scientific reports

Check for updates

OPEN

Disparities in medical recommendations from AI-based chatbots across different countries/regions

Khanisyah E. Gumilar[1,2]✉, Birama R. Indraprasta[3], Yu-Cheng Hsu[4,5], Zih-Ying Yu[4], Hong Chen[1], Budi Irawan[3], Zulkarnain Tambunan[3], Bagus M. Wibowo[3], Hari Nugroho[3], Brahmana A. Tjokroprawiro[3], Erry G. Dachlan[3], Pungky Mulawardhana[2], Eccita Rahestyningtyas[2], Herlangga Pramuditya[6], Very Great E. Putra[7], Setyo T. Waluyo[8], Nathan R. Tan[9], Royhaan Folarin[10], Ibrahim H. Ibrahim[1], Cheng-Han Lin[1], Tai-Yu Hung[1], Ting-Fang Lu[11], Yen-Fu Chen[11], Yu-Hsiang Shih[11], Shao-Jing Wang[11], Jingshan Huang[12], Clayton C. Yates[13], Chien-Hsing Lu[11]✉, Li-Na Liao[4]✉ & Ming Tan[1,14]✉

This study explores disparities and opportunities in healthcare information provided by AI chatbots. We focused on recommendations for adjuvant therapy in endometrial cancer, analyzing responses across four regions (Indonesia, Nigeria, Taiwan, USA) and three platforms (Bard, Bing, ChatGPT-3.5). Utilizing previously published cases, we asked identical questions to chatbots from each location within a 24-h window. Responses were evaluated in a double-blinded manner on relevance, clarity, depth, focus, and coherence by ten experts in endometrial cancer. Our analysis revealed significant variations across different countries/regions ($p < 0.001$). Interestingly, Bing's responses in Nigeria consistently outperformed others ($p < 0.05$), excelling in all evaluation criteria ($p < 0.001$). Bard also performed better in Nigeria compared to other regions ($p < 0.05$), consistently surpassing them across all categories ($p < 0.001$, with relevance reaching $p < 0.01$). Notably, Bard's overall scores were significantly higher than those of ChatGPT-3.5 and Bing in all locations ($p < 0.001$). These findings highlight disparities and opportunities in the quality of AI-powered healthcare information based on user location and platform. This emphasizes the necessity for more research and development to guarantee equal access to trustworthy medical information through AI technologies.

Keywords Artificial intelligence, Endometrial cancer, Bing, Bard, ChatGPT, Disparity

[1]Graduate Institute of Biomedical Science, China Medical University, Taichung, Taiwan. [2]Department of Obstetrics and Gynecology, Hospital of Universitas Airlangga-Faculty of Medicine, Universitas Airlangga, Jl. Dharmahusada Permai, Mulyorejo, Kec. Mulyorejo, Surabaya, Jawa Timur 60115, Indonesia. [3]Department of Obstetrics and Gynecology, Dr. Soetomo General Hospital-Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia. [4]Department of Public Health, China Medical University, No. 100, Sec. 1, Jingmao Rd, Beitun Dist, Taichung 406040, Taiwan, ROC. [5]School of Chinese Medicine, China Medical University, Taichung, Taiwan. [6]Department of Obstetrics and Gynecology, Dr. Ramelan Naval Hospital, Surabaya, Indonesia. [7]Department of Obstetrics and Gynecology, Dr. Kariadi Central General Hospital, Semarang, Indonesia. [8]Department of Obstetrics and Gynecology, Ulin General Hospital, Banjarmasin, Indonesia. [9]Department of Modern and Classical Languages and Literature, University of South Alabama, Mobile, AL, USA. [10]Department of Anatomy, Faculty of Basic Medical Sciences, Olabisi Onabanjo University, Sagamu, Nigeria. [11]Department of Obstetrics and Gynecology, Taichung Veteran General Hospital, 1650 Taiwan Boulevard Sector. 4, Taichung 40705, Taiwan, ROC. [12]School of Computing and College of Medicine, University of South Alabama, Mobile, AL, USA. [13]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA. [14]Institute of Biochemistry and Molecular Biology, Graduate Institute of Biomedical Sciences, China Medical University (Taiwan), No. 100, Sec. 1, Jingmao Rd, Beitun Dist, Taichung 406040, Taiwan, ROC. ✉email: khanisyah@fk.unair.ac.id; chlu@vghtc.gov.tw; linaliao@mail.cmu.edu.tw; mingtan@mail.cmu.edu.tw

The rise of large language models (LLM) is transforming the landscape of digital health, paving the way for innovative applications in medical research, diagnosis, and treatment[1]. In less than half a year, three major Artificial intelligence (AI)-based chatbots have been successfully released. It is known that Open AI's ChatGPT (hereafter referred to as ChatGPT-3.5) initially debuted in November 2022[2]. Early February 2023, Microsoft officially launched the Bing Chat or Bing AI chatbot (hereafter referred to as Bing). Then, in March 2023, Google released Google Bard (which is currently known as "Gemini", hereafter referred to as Bard)[3,4]. These chatbots immediately gained a lot of attention for their ability to understand and generate human-like text based on large amounts of training data[5–7]. The transformational power of AI has been reported in various medical features, ranging from diagnosis capabilities to therapy planning and patient care[8–10].

Meanwhile, equality in AI access and the quality of technological development should not be based on a country's location or economic level. The advancement of AI-based chatbots should be enjoyed fairly throughout the world. Technological advances caused uneven access in various parts of the world due to existing patterns of discrimination, inequality, and bias, which should be avoided[11].

Identifying information inequities is crucial for public health as it ensures equal access to reliable health information, reduces disparities, and allows policymakers to support disadvantaged populations effectively. Overcoming these inequities will lead to improved health literacy, informed decisions, and health outcomes[12,13]. Several studies have found that unequal access to medical information is associated with poor health outcomes[14] and unsatisfied patient experiences[15,16].

Gynecologic oncology, relies heavily on accurate and timely information, as inaccuracies can be harmful. Geographical inaccuracies can exacerbate global health disparities. This study examines whether AI-LLM provides consistent, unbiased medical information across different countries/regions. The goal is to identify potential discrepancies in AI-generated medical advice, ensuring fair and reliable health information access worldwide. The insights gained could guide the improvement of AI systems, fostering global health equity and trust in AI technology. Understanding these differences is vital for advancing and adopting AI-based technologies in medical practice, ultimately improving patient outcomes and experiences globally.

Methods

Ethics

Approval from the ethics committee was not required since no patients were involved in our study.

Materials

We used AI-based chatbots: Bing (https://www.bing.com/), Bard (https://gemini.google.com/app), and ChatGPT3.5 (https://chatgpt.com/) in this study. The patient scenarios were based on published case reports[17–22]. The VPN we used was the SoftEther VPN client downloaded from https://softether-vpn-client.en.softonic.com/.

Study design

Three freely accessible AI-based chatbots (Bing, Bard, and ChatGPT-3.5) were investigated. Each chatbot in Indonesia, Nigeria, Taiwan, and the United States was presented with six patient cases (from now on referred to as "patient") (Fig. 1). The patient case scenarios were publicly available published data and then adapted into vignettes and prompts. This step is processed into a question by adding a command sentence and a structured writing system. The selected patients are those with various ages, clinical stages, and histopathology results. We used The National Comprehensive Cancer Network (NCCN) guideline as the standard of care for endometrial cancer[23], which is widely accepted as treatment guidelines in many countries and provides treatment principles for endometrial cancer. Then we provide the patient vignette to the chatbot. After that, the identical question was asked "*What is the most appropriate adjuvant therapy in this case based on consensus from NCCN?*" (Supp. 1).

The vignette of each patient was input in each chatbot and location only once. All questions were tested on 3 chatbots in 4 countries on the same date (Sunday, November 26th, 2023). All the input across four locations was completed within a 24-h period to minimize the potential variation from ever-evolving AI chatbots. The replies were immediately entered into a database for review by a rater team, which included ten well-qualified gynecologic oncologists. To avoid bias, the responses from the AI chatbots were coded and randomized before being transferred to the raters for scoring. The raters rated the responses without knowing which response was from which chatbot and from what location. To evaluate the output from the chatbots, we used 5 parameters including "relevance", "clarity", "depth", "focus", and "coherence"[9,24–26] with a 5-point Likert scale (Table 1)[27,28].

Statistical analysis

Homogeneity Index of the raters: Pearson and Spearman correlation coefficient analyses were performed to study the homogeneity among scores assigned by the raters to the responses. The interpretation of correlation coefficients is as follows: values greater than or equal to 0.4 but less than 0.7 are considered moderate, values greater than or equal to 0.1 but less than 0.4 are regarded as weak, and values greater than or equal to 0.7 but less than 1.0 are indicative of a strong correlation[29–31]. Furthermore, we performed a one-way ANOVA test with Scheffe's post hoc analysis to evaluate the homogeneity of raters.

We investigated the performance of three AI chatbots in four countries/regions in providing medical recommendations for six patients. The medical advice was evaluated by ten experienced doctors in five parameters: (1) relevance; (2) clarity; (3) depth; (4) focus; and (5) coherence. The assessment parameters were scored by physicians on a Likert scale ranging from 1 to 5, where higher scores denoted superior performance. Subsequently, these evaluations were categorized into "poor," "fair," and "good" based on scores of 1–2, 3, and 4–5, respectively. Additionally, based on prior studies and recommendations, items responses were linearly converted from a

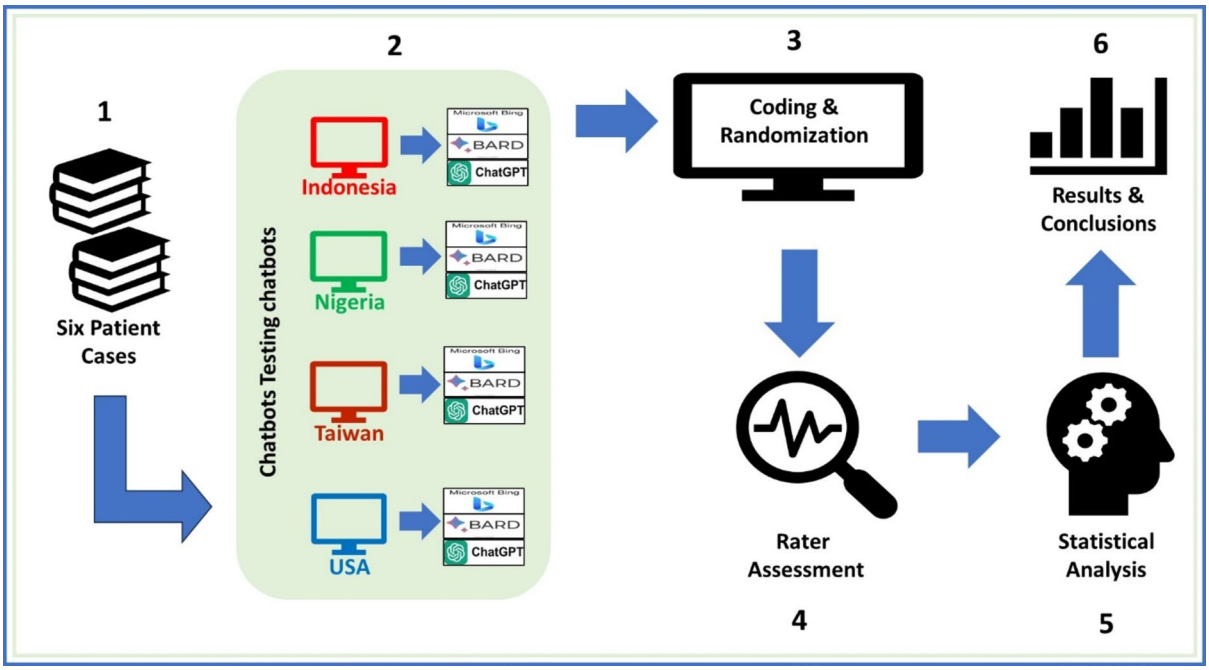

Figure 1. The research protocols. Based on public data, six endometrial cancer patient cases were examined as patient scenarios (1). Six patient cases were tested in Indonesia, Nigeria, Taiwan, and the United States using three AI-based chatbots (Bing, Bard, and ChatGPT-3.5) (2). Answers from each chatbot in the 4 locations were coded and randomized (3). A panel of ten gynecologic oncologists analysed and scored the responses from each chatbot (4). Data analysis and statistical computations were used to determine the importance and compare the results of each chatbot's responses in different places (5). Use diagrams and tables to present study findings and conclusions (6).

	Definition
Assessment parameters	
Relevance	Response is closely related or appropriate to the issue at hand
Clarity	Clear, easy to understand, free from ambiguity, and transparent
Depth	The answer provides detailed and specific information, not just a general or surface answer
Focus	Contains the main points or keywords expected
Coherence	All parts of the answer work together in a logical and structured way, with no conflicting parts
Scoring scale	
1 = Very poor	The answer does not fulfil the basic criteria, is highly irrelevant or no effort is evident
2 = Poor	The answer fulfils very few of the expected criteria, with many basic errors
3 = Average	The answer fulfils the basic criteria, but does not show more effort or understanding than expected
4 = Good	The answer fulfils all the basic criteria well and shows some aspects that are more than expected
5 = Outstanding	A perfect answer of flawless quality, showing exceptional understanding and complete mastery of the material

Table 1. Assessment parameters and scoring.

5-point Likert scale to a 0–100 scale (higher scores representing better performance). The total score was calculated by averaging the transformed values of the five parameters for subsequent analysis[32–35].

In statistical analyses, to examine the differences in the performance of chatbots across four locations, we employed a Chi-square test to assess variations in the proportions of three categories (i.e., good, fair, and poor) across the four regions. Additionally, Pearson and Spearman correlation coefficients, as well as a one-way ANOVA test with Scheffe's post hoc analysis were used to assess the homogeneity among scores assigned by each rater to all responses. To assess disparities in both the overall performance and each of the five parameters for the medical recommendations among the four regions, a one-way ANOVA test with Scheffe's post hoc analysis was employed. Recognizing potential confounding due to patient condition heterogeneity and the subjectivity of doctor assessments, multiple linear regression was utilized to explore variations in the total score across the four regions. The regression model featured the total score as the dependent variable and the four regions as independent variables, with chatbot, patient, and rater as covariates. All statistical analyses were conducted using SAS (Version 9.4, SAS, Cary, NC, USA), with a significance level set at 0.05.

Results

Our preliminary findings suggest that AI chatbots may provide different responses to general inquiries based on the user's location. When a query about domestic violence was sent to the chatbot Bing from different countries, Bing's responses varied significantly (Supp. 2). This prompted us to raise concerns about potential disparities or inequalities in accessing medical recommendations across different countries/regions. To investigate this, we initiated a systematic, double-blind study to determine if chatbots respond differently to identical questions posed from various geographical locations.

We relied on the expertise of ten gynecologic oncologists to rate all the responses provided by the chatbots. To validate our results, we analyzed the rater's scoring homogeneity. We found that most of the raters (8 out of 10) showed a correlation coefficient between 0.47 and 0.93 using Pearson's test (Fig. 2A), and a value of 0.39–0.91 using Spearman's test (Fig. 2B). Furthermore, we did a one-way ANOVA test with Scheffe's post hoc analysis to evaluate the homogeneity of raters (Fig. 2C). Except for raters 8 and 10 ($p < 0.05$), all other 8 raters showed no significant variation. These results indicate a moderate to high level of homogeneity among the raters, further validating our scoring. To examine if there is a disparity among the four locations, we analyzed the overall performance of the chatbots. Nigeria ranked first based on the combined performance of the three chatbots, with a significant first place ($p < 0.001$) over Indonesia, Taiwan, and the United States (Fig. 3A). Bing Nigeria performed significantly higher ($p < 0.05$) than Bing Indonesia, Taiwan, and USA (Fig. 3B). Similarly,

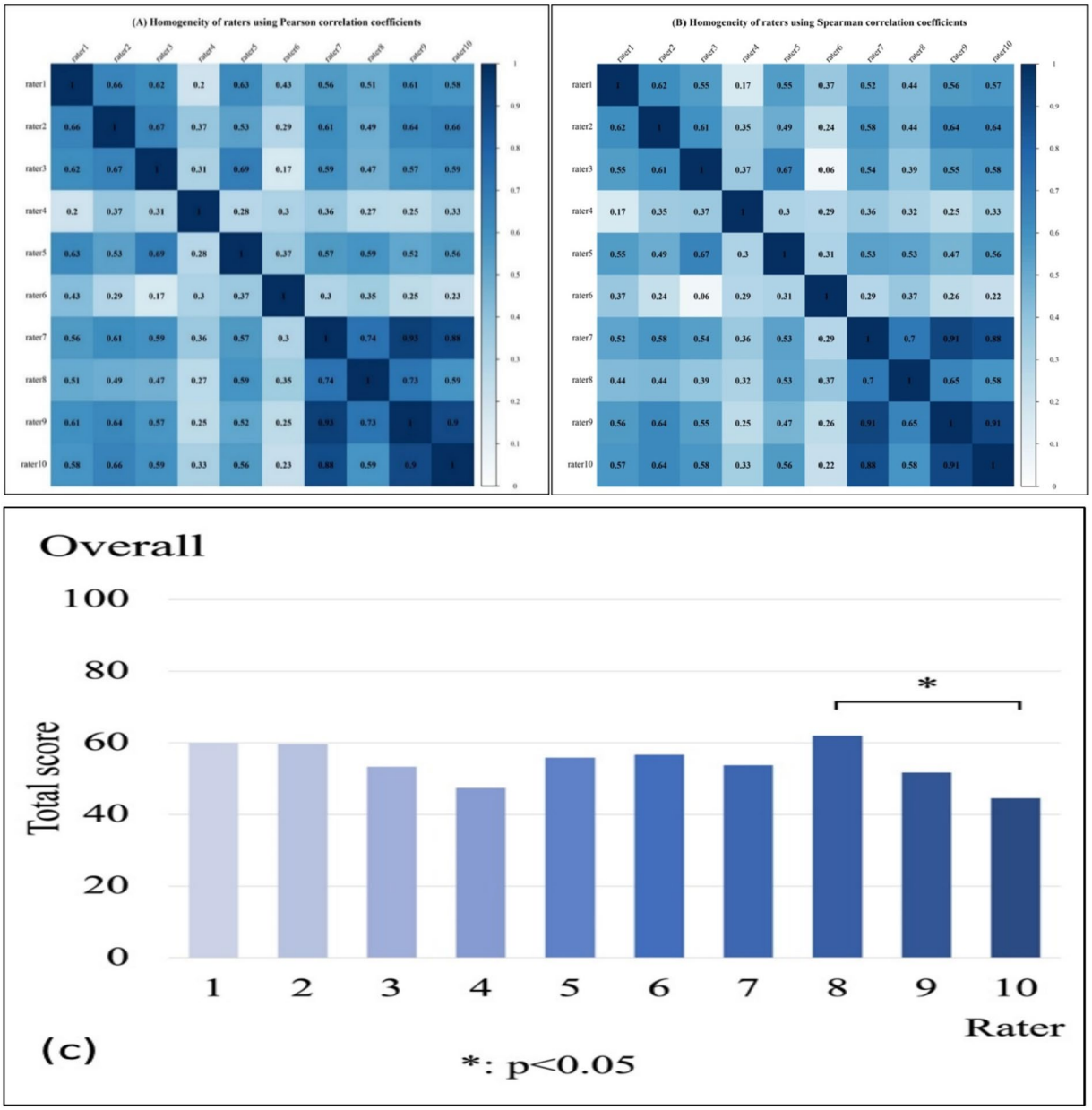

Figure 2. (**A** and **B**) Homogeneity of raters. (**C**) Distribution of overall performance among ten raters, with the p-value indicating the results of the one-way ANOVA test with Scheffe's post hoc analysis.

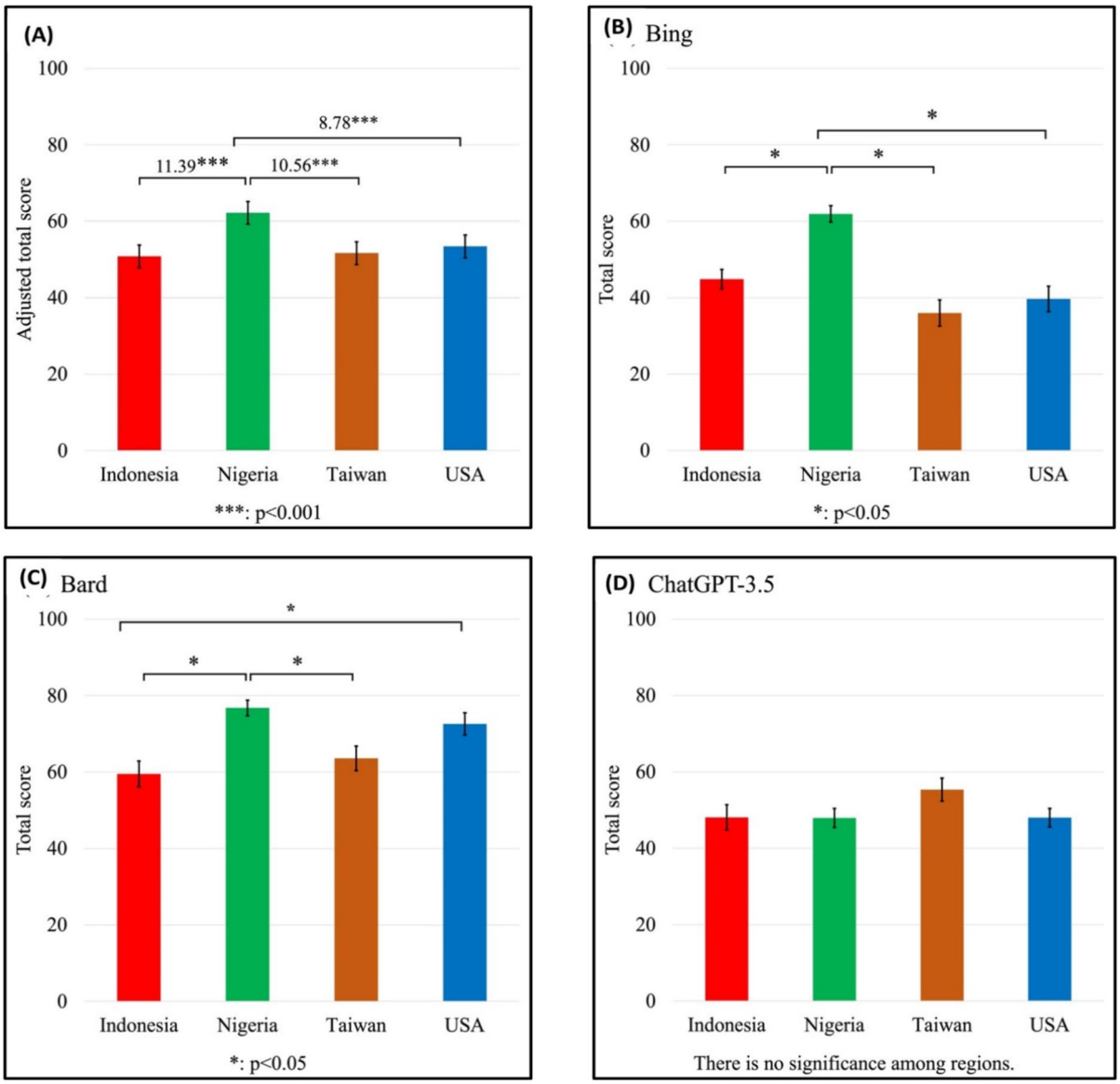

Figure 3. In (**A**), it displays the results of the multiple linear regression model (adjusted for chatbot, patient, and rater), where bars represent the regionally adjusted mean total scores, with error bars representing 95% confidence intervals. The numbers on the chart indicate the adjusted mean differences. Distribution of overall performance for AI chatbots by region: (**B**) Bing, (**C**) Bard, and (**D**) ChatGPT-3.5. In (**B**)–(**D**), bars denote the average total scores for the four regions, with error bars representing standard errors. The p-value in the figure indicates the results of the one-way ANOVA test with Scheffe's post hoc analysis.

Bard Nigeria performed significantly higher ($p < 0.05$) than Bard Indonesia and Taiwan, but not Bard USA. In addition, Bard in the USA performed significantly higher than Bard in Indonesia (Fig. 3C). The performance of ChatGPT-3.5 in Taiwan showed the highest score but was not significantly different from the other three locations (Fig. 3D). Taken together, these results indicate that the treatment recommendations provided by AI chatbots vary significantly by location.

To further confirm our findings, we employed a VPN to examine if there are disparities in responses based on questioning location. The question was asked in Taiwan, without a VPN or with a VPN to different countries. ChatGPT-3.5's responses from South Korea and Thailand with VPN differed dramatically from no VPN Taiwan results in terms of completeness and structured systematics (Supp. 3). These results indicate that by using a VPN to change location, one can obtain different responses provided by the chatbot, further supporting our results that response provided by AI chatbots vary by location.

To evaluate which chatbot's response achieves the highest score, we analyzed the scores of the chatbots in the five individual parameters. Overall, among the four locations, Bard scored higher than ChatGPT-3.5 and Bing (Fig. 4A) by a significant gap ($p < 0.001$), suggesting that Bard provides better recommendations according to NCCN guidelines. The results showed that Bing in Nigeria had the highest ratings on all parameters substantially (all $p < 0.001$) compared to Indonesia, Taiwan, and the United States (Fig. 4B). A similar performance was seen

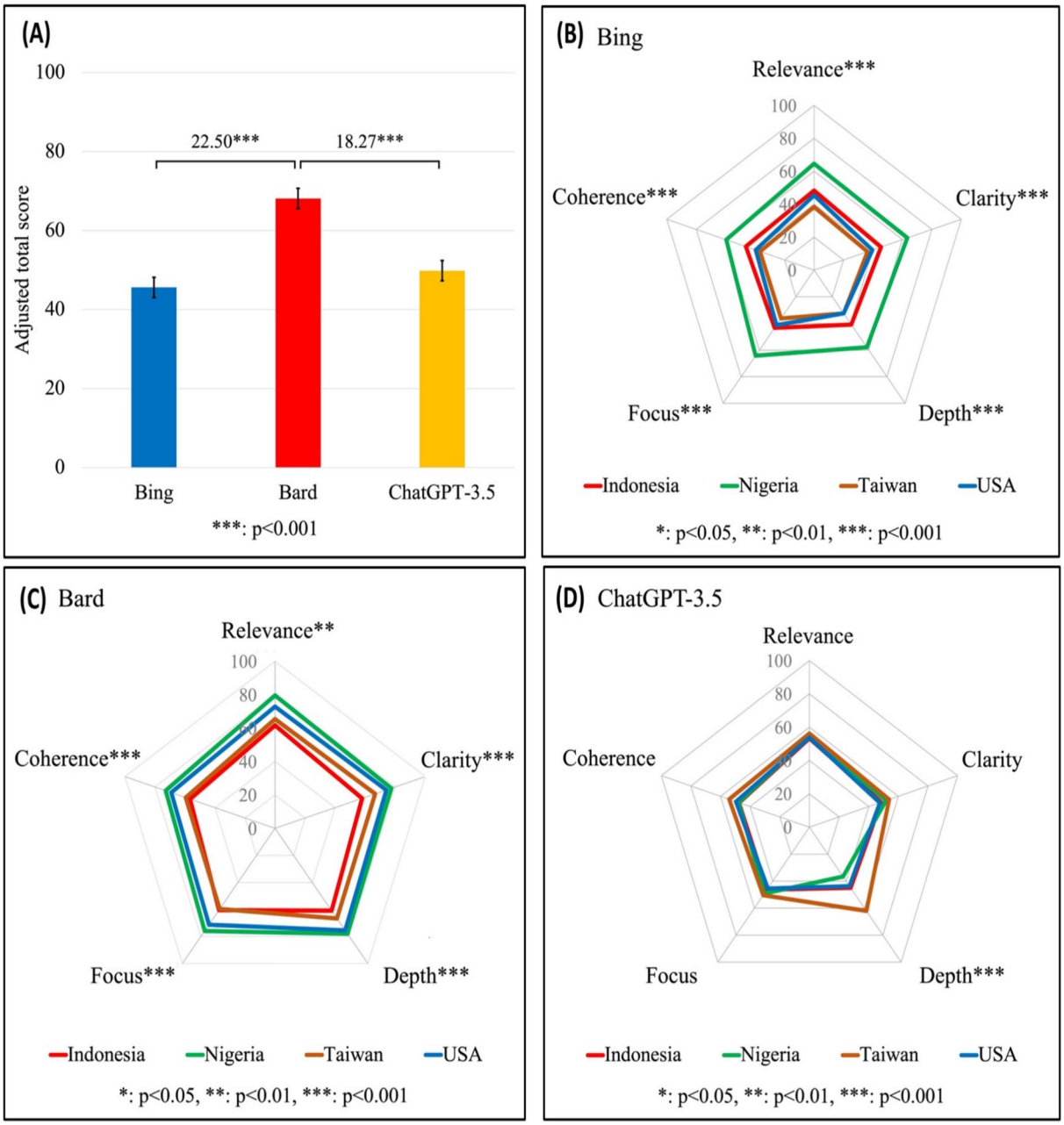

Figure 4. In (**A**), it presents the results of the multiple linear regression model (adjusted for region, patient, and rater), where bars represent the adjusted mean total scores for the three AI chatbots, with error bars representing 95% confidence intervals. The numbers on the chart indicate the adjusted mean differences. Distribution of AI chatbot performance across five parameters by region: (**B**) Bing, (**C**) Bard, and (**D**) ChatGPT-3.5. In (**B**)–(**D**), radar charts summarize the 100-point scale scores for each parameter, with the p-value for each parameter indicating the results of the one-way ANOVA test.

for Bard Nigeria, which outperformed Bard in Indonesia (Fig. 4C). ChatGPT-3.5 in Taiwan achieved the highest score in the "depth", where it outperformed the other three locations significantly ($p < 0.001$) (Fig. 4D). A more detailed information table about the chatbot performance across five parameters by locations are shown in Fig. 5. These results indicate that the scores of the responses provided by AI chatbots vary by platform and Bard outperformed the other two chatbots in this specific study.

Discussion

This study shows that AI chatbots such as Bing, Bard, and ChatGPT-3.5 give different responses to the same question depending on the user's location. We believe the variations in responses observed in this study are not due to time differences[36], as all questions were tested on the same date, within a 24-h window. The use of VPN in the study further supports that the different responses are indeed depending on geographical locations.

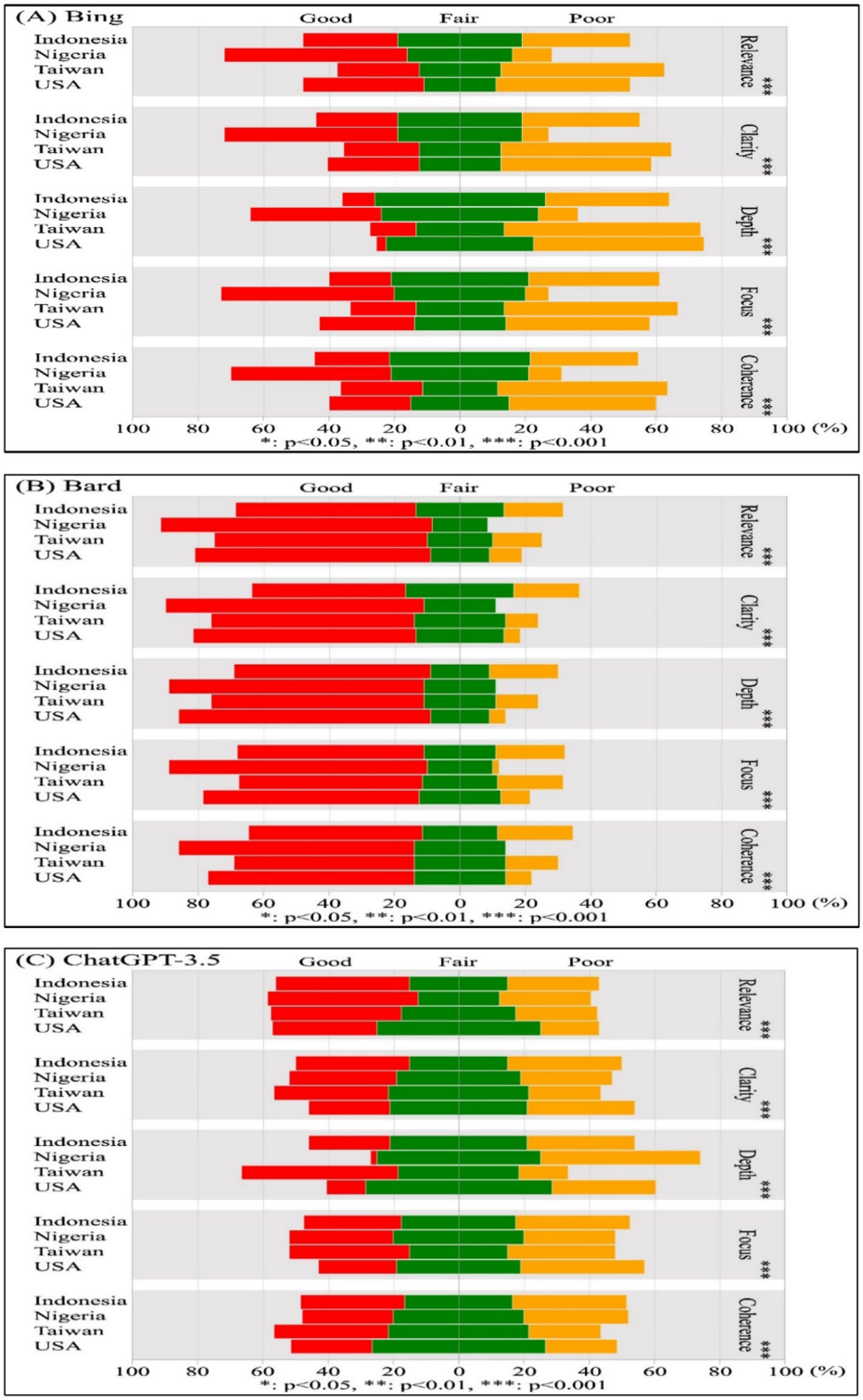

Figure 5. Distribution of AI-based chatbot performance across five parameters by region: (**A**) Bing, (**B**) Bard, and (**C**) ChatGPT-3.5. The bar denotes the percentage of good, fair, and poor for the four regions. The p-value in the figures indicates the results of the Chi-square test.

The variable performance of chatbots such as Bing, Bard, and ChatGPT-3.5 in various regions highlights the importance of considering local geographical and cultural factors in AI development. Our findings in this study will have several implications for future medical AI development:

Inequality of health conditions

These findings suggest that AI chatbots provide different endometrial cancer treatment recommendations based on geographic location. The observed differences in chatbot responses may exacerbate existing health inequalities. In regions where chatbot responses are less accurate or comprehensive, users may not receive important health information needed to make informed decisions. For example, if a chatbot provides inadequate advice regarding domestic violence or reproductive health in one region, this could lead to delayed or inappropriate action, negatively impacting health outcomes. The study's results highlight the necessity for standardizing AI chatbot responses to provide fair access to reliable health information globally.

Trust in AI to deliver health information

Disparities in the quality of responses from AI chatbots may undermine public trust in this technology. If users consistently receive inaccurate or incomplete information, they may become skeptical about the reliability of AI for health advice. This distrust may discourage people from using AI health tools, which are intended to improve access to medical information. Lack of trust in AI systems may hinder their adoption and effectiveness, ultimately limiting their potential to improve health literacy and outcomes. Addressing these inconsistencies is critical for building and maintaining trust in AI-based health solutions.

Challenges in implementing policies and regulations

The observed variability in AI chatbot responses poses challenges for policymakers and regulators. Ensuring that AI health technologies provide consistent and accurate information across different regions requires strong policies and regulations. However, developing and enforcing such standards can be difficult, especially in countries with diverse cultural and regulatory circumstances. Policymakers must acknowledge these complexities to establish regulations that support fairness and dependability in AI medical devices. Successfully addressing these challenges is critical to harnessing AI for equitable health access and improving public health outcomes globally. Public health stakeholders must be aware of these differences to reduce the potential for bias in AI health technologies.

In this study we also applied rigorous validation methods, including correlation and ANOVA tests, to ensure homogeneity of rater scores. The moderate to high agreement among the raters adds credibility to these findings. Validating the consistency of chatbot responses is critical to building trust in these AI systems.

We found that Bard outperformed ChatGPT-3.5 and Bing in our study in providing endometrial cancer adjuvant therapy recommendations. These advantages may be related to the algorithms used, the data train, or the way Bard processes and presents information. Chatbot performance is largely influenced by Natural Language Processing (NLP), Machine Learning (ML), and User experience (UX) considerations. In recent years, chatbot technology has advanced, making it more accessible and widely employed across a variety of businesses. NLP and ML are critical for making chatbots more conversational and human-like, while UX is required to ensure that users have a favorable engagement with the chatbot[26].

Nevertheless, some limitations of AI-based chatbots are still need to be addressed. According to a recent study, a chatbot does not adequately represent the demographic diversity of medical disorders when creating clinical vignettes, resulting in stereotypical presentations[37]. AI can also generate inaccurate or nonsensical content, a condition known as hallucinations or confabulation[38], implying that they might provide wrong answers, which is especially problematic for persons with low education or the capacity to judge the accuracy of information.

Several studies have compared Bing, Bard, and ChatGPT in a medical context, yielding various results about their performance. In a study on methotrexate use, ChatGPT-3.5, Bard, and Bing had correct answer rates of 100%, 73.91%, and 73.91%, respectively[39]. A comparative study in endodontics also found that ChatGPT-3.5 offered more reliable information than Bard and Bing[40]. Dhanvijay et al.'s comparative analysis, as well as a number of other studies with similar outcomes[26,41,42]. In other studies, Bard outperforms ChatGPT and Bing[7,43,44]. Interestingly, a Yemeni study found that Bing had the highest accuracy and specificity, outperforming Bard and ChatGPT-3.5[45].

Our study focuses on cancer, which is similar to Rahsepar et al.'s work[9] and Sensoy's[43] by using case scenarios in the form of vignettes such as Dhanvijay's et al.[41]. Our findings showed that Bard performed Bing and ChatGPT3.5 on the topic of endometrial cancer adjuvant therapy. The diversity in study results above suggests that each chatbot may have a distinct advantage in specific areas of the medical domain. In addition, since the chatbots are constantly evolving, testing at different times may produce different results.

AI-based chatbots are adapting and evolving constantly in response to user feedback and iterative training set upgrades. Although we limited our test to different locations within 24 h, it was very difficult to test the chatbots at the same time. This time difference may also cause some variance.

Our study was initiated when we accidentally asked two general questions (about domestic violence and a river incident, Suppl. 2) to Bing from different locations. We observed that the responses differed depending on where the questions were asked. This prompted us to consider the potential impact of such disparities when seeking medical information. The findings of our study suggest that such disparities in responses from AI-based chatbots extend beyond medical information to disparities in AI responses to general information related to people's daily lives.

Based on the findings of this study, we urge AI developers to take on more social responsibilities and put in extra effort to mitigate disparities when creating AI products.We also urge AI industry regulators and policymakers to establish stricter policies for regulating the AI industry, in order to prevent AI products from exacerbating existing disparities. Proactive identification of disparities, targeted mitigation strategies, and increased transparency in model training and data sourcing are essential to ensure that AI benefits everyone, rather than just a few.

Data availability

We have ensured that all important data required has been included on the Supplementary file. The exception is the raw values provided by individual doctors, which can be provided upon request.

Received: 6 March 2024; Accepted: 15 July 2024
Published online: 24 July 2024

References

1. The Lancet Digital H. Large language models: A new chapter in digital health. *Lancet Digit. Health* **6**(1), e1 (2024).
2. Scholar D. *What is ChatGPT : The History of ChatGPT - OpenAI [2023]*. (Accessed 1 Nov 2023) https://digitalscholar.in/history-of-chatgpt/.
3. Team S. *Bing AI: Exploring Bing Chat, an AI-Powered Search Engine*. (Accessed 1 Nov 2023) https://www.semrush.com/blog/bing-ai/.
4. Grant N. *Google Releases Bard, Its Competitor in the Race to Create A.I. Chatbots*. (Accessed 1 Nov 2023) https://www.nytimes.com/2023/03/21/technology/google-bard-chatbot.html.
5. Seth, I. *et al.* Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: A ChatGPT case study focused on breast augmentation. *Aesthet. Surg. J.* **43**(10), 1126–1135 (2023).
6. Gupta, R. *et al.* Utilization of ChatGPT for plastic surgery research: Friend or foe?. *J. Plast. Reconstr. Aesthet. Surg.* **80**, 145–147 (2023).
7. Seth, I. *et al.* Comparing the efficacy of large language models ChatGPT, BARD, and bing AI in providing information on rhinoplasty: An observational study. *Aesthet. Surg. J. Open Forum* **5**, ojad084 (2023).
8. Zuniga Salazar, G. *et al.* Efficacy of AI chats to determine an emergency: A comparison between OpenAI's ChatGPT, google bard, and microsoft bing AI chat. *Cureus* **15**(9), e45473 (2023).
9. Rahsepar, A. A. *et al.* How AI responds to common lung cancer questions: ChatGPT vs google bard. *Radiology* **307**(5), e230922 (2023).
10. Lim, Z. W. *et al.* Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **95**, 104770 (2023).
11. *Gates Foundation Selects Nearly 50 Global Health and Development Projects That Will Contribute to Shaping Equitable Access to AI*. (2023) (Accessed 7 Dec 2023) https://www.gatesfoundation.org/ideas/media-center/press-releases/2023/08/grand-challenges-rfp-recipients-ai-large-language-models.
12. Perez, G. *et al.* Tackling health inequalities in a public health organization: The case of the Barcelona Public Health Agency. *Int. J. Equity Health* **21**(1), 129 (2022).
13. Nichols, L. M. & Ballard, D. J. Leveraging community information to improve health equity. *Mayo Clin. Proc.* **97**(1), 10–11 (2022).
14. Root, T. Causes of health inequity. In *Communities in Action: Pathways to Health Equity* (eds Baciu, A. *et al.*) 99–184 (The National Academies Press, 2017).
15. Ferreira, D. C., Vieira, I., Pedro, M. I., Caldas, P. & Varela, M. Patient satisfaction with healthcare services and the techniques used for its assessment: A systematic literature review and a bibliometric analysis. *Healthcare* **11**(5), 639 (2023).
16. Marzban, S., Najafi, M., Agolli, A. & Ashrafi, E. Impact of patient engagement on healthcare quality: A scoping review. *J. Patient Exp.* **9**, 23743735221125440 (2022).
17. Chepkemoi, L., Ajayi, O., Anabaraonye, N. & Balogun, O. D. Combining concurrent radiotherapy and immunotherapy for synergistic effects in recurrent endometrial cancer—A case report. *Gynecol. Oncol. Rep.* **44**, 101090 (2022).
18. Kakibuchi, A. *et al.* Robot-assisted laparoscopic hysterectomy for early-stage endometrial cancer with massive uterine leiomyomas: A case report. *Int. J. Surg. Case Rep.* **97**, 107473 (2022).
19. Kuno, I., Yoshida, H., Kohno, T., Ochiai, A. & Kato, T. Endometrial cancer arising after complete remission of uterine malignant lymphoma: A case report and mutation analysis. *Gynecol. Oncol. Rep.* **28**, 50–53 (2019).
20. Mandato, V. D. *et al.* Solitary vulvar metastasis from early-stage endometrial cancer: Case report and literature review. *Medicine* **100**(22), e25863 (2021).
21. Si, M. *et al.* Idiopathic retroperitoneal fibrosis with endometrial cancer: A case report and literature review. *BMC Womens Health* **22**(1), 399 (2022).
22. Tsuji, S., Hori, K., Tashima, L., Yoshimura, M. & Ito, K. Multiple metastases after laparoscopic surgery for early-stage endometrial cancer: A case report. *Int. J. Surg. Case Rep.* **76**, 552–556 (2020).
23. Abu-Rustum, N. *et al.* Uterine neoplasms, version 1.2023, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **21**(2), 181–209 (2023).
24. Gordon, E. B. *et al.* Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J. Am. Coll. Radiol.* (2023).
25. Wu, T. *et al.* A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J. Automatica Sinica* **10**(5), 1122–1136 (2023).
26. Bhardwaz, S. & Kumar, J. An extensive comparative analysis of chatbot technologies - ChatGPT, google BARD and microsoft bing. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 673–679 (2023).
27. Sikander, B., Baker, J. J., Deveci, C. D., Lund, L. & Rosenberg, J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: A blinded, randomized, non-inferiority controlled study. *Cureus* **15**(11), e49019 (2023).
28. Veras, M. *et al.* Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: Protocol for a crossover randomized controlled trial. *JMIR Res. Protoc.* **12**, e51873 (2023).
29. Akoglu, H. User's guide to correlation coefficients. *Turk. J. Emerg. Med.* **18**(3), 91–93 (2018).
30. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: Appropriate use and interpretation. *Anesth Analg.* **126**(5), 1763–1768 (2018).
31. Dancey, C. & Reidy, J. *Statistics Without Maths for Psychology* (Pearson, 2020).
32. Ma, D. C. *et al.* Patient experience performance at a primary cancer center versus affiliated community facilities. *Adv. Radiat. Oncol.* **8**(5), 101240 (2023).
33. Kapoor, N. *et al.* Patient experience scores for radiologists: Comparison with nonradiologist physicians and changes after public posting in an institutional online provider directory. *Am. J. Roentgenol.* **219**(2), 338–345 (2022).
34. Vaidya, T. S. *et al.* Appearance-related psychosocial distress following facial skin cancer surgery using the FACE-Q skin cancer. *Arch. Dermatol. Res.* **311**(9), 691–696 (2019).
35. Kamo, N. *et al.* Evaluation of the SCA instrument for measuring patient satisfaction with cancer care administered via paper or via the Internet. *Ann. Oncol.* **22**(3), 723–729 (2011).
36. Bajcetic, M. *et al.* Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. *Ann. Anat.* **254**, 152261 (2024).
37. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *Lancet Digit. Health* **6**(1), e12–e22 (2024).

38. Smith, A. L., Greaves, F. & Panch, T. Hallucination or confabulation? Neuroanatomy as metaphor in Large Language Models. *PLoS Digit. Health* **2**(11), e0000388 (2023).
39. Coskun, B. N., Yagiz, B., Ocakoglu, G., Dalkilic, E. & Pehlivan, Y. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatol. Int.* **44**, 509–515 (2023).
40. Mohammad-Rahimi, H. *et al.* Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int. Endod. J.* **57**, 305–314 (2023).
41. Dhanvijay, A. K. D. *et al.* Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. *Cureus* **15**(8), e42972 (2023).
42. Kumari, A. *et al.* Large language models in hematology case solving: A comparative study of ChatGPT-3.5, google bard, and microsoft bing. *Cureus* **15**(8), e43861 (2023).
43. Sensoy, E. & Citirik, M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. *Int. Ophthalmol.* **43**(12), 4905–4909 (2023).
44. Fijacko, N., Prosen, G., Abella, B. S., Metlicar, S. & Stiglic, G. Can novel multimodal chatbots such as Bing Chat Enterprise, ChatGPT-4 Pro, and Google Bard correctly interpret electrocardiogram images?. *Resuscitation* **193**, 110009 (2023).
45. Al-Ashwal, F. Y., Zawiah, M., Gharaibeh, L., Abu-Farha, R. & Bitar, A. N. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-35, ChatGPT-4, bing AI, and bard against conventional drug-drug interactions clinical tools. *Drug Healthc. Patient Saf.* **15**, 137–147 (2023).

Author contributions

MT, KEG, and BRI contributed to the conception of the study. MT, KEG, and BRI contributed to the methodology and study design. KEG, BRI, C-HL, LNL, YCH, NT, and ZYY contributed to data curation and validation. LNL, YCH, and ZYY contributed to the statistical analysis of the data. LNL, MT, and KEG contributed to visualization which includes figures, charts, and tables of the data. KEG and BRI contributed to project administration. KEG, BRI, and C-HL contributed to resources. KEG, LNL, CY, and MT contributed to the writing and revising the manuscript. Supervision of this research which includes responsibility for the research activity planning and execution was oversighted by MT. KEG, LNL and MT were responsible for the decision to submit the manuscript. All authors read and approved the final version of the manuscript.
During the preparation of this work, the authors used ChatGPT-4.0 and Grammarly to edit and proofread the manuscript to improve readability. After using this tool/service, the authors reviewed, verified, and edited the content as needed. The authors take full responsibility for the content of the publication.

Funding

This research was partly funded by the China Medical University Ying-Tsai Scholar Fund CMU109-YT-04 (to MT). KEG is a recipient of a scholarship from the Taiwan Ministry of Education.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67689-0.

Correspondence and requests for materials should be addressed to K.E.G., C.-H.L., L.-N.L. or M.T.