



# Evolution of a Large Language Model for Preoperative Assessment Based on the Japanese Circulation Society 2022 Guideline on Perioperative Cardiovascular Assessment and Management for Non-Cardiac Surgery

Takahiro Kamihara, MD, PhD; Masanori Tabuchi; Takuya Omura, MD, PhD;  
Yumi Suzuki, MD, PhD; Tsukasa Aritake, MD; Akihiro Hirashiki, MD, PhD;  
Manabu Kokubo, MD, PhD; Atsuya Shimizu, MD, PhD

**Background:** The Japanese Circulation Society 2022 Guideline on Perioperative Cardiovascular Assessment and Management for Non-Cardiac Surgery standardizes preoperative cardiovascular assessments. The present study investigated the efficacy of a large language model (LLM) in providing accurate responses meeting the JCS 2022 Guideline.

**Methods and Results:** Data on consultation requests, physicians' cardiovascular records, and patients' response content were analyzed. Virtual scenarios were created using real-world clinical data, and a LLM was then consulted for such scenarios.

**Conclusions:** Google BARD could accurately provide responses in accordance with the JCS 2022 Guideline in low-risk cases. Google Gemini has significantly improved its accuracy in intermediate- and high-risk cases.

**Key Words:** Guidelines; Large language model; Virtual scenarios

The aging population has resulted in an increase in the number of surgeries among older individuals. However, older patients have a higher surgical risk than younger ones. Therefore, surgeons consult cardiologists before performing surgeries on older individuals. The Japanese Circulation Society introduced the Japanese Circulation Society 2022 Guideline on Perioperative Cardiovascular Assessment and Management for Non-Cardiac Surgery in 2022 (JCS 2022 Guideline).<sup>1</sup> The Japanese version was published on March 11, 2022, and the English version was released on August 22, 2023. Standardization of medical care is an important issue, and medical care should be provided based on treatment guidelines. The JCS 2022 Guideline shows the importance of decision-making in the preoperative medical team, which includes surgeons and cardiologists. However, several hospitals do not have cardiologists. Several studies have examined the usefulness of large language models (LLMs) in various fields.<sup>2–11</sup> However, no study has validated the efficacy of LLMs in providing responses about preoperative evaluation meeting the guidelines.

Therefore, the present study aimed to evaluate the efficacy of a LLM in providing accurate responses meeting the JCS 2022 Guideline. This research used Google BARD and Google Gemini, which are LLMs that can be used by anyone for free and refer to the latest guidelines on the Internet. The Google BARD platform has been updated to Google Gemini. An evaluation of the reliability of LLMs was conducted before and after the update in February 2024.

## Methods

The ethics committee of the National Center for Geriatrics and Gerontology approved this study (no. 1668-2). Further, the External Service Usage Manager of the Department of Cardiology at the National Center for Geriatrics and Gerontology (for external services that do not handle confidential information) was approved by the external service usage permission application. We analyzed data on consultation requests, cardiovascular records of the physicians, and response content of the patients in which surgery

Received March 4, 2024; accepted March 4, 2024; J-STAGE Advance Publication released online March 15, 2024 Time for primary review: 1 days

Department of Cardiology (T.K., A.H., M.K., A.S.), Department of Nursing (M.T.), Department of Metabolism (T.O.), Department of Surgery (Y.S., T.A.), National Center for Geriatrics and Gerontology, Obu, Japan

The first two authors contributed equally to this work (T.K., M.T.).

Mailing address: Takahiro Kamihara, MD, PhD, Department of Cardiology, National Center for Geriatrics and Gerontology, 7-430 Morioka-cho, Obu, Aichi 474-8511, Japan. email: kamihara@ncgg.go.jp

All rights are reserved to the Japanese Circulation Society. For permissions, please email: cr@j-circ.or.jp

ISSN-2434-0790



was performed from October 1, 2022, to December 31, 2022. The High Care Unit was considered important for management after surgery by the surgical department. Further, the request and response contents were examined using text mining with KH Coder 3.Beta.07f.<sup>12-14</sup> Because the original clinical data were written in Japanese, the analysis up to this point was performed in Japanese. Based on the results of the analysis of the request content, virtual scenarios were created in English, and the patient in these scenarios was consulted with both Google BARD and Google Gemini in English. The responses created by Google BARD and Google Gemini were evaluated by a cardiovascular specialist in terms of 'comprehensibility', 'appropriateness', 'absence of relevant content', 'confabulation', and 'clinical decisions'. These evaluation points were used in a previous cardiology study.<sup>5</sup> **Figure 1A** shows the overall research and evaluation methods.

## Results

**Figure 1B** depicts a self-organizing map of consultation requests from surgeons for preoperative consultations. At the top of the figure, the following three clusters were created: 'abdominal surgery', 'anesthesia type', and 'heart failure (HF)'. The clusters for 'past medical history', 'chest pain', and 'acute myocardial infarction (AMI)/angina pectoris (AP)' were adjacent to these clusters. This suggests that surgeons in the digestive surgery department were more likely to place more emphasis on these items. The cluster for 'echocardiogram', 'hypertension (HT)', and 'blood pressure (BP)' was adjacent to the cluster for 'orthopedic surgery'. Both surgeons and orthopedic surgeons might emphasize 'arrhythmia', 'ultrasound cardiography (UCG)', and 'electrocardiogram (ECG)'. **Figure 1C** shows a self-organizing map of the responses from the cardiovascular department. The word 'guideline' was located in the upper right corner, surrounded by words related to the 'revised Cardiac Risk Index (RCRI)' and 'risk'. Other terms were listed outside these clusters. **Figure 1D,E** shows the results of the response analysis based on the surgical risk itself. **Figure 1D** presents the results of the response analysis of the consultation comments to the cardiovascular department. **Figure 1E** depicts the results of the response analysis from the cardiovascular department. **Figure 2A** shows the clinical characteristics of the patients by plotting sex, age, and left ventricular ejection fraction (LVEF). Based on the clinical characteristics of the patients, as shown in **Figure 2A**, and **Figure 1B,D**, we created virtual cases and consulted Google BARD and Google Gemini (**Figure 2B**). In case 1, the patient was young and had a relatively high LVEF. Hence, the patient generally belonged to the low-risk group in the graph, as shown in **Figure 2A**. The patient had few significant medical histories but had no history of medication use. In case 2, the patient had an average age and LVEF. Thus, the patient belonged to the intermediate-risk group, based on the graph shown in **Figure 2A**. The patient presented with HT, which was treated with olmesartan. In case 3, the patient was old and had a low LVEF. Therefore, the patient belonged to the high-risk group, based on the graph shown in **Figure 2A**. The patient had a history of myocardial infarction and had been taking several medications. The validity of the responses of Google BARD and Google Gemini for each case was evaluated by a cardiovascular specialist. **Figure 3** shows the summary of the virtual cases. **Figure 4** presents a concise

summary of the study results.

## Discussion

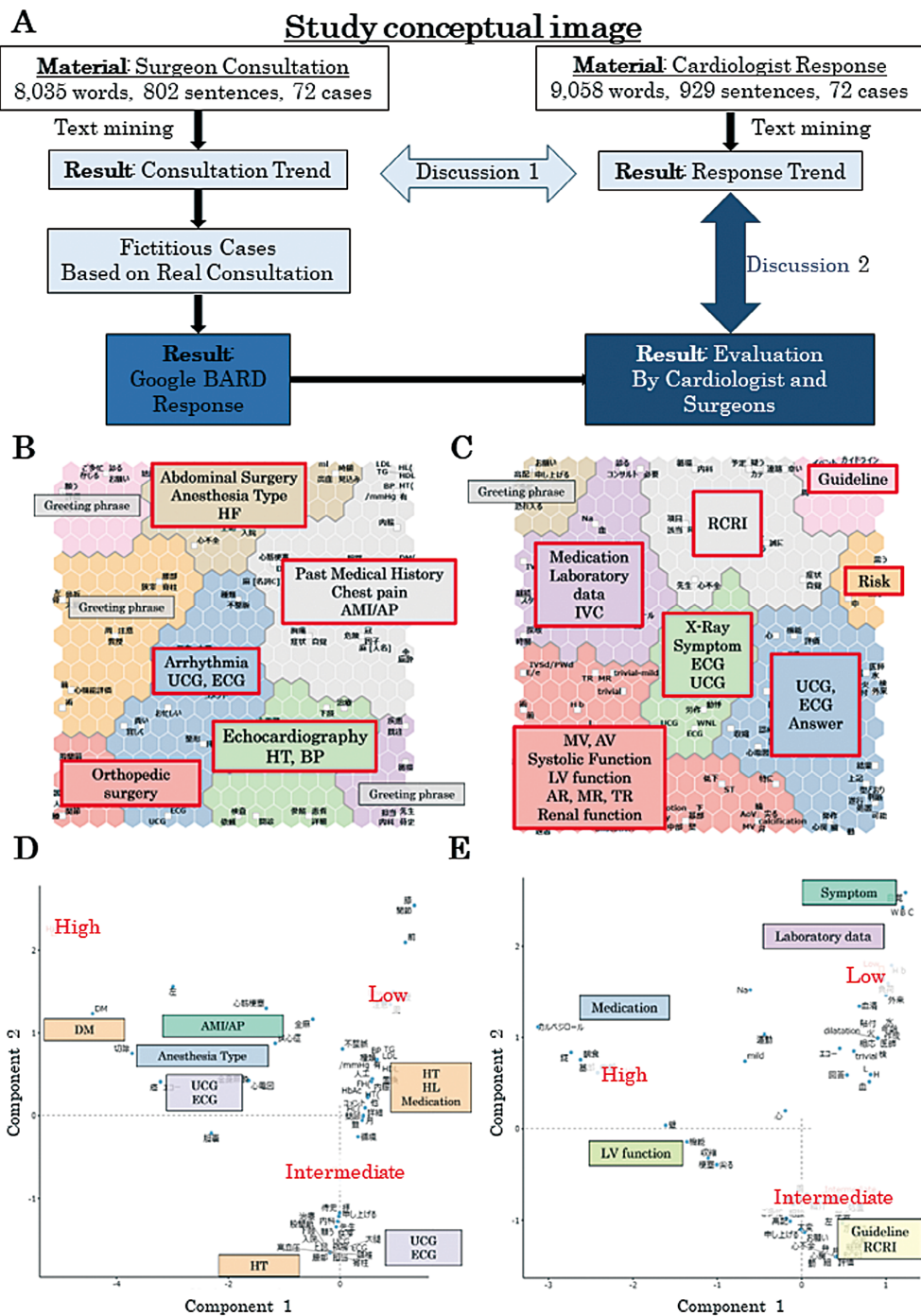
The present study aimed to assess the efficacy of LLM in accurately generating responses to preoperative consultations consistent with the JCS 2022 Guideline, which was published in English on August 22, 2023. Google BARD and Google Gemini are the only LLM that can generate responses based on real-time information and can be used for free. Hence, rather than ChatGPT, it was utilized in this research.

As shown in **Figure 1A**, the consultation content from surgical departments (**Figure 1B**), and the medical records and response content from the cardiovascular department (**Figure 1C**) were compared. Results showed that the block for abdominal surgery in the upper right corner of **Figure 1B** was connected to medical history and other items. However, in the case of requests from orthopedic surgery, it was connected to BP and HT. Based on these results, abdominal surgery focused on medical history. Further, orthopedic surgeons often perform surgeries on fractures. Hence, they are focused on BP increases caused by pain. As the surgical risk increased, the surgical departments were more likely to mention diabetes (**Figure 1D**). In contrast, as shown in **Figure 1C,E**, the cardiovascular department responded in accordance with the JCS 2022 Guideline, picking up RCRI items and responding. In the case of high-risk surgery, the cardiovascular department often encouraged the external surgical departments to continue taking medications, particularly  $\beta$ -blockers, in accordance with the JCS 2022 Guideline (**Figure 1E**).

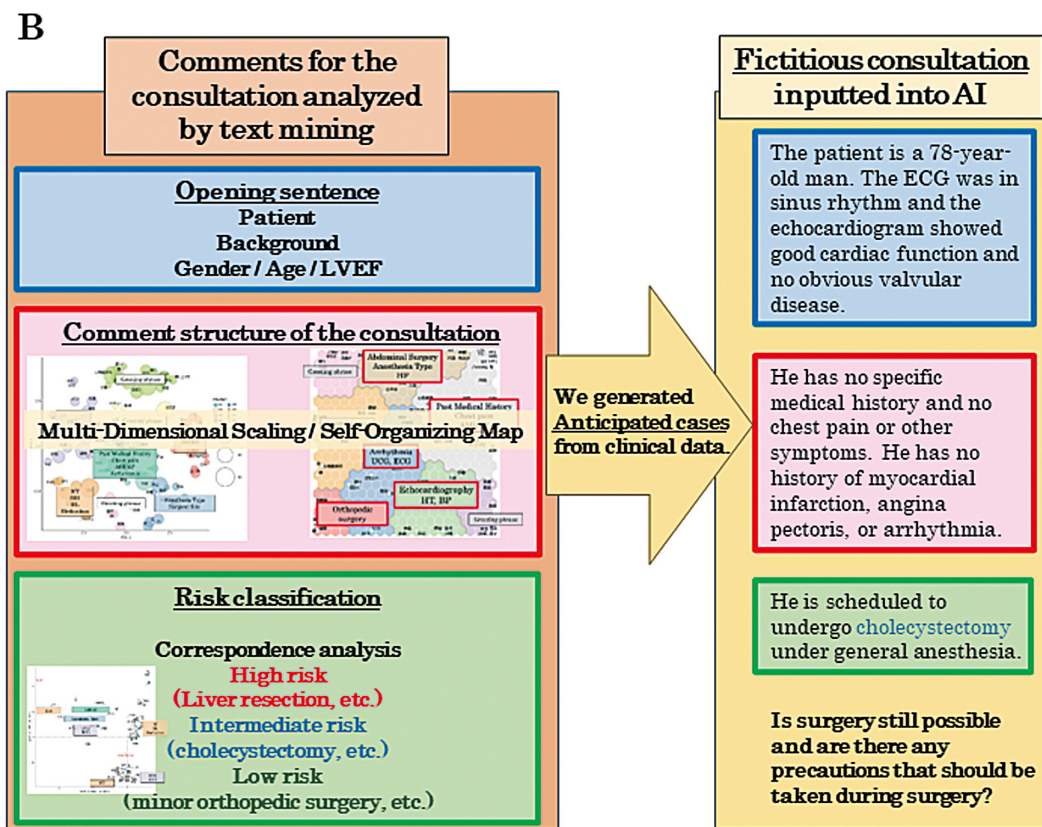
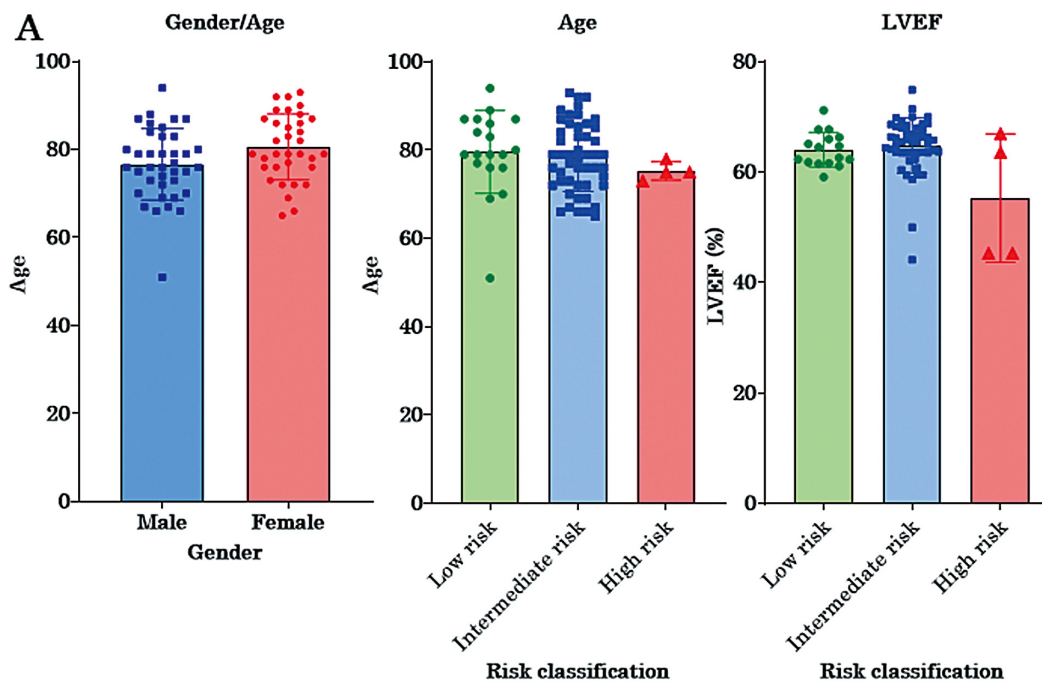
However, to evaluate the use of LLMs as an alternative to preoperative cardiovascular consultations in the future, the accuracy of LLMs in answering consultation requests from surgeons was investigated. Therefore, based on **Figure 1B,D**, and **Figure 2A**, virtual scenarios were created by dividing cases into high, intermediate, and low risks, according to the patient's age, and cardiac function, which are items that were often described by surgeons in clinical practice, and the surgical risk in accordance with the JCS 2022 Guideline. LLM was consulted for these virtual scenarios.

**Figure 3** shows the virtual scenarios. **Figure 4** shows the results of the evaluation by one cardiovascular specialist. Google BARD can generate fairly reasonable responses in low-risk cases. However, in intermediate- and high-risk cases, the amount of patient information, medical history, and cardiac function information increases. The present study showed that Google BARD could not generate satisfactory responses. In addition, due to unknown reasons, adding information about valvular disease or low cardiac function caused Google BARD to start quoting other information from the Internet that was not in accordance with the guidelines, or to fabricate information. However, Google Gemini had the ability to provide answers to questions that were not adequately addressed in Google BARD, as shown in **Figure 4**. It is anticipated that there will be significant advancements in LLM in the future. Although actual cardiologists might be able to deduce the surgeon's intentions (**Figure 1C,E**), search for relevant information in medical records, and make recommendations for medication, this remains challenging for LLMs.

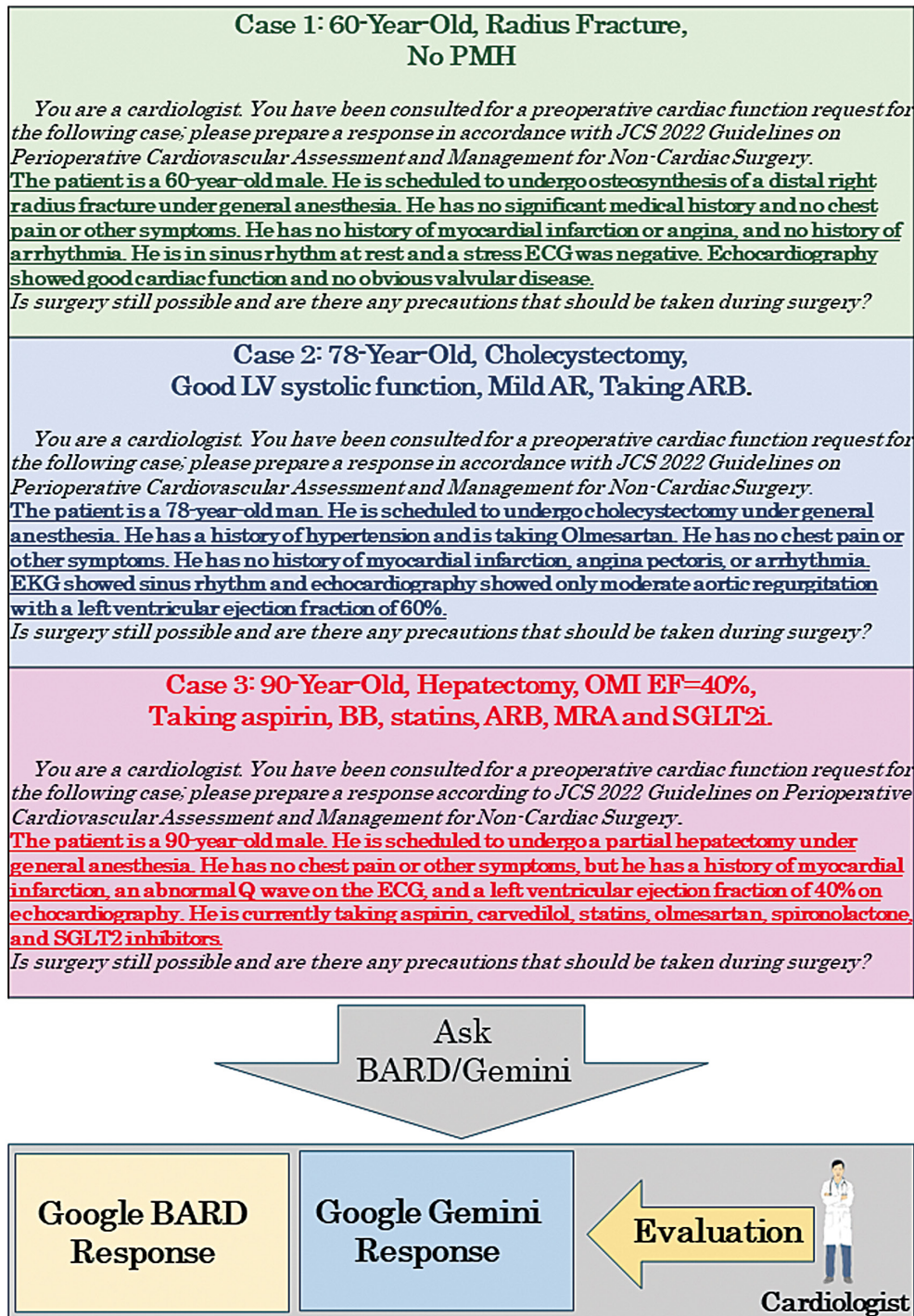
Despite its limitations, the transition from Google BARD to Google Gemini has significantly enhanced the




































**Figure 1.** (A) The original clinical data were written in Japanese. Thus, the analysis was performed in Japanese. Based on the results of the analysis on consultation requests, a virtual scenario was created in English, and preoperative consultation was performed in English using Google BARD/Gemini. The responses created by Google BARD/Gemini were evaluated by a cardiovascular specialist in terms of understandability, appropriateness, absence of missing information, fabrication, and clinical evaluation. (B) Self-organizing maps were created using KH Coder for the consultation requests from surgery for preoperative consultations. The maps were divided into three main clusters, which were as follows: abdominal surgery; anesthesia type; and heart failure (HF). The areas related to 'previous medical history', 'chest pain', and 'acute myocardial infarction (AMI)/angina pectoris (AP)' were adjacent to these clusters. This suggests that surgical departments focus on these items. The cluster for 'echocardiography', 'hypertension (HT)', and 'blood pressure (BP)' was adjacent to the cluster for 'orthopedic surgery'. Hence, 'arrhythmia', 'ultrasound echocardiography (UCG)', and 'electrocardiograph (ECG)' can be important for both surgical and orthopedic departments. (C) Self-organizing map of the response content from the cardiovascular department. The word 'guideline' was located in the upper right corner, surrounded by words related to 'revised Cardiac Risk Index (RCRI)' and 'risk'. Other words related to other matters were listed outside of these areas. (D) Response to the request to the cardiovascular department, analyzed according to surgical risk. (E) Response from the cardiovascular department, analyzed according to surgical risk.




**Figure 2.** (A) A graph of sex, age, and left ventricular ejection fraction (LVEF) was created to capture the clinical characteristics of actual patients. (B) Virtual cases were created based on the clinical characteristics of the patients, as depicted in Figure 1B,C, and Figure 2A. AI, artificial intelligence.




**Figure 3.** Three virtual cases were created for consultation with Google BARD and Google Gemini based on the clinical data. The sentences in black letters are instructions to Google BARD/Gemini and questions common to each case. In case 1, the patient was at low risk (as shown in the graph in **Figure 2A**), young, and had a relatively high left ventricular ejection fraction (LVEF). The patient had no significant past medical history (PMH), as shown in **Figure 1B**. In case 2, the patient was at average risk (as shown in the graph in **Figure 2A**), with blood pressure of 140/90 mmHg, and a history of hypertension. In case 3, the patient was at high risk (as shown in the graph in **Figure 2A**), aged 75 years, and had a LVEF of 40%. The patient had a history of myocardial infarction and was taking several medications. The responses created by Google BARD and Google Gemini were evaluated by a cardiovascular specialist in terms of 'comprehensibility', 'appropriateness', 'absence of relevant content', 'confabulation', and 'clinical decisions'. Images reproduced with permission from Servier Medical Art under a Creative Commons Attribution 3.0 Unported License (<https://creativecommons.org/licenses/by/3.0/>). AR, aortic valve regurgitation; ARB, angiotensin receptor blocker; BB,  $\beta$ -blocker; ECG, electrocardiogram; EF, ejection fraction; EKG, electrocardiogram; LV, left ventricle; MRA, mineralocorticoid receptor antagonist; OMI, old myocardial infarction; SGLT2i, Sodium-Glucose Transport Protein 2 inhibitor.

<p><b>Fictitious consultation</b></p> <p><b>Indicator</b></p>	<p><b>60-Year-Old</b> Radius Fracture No PMH</p> 	<p><b>78-Year-Old</b> Cholecystectomy Good LV systolic function Mild AR Taking ARB.</p> 	<p><b>90-Year-Old</b> Hepatectomy OMI EF=40% Taking aspirin, BB, statins, ARB, MRA and SGLT2i.</p> 
<p><b>Comprehensibility</b></p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>
<p><b>Appropriateness</b></p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>
<p><b>Absence of relevant content</b></p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ×</p>
<p><b>Confabulation</b></p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>
<p><b>Clinical decisions</b></p>	<p> BARD ○</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>	<p> BARD ×</p> <p> Gemini ○</p>



The cardiologists agreed that BARD/Gemini response was appropriate.



The cardiologists did not agree that BARD/Gemini response was appropriate.

The items highlighted by orange squares were inaccurate in BARD but correct in Gemini.

**Figure 4.** Results of the evaluation of the responses created by Google BARD and Google Gemini for preoperative consultations. The evaluations were made in terms of ‘comprehensibility’, ‘appropriateness’, ‘absence of relevant content’, ‘confabulation’, and ‘clinical decisions’. The colors in the figure represent the following: Green, the cardiologist considered the item to be appropriate; Red, the cardiologist considered the item to be inappropriate; Orange squares, items highlighted with orange squares were inaccurate in Google BARD but correct in Google Gemini. Images reproduced with permission from Servier Medical Art under a Creative Commons Attribution 3.0 Unported License (<https://creativecommons.org/licenses/by/3.0/>). AR, aortic valve regurgitation; ARB, angiotensin receptor blocker; BB,  $\beta$ -blocker; EF, ejection fraction; LV, left ventricle; MRA, mineralocorticoid receptor antagonist; OMI, old myocardial infarction; PMH, past medical history; SGLT2i, Sodium-Glucose Transport Protein 2 inhibitor.

ability to comply with guidelines, and practical implementation is imminent.

### Conclusions

Google Gemini compensated for deficiencies in Google BARD and can now handle cardiology consults for intermediate-and high-risk patients, in addition to low-

risk patients. This improvement is promising for future practical use.

### Acknowledgments

We thank Google AI for providing the Google BARD and Google Gemini language models. We used Google BARD and Gemini to generate the consultation response. This work was supported by JSPS KAKENHI Grant Number 23K19602 (to T.K.) and a Chukyo

longevity medical and promotion foundation research grant number 2023.03.09 (to T.K.).

### Sources of Funding

T.K. received grants from JSPS KAKENHI (grant number 23K19602) and Chukyo Longevity Medical and Promotion Foundation Research (grant number 2023.03.090).

### Disclosures

The authors declare that there are no conflicts of interest.

### IRB Information

The protocol for the research project was approved by the ethics committee of the National Center for Geriatrics and Gerontology (no. 1668-2) and the External Service Usage Manager of the Department of Cardiology in the National Center for Geriatrics and Gerontology within which the work was undertaken. The present study was performed in accordance with the provisions of the Declaration of Helsinki (as revised in Tokyo 2004).

### Data Availability

All data generated or analyzed during this study are included in this article. Further inquiries can be directed to the corresponding author.

### References

- Hiraoka E, Tanabe K, Izuta S, Kubota T, Kohsaka S, Kozuki A, et al. JCS 2022 guideline on perioperative cardiovascular assessment and management for non-cardiac surgery. *Circ J* 2023; **87**: 1253–1337, doi:10.1253/circj.CJ-22-0609.
- Acar AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg* 2023; **125**: 101724, doi:10.1016/j.jormas.2023.101724.
- Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. *Eur Arch Otorhinolaryngol* 2024; **281**: 2137–2143, doi:10.1007/s00405-023-08381-3.
- Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. *J Med Internet Res* 2023; **25**: e51580, doi:10.2196/51580.
- Hillmann HAK, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace* 2023; **26**: euad369, doi:10.1093/europace/euad369.
- Iannantuono GM, Bracken-Clarke D, Karzai F, Choo-Wosoba H, Gulley JL, Floudas CS. Comparison of large language models in answering immuno-oncology questions: A cross-sectional study. *medRxiv* 2023, doi:10.1101/2023.10.31.23297825.
- McGowan M, Correia Martins F, Keen JL, Whitehead A, Davis E, Pathiraja P, et al. Can natural language processing be effectively applied for audit data analysis in gynaecological oncology at a UK cancer centre? *Int J Med Inform* 2024; **182**: 105306, doi:10.1016/j.ijmedinf.2023.105306.
- Nguyen D, Swanson D, Newbury A, Kim YH. Evaluation of ChatGPT and Google Bard using prompt engineering in cancer screening algorithms. *Acad Radiol* 2023, doi:10.1016/j.acra.2023.11.002.
- Patil NS, Huang RS, Caterine S, Yao J, Larocque N, van der Pol CB, et al. Artificial intelligence Chatbots' understanding of the risks and benefits of computed tomography and magnetic resonance imaging scenarios. *Can Assoc Radiol J* 2024, doi:10.1177/08465371231220561.
- Roberts RHR, Ali SR, Dobbs TD, Whitaker IS. Can large language models generate outpatient clinic letters at first consultation that incorporate complication profiles from UK and USA aesthetic plastic surgery associations? *Aesthet Surg J Open Forum* 2024; **6**: ojad109, doi:10.1093/asjof/ojad109.
- Thibaut G, Dabbagh A, Liverneaux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? *Int Orthop* 2024; **48**: 151–158, doi:10.1007/s00264-023-06034-y.
- Kazawa K, Akishita M, Ikeda M, Iwatsubo T, Ishii S. Experts' perception of support for people with dementia and their families during the COVID-19 pandemic. *Geriatr Gerontol Int* 2022; **22**: 26–31, doi:10.1111/ggi.14307.
- Maeda W, Hirakawa Y, Muraya T, Miura H. Text mining analysis of newspaper editorials concerning the COVID-19 pandemic from a healthcare perspective. *J Rural Med* 2022; **17**: 279–282, doi:10.2185/jrm.2021-063.
- Mori Y, Miyatake N, Suzuki H, Mori Y, Okada S, Tanimoto K. Comparison of impressions of COVID-19 vaccination and influenza vaccination in Japan by analyzing social media using text mining. *Vaccines (Basel)* 2023; **11**: 1327, doi:10.3390/vaccines11081327.