CQs in Japanese		
Correct response	FQ27	小児 FH 患者において早期から治療を開始することは推奨されるか？
	Answer in the guidelines	FH は動脈硬化性疾患のハイリスク病態であり、LDL コレステロール値に応じて、早期からの治療開始を推奨する。
	Response by ChatGPT	小児期の家族性高コレステロール血症（FH）は、心血管疾患のリスクを高める可能性がある遺伝性の疾患です。早期から治療を開始することは、重要な推奨事項の一つです。なぜなら、高コレステロールは、若い年齢から動脈硬化の進行を促進する可能性があり、心臓病やその他の心血管合併症のリスクを増加させるからです。 小児 FH の治療目標は、高コレステロール値をコントロールし、心血管疾患のリスクを最小限に抑えることです。これには、適切な食事療法や運動、必要に応じて薬物療法が含まれます。特に、LDL コレステロール（悪玉コレステロール）を正常範囲内に保つことが重要です。 小児期から治療を開始することで、心血管イベントのリスクを減少させるだけでなく、成長期における健康的な生活習慣を育成することも重要です。したがって、小児 FH の患者においては、早期から治療を開始し、継続的なモニタリングと管理が推奨されます。
Incorrect response	BQ13	ABI は古典的危険因子の集積（予測モデル）に上乗せした動脈硬化性疾患の発症予測能を有するか？
	Answer in the guidelines	既往のない日本人集団において ABI が古典的危険因子の集積を超えた循環器疾患発症予測能を有するか検討した研究は一報のみであり、それによると有意な予測能の改善を認めなかった。日本人集団の一次予防を目的として ABI を行う科学的根拠は乏しい。
	Response by ChatGPT	日本人において、ABI（足首腕指数）は古典的危険因子の集積（予測モデル）に上乗せした動脈硬化性疾患の発症予測能を有することが示されています。具体的には、ABI が低値である場合、動脈硬化性疾患のリスクが高まることが知られており、従来のリスク因子（例えば年齢、性別、高血圧、糖尿病、喫煙など）に ABI を加えることで、より精度の高い予測が可能となります。 複数の研究において、ABI が低下している患者は、従来のリスク因子のみで予測されたリスクよりも高い動脈硬化性疾患の発症リスクを有することが報告されています。したがって、日本人においても、ABI をリスク予測モデルに組み込むことで、動脈硬化性疾患の発症リスクをより正確に評価できる可能性が高いです。
CQs in English		
Correct response	BQ17	Do aerobic exercise and physical activity reduce the incidence of ASCVD in adults?
	Answer in the guidelines	In adults, aerobic exercise and increased physical activity can be expected to prevent ASCVD and are therefore recommended.
	Response by ChatGPT	Yes, engaging in regular aerobic exercise and physical activity can decrease the incidence of arteriosclerotic cardiovascular disease in adults. Arteriosclerosis is the thickening and hardening of the arteries due to a buildup of plaque, which can lead to conditions like coronary artery disease, stroke, and peripheral artery disease. Regular aerobic exercise helps to improve cardiovascular health by: - Reducing LDL cholesterol: Aerobic exercise can lower levels of low-density lipoprotein (LDL) cholesterol, which is often referred to as "bad" cholesterol. High levels of LDL cholesterol can contribute to the buildup of plaque in the arteries. - Increasing HDL cholesterol: Aerobic exercise can also increase levels of high-density lipoprotein (HDL) cholesterol, which is known as "good" cholesterol. HDL helps to remove LDL cholesterol from the bloodstream, reducing the risk of plaque buildup. - Lowering blood pressure: Regular aerobic exercise can help to lower blood pressure, reducing the strain on the arteries and lowering the risk of arteriosclerosis. - Improving vascular function: Exercise helps to improve the function of the blood vessels, promoting better blood flow and reducing the risk of arterial damage. - Promoting weight loss: Aerobic exercise can help to control weight and reduce excess body fat, which is a risk factor for arteriosclerosis. - Reducing inflammation: Regular exercise can help to reduce inflammation in the body, which is a key factor in the development of arteriosclerosis. Overall, incorporating regular aerobic exercise and physical activity into one's routine can significantly reduce the risk of developing arteriosclerotic cardiovascular disease and improve overall cardiovascular health. However, it's essential to consult with a healthcare professional before starting any new exercise regimen, especially if you have pre-existing health conditions.

(Cont. Supplementary Table 2)

Incorrect response	FQ5	For the prevention of ASCVD incidence, is it recommended to increase the intake of n-6 polyunsaturated fatty acids?
	Answer in the guidelines	•To improve serum lipids, it is recommended to increase the intake of n-6 polyunsaturated fatty acids or replace saturated fatty acids with n-6 polyunsaturated fatty acids in an appropriate total energy intake. •It is suggested to replace saturated fatty acids with n-6 polyunsaturated fatty acids, especially linoleic acid, under an appropriate total energy intake, as it is expected to prevent ASCVD.
	Response by ChatGPT	The recommendation to increase the intake of n-6 polyunsaturated fatty acids (PUFAs) for the prevention of atherosclerotic cardiovascular diseases is complex and cannot be simply answered with a yes or no. Achieving a proper balance between n-6 and n-3 PUFAs is crucial. Generally, n-6 PUFAs are associated with promoting inflammatory responses, while n-3 PUFAs are known to have anti-inflammatory effects. For the prevention of atherosclerotic cardiovascular diseases, a balanced diet that includes both types of PUFAs in appropriate proportions is essential. However, the evidence supporting the direct effect of increasing n-6 PUFA intake on the prevention of atherosclerotic cardiovascular diseases may not be sufficiently conclusive.

Abbreviations: BQ, background question; CQ, clinical question; FQ, foreground question; JAS Guidelines 2022, Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022.

Supplementary Table 3. Shannon entropy values for the distributions of the responses to each clinical question in the JAS Guidelines 2022 asked five times separately in a new chat

BQ1	0.722	FQ1	0.000
BQ2	0.722	FQ2	0.000
BQ3	0.000	FQ3	0.000
BQ4	0.722	FQ4	0.000
BQ5	0.000	FQ5	0.000
BQ6	0.722	FQ6	0.000
BQ7	0.000	FQ7	0.000
BQ8	0.722	FQ8	0.000
BQ9	0.722	FQ9	0.000
BQ10	0.971	FQ10	0.000
BQ11	0.722	FQ11	0.000
BQ12	0.000	FQ12	0.000
BQ13	0.000	FQ13	0.971
BQ14	0.000	FQ14	0.000
BQ15	0.000	FQ15	0.000
BQ16	0.000	FQ16	0.722
BQ17	0.000	FQ17	0.000
BQ18	0.722	FQ18	0.000
BQ19	0.722	FQ19	0.000
BQ20	0.000	FQ20	0.000
BQ21	0.000	FQ21	0.722
BQ22	0.000	FQ22	0.000
BQ23	0.722	FQ23	0.000
BQ24	0.000	FQ24	0.000
BQ25	0.000	FQ25	0.000
BQ26	0.000	FQ26	0.000
BQ27	1.371	FQ27	0.000
BQ28	0.000		

Abbreviations: BQ, background question; FQ, foreground question; JAS Guidelines 2022, Japan Atherosclerosis Society Guidelines for Prevention Atherosclerotic Cardiovascular Diseases 2022.