Patient Education in Bariatric Surgery: Can Artificial Intelligence-based Chatbots Bridge the Knowledge Gap?

Amirreza Izadi, Hesam Mosavari, Ali Hosseininasab, Ali Jaliliyan, Arzhang Jafari, Mohammadhosein Akhlaghpasand, Aghil Rostami, Maziar Moradi-Lakeh, Foolad Eghbali

Submitted to: Journal of Medical Internet Research
on: October 05, 2024

Table of Contents

Patient Education in Bariatric Surgery: Can Artificial Intelligence-based Chatbots Bridge the Knowledge Gap?

Amirreza Izadi[1, 2*] MD; Hesam Mosavari[1, 2*] MD; Ali Hosseininasab[1, 2] MD; Ali Jaliliyan[1, 2] MD; Arzhang Jafari[2] MS; Mohammadhosein Akhlaghpasand[1, 2] MD; Aghil Rostami[3] MSc; Maziar Moradi-Lakeh[4] MD; Foolad Eghbali[2] MD

[1]Artificial Intelligence in Health Research Center, Iran University of Medical Sciences, Tehran, Iran Tehran IR
[2]Surgery Research Center, School of Medicine, Rasool-E Akram Hospital, Iran University of Medical Sciences, Tehran, Iran. Tehran IR
[3]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran Tehran IR
[4]Gastrointestinal and Liver Diseases Research Center, Iran University of Medical Sciences, Tehran, Iran Tehran IR
[*]these authors contributed equally

Corresponding Author:
Foolad Eghbali MD
Surgery Research Center, School of Medicine, Rasool-E Akram Hospital, Iran University of Medical Sciences, Tehran, Iran.
Department of General Surgery, Hazrat-e Rasool General Hospital Niyayesh St, Tehran, Iran
Tehran
IR

Abstract

Background: The global obesity epidemic continues to pose significant health challenges, with an increasing number of individuals seeking Metabolic and Bariatric Surgery (MBS), particularly laparoscopic sleeve gastrectomy (LSG). However, many MBS centers face resource limitations that limit the availability of adequate patient education, leading to knowledge gaps among patients who require more information. This consequently leads patients to seek information online increasingly. Artificial intelligence (AI) based chatbots like ChatGPT offer a promising tool to provide reliable and accessible medical information. However, concerns about the accuracy, reliability, and comprehensiveness of AI-generated responses still need to be addressed.

Objective: This study aims to evaluate the effectiveness of AI-based chatbots in answering frequently asked patient questions about LSG and compare their performance with that of bariatric surgery experts.

Methods: The study included four fellowship-trained minimally invasive surgeons (MISs), nine minimally invasive surgery fellows (MIFs), and two general practitioners (GPs) involved in MBS multidiciplinary team, forming the expert group. Seven AI chatbots, including ChatGPT versions 3.5 and 4, Bard, Bing, Claude, Llama, and Perplexity, were selected based on public availability. Forty patient questions about LSG were derived from social media, MBS organizations, and online patient forums. Experts and chatbots answered these questions, and their responses were scored for accuracy and comprehensiveness using a 5-point scale. Statistical analyses were performed to compare group performance.

Results: Chatbots demonstrated a higher overall performance score (2.55 ± 0.95) compared to the expert group (1.92 ± 1.32, $P < .001$). Among chatbots, ChatGPT-4 achieved the highest performance (2.94 ± 0.24), while Llama had the lowest (2.15 ± 1.23). Expert group scores were highest for MISs (2.36 ± 1.09), followed by GPs (1.90 ± 1.36) and MIFs (1.75 ± 1.36). The readability of chatbot responses was assessed using Flesch-Kincaid scores, revealing that most responses required reading levels between the 11th grade and college level. Furthermore, chatbots exhibited fair reliability and reproducibility in response consistency, with ChatGPT-4 showing the highest test-retest reliability.

Conclusions: AI-based chatbots provide reliable and comprehensive answers to common patient questions about LSG. These chatbots could play a significant role in patient education. Still, concerns over AI limitations, including readability and the potential for misinformation, must be addressed to ensure effective integration into healthcare.

(JMIR Preprints 05/10/2024:67101)
DOI: https://doi.org/10.2196/preprints.67101

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original paper

Authors

Amirreza Izadi, MD[1,2], Hesam Mosavari, MD[1,2], Ali Hosseininasab, MD[1,2], Ali Jaliliyan, MD[1,2], Arzhang Jafari, MS[1], Mohammadhosein Akhlaghpasand, MD[1,2], Aghil Rostami, MSc[3], Maziar Moradi-Lakeh, MD[4], Foolad Eghbali, MD[1,5]*

1. Surgery Research Center, School of Medicine, Rasool-E Akram Hospital, Iran University of Medical Sciences, Tehran, Iran.
2. Artificial Intelligence in Health Research Center, Iran University of Medical Sciences, Tehran, Iran.
3. Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
4. Gastrointestinal and Liver Diseases Research Center, Iran University of Medical Sciences, Tehran, Iran
5. Division of Minimally Invasive and Bariatric Surgery, Department of Surgery, Minimally Invasive Surgery Research Center, Rasool-E Akram Hospital, School of Medicine, Iran University of Medical Sciences, Tehran, Iran

***Corresponding author: Foolad Eghbali**

Address: Department of General Surgery, Hazrat-e Rasool General Hospital Niyayesh St, Tehran, Iran 14456 13131.

E-mail: foolade@yahoo.com

Tel: +98 912 826 3400

Authors information and contributions

Amirreza Izadi: izadiamirreza98@gmail.com/ 0009-0005-2286-7371 / Conceptualization, Literature Review, Methodology, Data Collection, Data Analysis, Writing Draft, Review & Editing

Hesam Mosavari: hesammosavari@gmail.com / 0000-0001-9701-8490 / Conceptualization, Literature Review, Methodology, Data Collection, Data Analysis, Writing Draft, Review & Editing

Ali Hosseininasab: alihosseininasab@gmail.com / 0009-0002-3954-4104 / Data Collection, Writing Draft, Review & Editing

Ali Jaliliyan: jaliliyan.a2020@gmail.com/ 0009-0005-0842-9041 / Data Collection, Data Analysis, Review & Editing

Arzhange Jafari: arzhang.jafarii@gmail.com / 0009-0002-6121-1488 / Literature Review, Data Collection, Writing Draft

Mohammadhosein Akhlaghpasand: Akhlaghpasandm@yahoo.com / 0000-0001-7514-5931 / Literature Review, Data Collection, Writing Draft

Aghil Rostami: aghil.rostamii@yahoo.com / 0000-0001-6651-1354 / Data Analysis, Review & Editing

Maziar Moradi-Lakeh: mazmoradi@gmail.com / 0000-0001-7381-5305 / Data Analysis, Review & Editing, Project Supervision

Foolad Eghbali: foolade@yahoo.com / 0000-0002-5720-3488 / Conceptualization, Data Collection, Writing Draft, Review & Editing, Project Supervision

Patient Education in Bariatric Surgery: Can Artificial Intelligence-based Chatbots Bridge the Knowledge Gap?

Abstract

Background: The global obesity epidemic continues to pose significant health challenges, with an increasing number of individuals seeking Metabolic and Bariatric Surgery (MBS), particularly laparoscopic sleeve gastrectomy (LSG). However, many MBS centers face resource limitations that limit the availability of adequate patient education, leading to knowledge gaps among patients who require more information. This consequently leads patients to seek information online increasingly. Artificial intelligence (AI) based chatbots like ChatGPT offer a promising tool to provide reliable and accessible medical information. However, concerns about the accuracy, reliability, and comprehensiveness of AI-generated responses still need to be addressed.

Objective: This study aims to evaluate the effectiveness of AI-based chatbots in answering frequently asked patient questions about LSG and compare their performance with that of bariatric surgery experts.

Methods: The study included four fellowship-trained minimally invasive surgeons (MISs), nine minimally invasive surgery fellows (MIFs), and two general practitioners (GPs) involved in MBS multidiciplinary team, forming the expert group. Seven AI chatbots, including ChatGPT versions 3.5 and 4, Bard, Bing, Claude, Llama, and Perplexity, were selected based on public availability. Forty patient questions about LSG were derived from social media, MBS organizations, and online patient forums. Experts and chatbots answered these questions, and their responses were scored for accuracy and comprehensiveness using a 5-point scale. Statistical analyses were performed to compare group performance.

Results: Chatbots demonstrated a higher overall performance score (2.55 ± 0.95) compared to the expert group (1.92 ± 1.32, P < .001). Among chatbots, ChatGPT-4 achieved the highest performance (2.94 ± 0.24), while Llama had the lowest (2.15 ± 1.23). Expert group scores were highest for MISs (2.36 ± 1.09), followed by GPs (1.90 ± 1.36) and MIFs (1.75 ± 1.36). The readability of chatbot responses was assessed using Flesch-Kincaid scores, revealing that most responses required reading levels between the 11th grade and college level. Furthermore, chatbots exhibited fair reliability and reproducibility in response consistency, with ChatGPT-4 showing the highest test-retest reliability.

Conclusion: AI-based chatbots provide reliable and comprehensive answers to common patient questions about LSG. These chatbots could play a significant role in patient education. Still, concerns over AI limitations, including readability and the potential for misinformation, must be addressed to ensure effective integration into healthcare.

Keywords: Bariatric surgery, Artificial intelligence, chatbots, patient education, laparoscopic sleeve gastrectomy.

Introduction

Obesity and bariatric surgery

The obesity epidemic poses a significant global health challenge; more than 1 billion people worldwide suffer from obesity. With a rising trend, WHO estimates that by 2025, approximately 167 million people will become less healthy because of obesity and obesity-related diseases [1]. Obesity is a risk factor for a variety of conditions, for instance, diabetes, cardiovascular disease, cancer, non-alcoholic fatty liver disease, cirrhosis, and liver failure [2]. It contributes substantially to morbidity and mortality rates [3], leading to increased healthcare expenditures and strain on human resources within the healthcare system, affecting not only those directly impacted but also society [4–6].

Metabolic and Bariatric Surgery (MBS) has emerged as a solution, offering substantial and durable weight loss [7,8]. Among different approaches in MBS, laparoscopic sleeve gastrectomy (LSG) is a fairly safe weight loss surgical procedure that represents 46% of all bariatric surgeries worldwide [9,10], with a significant positive impact on improving obesity comorbidities such as hypertension, type 2 diabetes, dyslipidemias, and obstructive sleep apnea [10,11].

Patient education and online resources

Patients frequently seek to comprehend and gain insight into their medical conditions and treatment options. Several studies have underscored the significance of patient education, particularly in chronic diseases [12]. While MBS represents a singular operative intervention, multidisciplinary and regular post-operative care in MBS patients is consistent with the model of care for chronic disease. Patient education improves the degree of weight loss achieved regardless of the procedure [13]. It also plays a crucial role in adherence to lifestyle modifications and improves compliance and weight loss outcomes following MBS [14,15]. Sufficient education can empower patients to make informed decisions, reduce anxiety, and actively engage in treatment deliberations, aiding them in comprehending various facets of the proposed care plan and ultimately refining the course of recovery and surgical outcomes [12,15,16].

Traditionally, patients have sought medical information from physicians, but inherent limitations constrain this avenue, especially in access to physicians and its costs. Hence, patients increasingly turn to online sources for information about their conditions, especially when they feel healthcare providers fail to answer [17,18]. The ambiguity inherent in acquiring knowledge and obtaining information from the internet is highlighted by the fact that various search engines produce significantly different results for identical search queries, and slight variations in wording can lead to significantly divergent outcomes. Consequently, it becomes challenging to anticipate what users will encounter during their searches. Furthermore, ranking a website in search results is complex, relying on several factors, such as the website's popularity and inbound links. This process is subject to manipulation, and a site's quality does not always match its ranking or search visibility [19,20]. This algorithm makes health information vulnerable to marketing, propaganda, and the spreading of false information. There is also heterogeneity in the reliability and readability of health-related content for the general population [21,22].

Large language models for patient education

In response to this complicated web of challenges, the potential exhibited by chatbots could be utilized to provide patients with sufficient and accurate information. Large language models (LLMs) have recently become publicly available through conversational AI models called “Chatbots.” These models are trained using various online resources, including books and articles [23]. This training enables chatbots to comprehend the intricacies of users' inquiries and adeptly address a wide range of user queries, including those related to medical concerns [24].

Previous studies have highlighted the valid clinical insights of AI-based chatbots [25,26]. Despite their promising performance, numerous articles highlighted their risks and weaknesses, such as inaccurate references, the quality of training datasets, and AI hallucinations, prompting warnings against using them to obtain medical information and research [27–30]. Furthermore, as an example, ChatGPT and GPT-4 were trained on data only until 2021 and generally lack access to information hidden behind paywalls [23]. Due to the proprietary nature of their training, it is challenging to model inherent biases and errors within the model beforehand [31].

Integrating AI in this delicate domain necessitates carefully examining its capabilities, reliability, risks, and benefits toward patients' health and well-being. This study aims to critically evaluate the performance of the chatbots in providing adequate and accurate answers to patient questions on LSG, comparing their responses to bariatric surgery experts as the conventional way of ascertaining medical and health information. By doing so, it seeks to shed light on the chatbot's strengths, limitations, and potential to complement conventional patient education methods.

Materials and Methods

Study design

This observational study was designed to compare the performance of chatbots and bariatric surgery experts at the Center of Minimally Invasive and Bariatric Surgery, Rasool-E Akram Hospital, affiliated with Iran University of Medical Sciences in answering patient frequently asked questions (FAQs) about LSG. This study was conducted using the principles outlined in the Declaration of Helsinki and received approval from the Institutional Review Board (IRB) under the ethics code IR.IUMS.REC.1402.982.

Participant

Four fellowship-trained, board-certified minimally invasive surgeons (MISs), nine minimally invasive surgery fellows (MIFs), and two general practitioners (GPs) participated in the study. For MISs, inclusion required active practice and completing over 50 LSG surgeries in the previous year. The authors of this study were excluded. MIFs were included if they had completed at least six months of training and had performed over 50 LSG surgeries. GPs (specially trained and responsible for initial patient consultation and guiding them throughout getting approved for MBS) were included if they had more than one year of experience. We will use the terms “expert group” and “bariatric surgery experts,” referring to all MISs, MIFs, and GPs.

Chatbot Selection

We selected seven publicly available chatbots: ChatGPT versions 3.5 and 4, Bard (dated October 30, 2023), Bing (latest version as of December 2023), Anthropic's Claude (version 2.0), Llama (version 2b7), and Perplexity (latest version as of December 2023). We chose chatbots based on their public availability and general popularity. We excluded chatbots designed for narrow, specialized purposes to focus the analysis on versatile, general-purpose conversational agents.

Questions Selection

We created the question set by gathering input from several sources that patients commonly access to learn about bariatric surgery. We retrieved questions from social media platforms (e.g., Twitter, Instagram) and the FAQ section of leading organizations in MBS like the International Federation for the Surgery of Obesity and Metabolic Disorders (IFSO), the American Society for Metabolic and Bariatric Surgery (ASMBS), Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), and the Iran Society for Metabolic and Bariatric Surgery (ISMBS) websites. Additionally, we reviewed online patient forums and education materials to find common patient questions.

Through this process, we compiled an initial pool of 200 questions on all aspects of LSG. After removing duplicate, unclear, too general, overly specialized, and ambiguous questions, we narrowed the list to 80 viable questions for the question set. We categorized 80 questions into five topics, each with at least ten questions: outcome and expectation, pre-operative care, recovery and post-operative care, risks and complications, and lifestyle modifications.

A team of three fellowship-trained, board-certified MISs reviewed the final pool of 80 questions to identify and eliminate any ambiguities. By refining together and focusing on clarity while maintaining the original objectives and topics covered, we arrived at 40 clear, focused, patient-centered questions for LSG surgery. We also considered accessibility by trying to avoid complex medical terminology and encouraging patient empowerment through the style and content of the questions. The English version of all Questions is available in Multimedia Appendix 1.

Questions originally in English were translated into Persian using a standard forward-backward translation process, which included a cognitive interview with five surgical residents and five bariatric surgery candidates. This focus group offered feedback, which we used to identify and amend ambiguous phrases or sentences to create a final version aligned with patients' real concerns. Subsequently, two MISs meticulously reviewed the translated questions to preserve the original aims and topics, making minor changes to enhance clarity and readability. Questions sourced from Persian materials underwent a similar translation process into English, facilitated by cognitive interviews with five surgical residents proficient in both languages. Again, two MISs meticulously reviewed the translated questions to preserve the original aims and topics. English questions were used in their original form as they were extracted.

Questions were presented to expert groups in Persian. The answers by expert groups were also in Persian, and the Persian transcripts were examined. Questions were presented to the chatbots in English, and English answers were analyzed.

Answer Sheet

Two MISs compiled the answer sheet, including critical clinical practice recommendations based on MBS guidelines from SAGES, IFSO, and ASMBS [7,32–34]. If the answer to a question could not be found in the guidelines, we utilized the most recent high-quality systematic reviews and meta-analyses and the answers provided in the FAQ section of the ASMBS or IFSO official websites [35,36]. An independent fellowship-trained, board-certified minimally invasive surgeon assessed the answer sheet to determine its completeness and comprehensiveness. Feedback on item content, wording, and relevance was integrated to improve the final answer sheet utilized in the present study to compare and evaluate the responses of bariatric surgery experts and chatbots. The English version of the answer sheet is available in Multimedia Appendix 1.

Taking the Exam

We invited participants to take an exam at a designated location. Their verbal responses were audio recorded. Two surgical interns independently transcribed the recordings word-for-word into text documents. One author then reviewed the original audio recordings and final transcripts for accuracy. Chatbots were presented with questions one at a time on separate tabs to prevent previous answers from influencing responses. Each chatbot was asked the same set of questions twice within the study period to evaluate the consistency and reliability of responses. All responses were unidentified and compiled into spreadsheets for scoring.

Scoring

Two MISs independently rated the accuracy and comprehensiveness of each response based on a scoring sheet on a 5-point rating scale. The scale was defined as 3 = Comprehensive and accurate; 2 = Accurate but incomplete; 1 = Partially accurate; 0 = Inaccurate or missing key information; -1 = Completely inaccurate, potentially harmful, or misleading. Any differences in ratings among the two reviewers were resolved through discussion and, if necessary, the involvement of a third reviewer to reach a consensus. A panel of three MISs agreed to score the missing answer and "I don't know" responses as completely inaccurate and potentially harmful. The final score for each individual chatbot and bariatric surgery expert was calculated by summing the ratings across all questions. Mean and standard deviations (SD) were used to report the overall performance of each group of bariatric surgery experts (e.g., MISs, MIFs, GPs) as well as the chatbot group. Each chatbot's mean and SD reflect the average score obtained from two test iterations.

The readability assessment of chatbots' responses (in English) was done using multiple readability metrics, including Flesch Kincaid Grade Level (FKGL), Flesch-Kincaid Reading Ease (FRE), Flesch Readability Category (FRC), Flesch School Level (FSL), Gunning Fog Score (FGI), Coleman-Liau Index (CLI), Automated Readability Index Score (ARI-S), and Automated Readability Index Age

(ARI-A). The readability assessment was performed using the readability library version 0.3.1 (Andreasvc, 2019) and Python version 3.12.1 (Python Software Foundation, 2023).

Statistical analysis

We conducted all analyses using IBM SPSS Statistics for Windows, Version 27.0 (IBM Corp., 2020). We used descriptive statistics to characterize the distribution of performance scores for each expert group and chatbot. In assessing the normality of the data distribution, we employed Kolmogorov-Smirnov, using an alpha level of 0.05 for significance. We compared the mean accuracy of overall scores and scores for each subcategory between groups using one-way ANOVA analysis for normally distributed data and the Kruskal-Wallis test for non-normally distributed data. The analysis was used to assess group performance and identify significant differences in accuracy between MISs, MIFs, GPs, and chatbots.

Additionally, the percentage of irrelevant/harmful responses was compared to assess the critical error difference between groups. We performed a test-retest analysis using the Intraclass Correlation Coefficient (ICC) to assess the repeatability and consistency of chatbot responses over time. A One-Way Random Effects Model with Consistency was employed to measure the degree of agreement between the ratings. ICC values were interpreted as follows: values less than 0.40 indicated poor reliability, values between 0.40 and 0.59 indicated fair reliability, values between 0.60 and 0.74 indicated good reliability, and values of 0.75 or higher indicated excellent reliability.

Results

In this study sample, we compared the performance of bariatric surgery experts and AI-based chatbots in answering 40 frequent patients' questions, covering a range of inquiries, including outcome and expectation, pre-operative care, recovery and post-operative care, risks and complications, and lifestyle modifications (Table 1).

Chatbot group performance versus expert group performance

The overall chatbots' performance score (mean 2.55, SD 0.95) was higher than the overall bariatric surgery experts' performance score (mean 1.92, SD 1.32), which was statistically significant ($P < .001$) (Table 1, Figure 1).

Expert group performance

Among the bariatric surgery experts, MISs provided the most accurate and comprehensive answers (mean 2.36, SD 1.09), followed by GPs (mean 1.90, SD 1.36) and MIFs (mean 1.75, SD 1.36) (Table 1, Figure 1). There was a statistically significant difference between MISs and MIFs bariatric surgery experts ($P = .011$) (Table 2).

Chatbot group performance

The range of performance scores for chatbots was from 2.15 (Llama) to 2.94 (ChatGPT-4). Chat-GPT-4 achieved the highest performance score (mean 2.94, SD 0.24), closely followed by Chat-GPT-3.5 (mean 2.91, SD 0.33) (Table 1). There was no statistically significant difference between the performance scores of different chatbots (Table 2).

Top-performing chatbots versus expert groups

The top two chatbots - ChatGPT-4 and ChatGPT-3.5 - achieved higher performance scores than MISs (the top-performer among expert groups) (mean 2.94, SD 0.24; mean 2.91, SD 0.33; vs mean 2.36, SD 1.09, respectively) (Table 1). However, there was no statistically significant difference between them (Table 2). Additionally, there was a statistically significant difference between the other two expert groups (e.g., MIFs, GPs) and the top two chatbots. The pairwise comparison p-values between different chatbots and expert groups are shown in Table 2.

Worst performing chatbots versus expert groups

The worst-performing chatbot, Llama (mean 2.15, SD 1.23), achieved a lower performance score than MISs (mean 2.36, SD 1.09) but higher than other expert groups (Table 1). There was no statistical difference between Llama and any of the expert groups (Table 2).

Domain-based Performance

In domain-based analysis, expert groups' performance scores ranged from 1.62 (SD 1.39) (Lifestyle modification) to 2.34 (SD 1.00) (Outcome & expectation) (Table 1, Figure 2). The overall Chatbot performance scores were significantly higher than the average scores of expert groups across all categories ($P < .05$). MISs achieved a higher performance score among expert groups, but the difference was not statistically significant. ChatGPT-3.5 and ChatGPT-4 achieved higher performance than other chatbots, but no statistically significant differences were observed (Table 2). Although the top-performing chatbots (e.g., ChatGPT-4, ChatGPT-3.5) scored higher than MISs, MISs scored higher than the worst-performing chatbots (Llama) in all categories except for the recovery and post-operative care category. MISs also scored higher than Perplexity in outcome and expectation and pre-operative care categories, but these differences were not statistically significant (Table 2, Multimedia Appendix 2). The only statistically significant differences were evident between MIFs and top-performing chatbots in recovery and post-operative care, risk and complications, and lifestyle modifications categories (Multimedia Appendix 2).

Table 1. The score of experts and chatbots for answering common questions about laparoscopic sleeve gastrectomy in total and each category.

	Total	Outcome & expectation	Pre-operative care	Recovery & post-operative care	Risks & complications	Lifestyle modifications
MISs[a]	2.36 (±1.09)[d]	2.67 (±0.58)	2.42 (±0.97)	2.50 (±1.21)	1.86 (±1.32)	2.25 (±1.13)
MIFs[b]	1.75 (±1.36)	2.26 (±1.08)	1.42 (±1.46)	1.90 (±1.41)	1.64 (±1.29)	1.38 (±1.39)
GPs[c]	1.90 (±1.36)	2.11 (±1.18)	1.58 (±1.50)	2.33 (±1.24)	1.79 (±1.37)	1.56 (±1.50)
Bing	2.46 (±1.03)	2.50 (±0.98)	1.58 (±1.38)	2.78 (±0.55)	2.29 (±1.33)	2.83 (±0.51)
Bard	2.44(±1.02)	2.94 (±0.23)	2.08 (±1.08)	2.11 (±1.41)	2.57 (±0.65)	2.39 (±1.09)
Claude	2.59 (±0.87)	2.72 (±0.57)	2.08 (±1.08)	2.67 (±0.84)	2.64 (±0.84)	2.67 (±0.97)
Llama	2.15 (±1.23)	2.28 (±1.23)	2.00 (±1.28)	2.17 (±1.34)	2.29 (±0.99)	2.00 (±1.37)
Perplexity	2.36 (±1.15)	2.39 (±1.09)	2.25 (±0.97)	2.50 (±1.15)	2.14 (±(1.46)	2.44 (±1.15)
ChatGPT-3.5	2.91 (±0.33)	3.00	2.58 (±0.51)	3.00	3.00	2.89 (±0.47)
ChatGPT-4	2.94 (±0.24)	3.00	2.67 (±0.49)	3.00	2.93 (±0.27)	3.00

Expert group	1.92 (±1.32)	2.34 (±1.00)	1.69 (±1.41)	2.10 (±1.36)	1.71 (±1.29)	1.62 (±1.39)
Chatbot group	2.55 (±0.95)	2.62 (±0.87)	2.18 (±1.04)	2.60 (±0.97)	2.55 (±0.95)	2.60 (±0.94)

[a]Minimally Invasive Surgeons

[b]Minimally Invasive Surgery Fellows

[c]General Practitioners

[d] The numbers represent the mean ± standard deviation

Table 2. Pairwise comparison p-values between groups, derived from ANOVA and Tukey's post hoc tests

	MISs[a]	**MIFs**	**GPs**	**Bing**	**Bard**	**Claude**	**Llama**	**Perplexity**	**ChatGPT-3.5**
MIFs[b]	0.011*								
GPs[c]	0.507	0.997							
Bing	1.000	0.029*	0.422						
Bard	1.000	0.038*	0.481	1.000					
Claude	0.979	0.007*	0.193	1.000	1.000				
Llama	0.990	0.517	0.987	0.946	0.967	0.729			
Perplexity	1.000	0.084	0.668	1.000	1.000	0.994	0.996		
ChatGPT-3.5	0.262	<0.001 *	0.014*	0.699	0.637	0.993	0.111	0.452	
ChatGPT-4	0.216	<0.001 *	0.014*	0.637	0.574	0.900	0.091	0.394	1.000

[a]Minimally Invasive Surgeons

[b]Minimally Invasive Surgery Fellows

[c]General Practitioners

* statistically significant

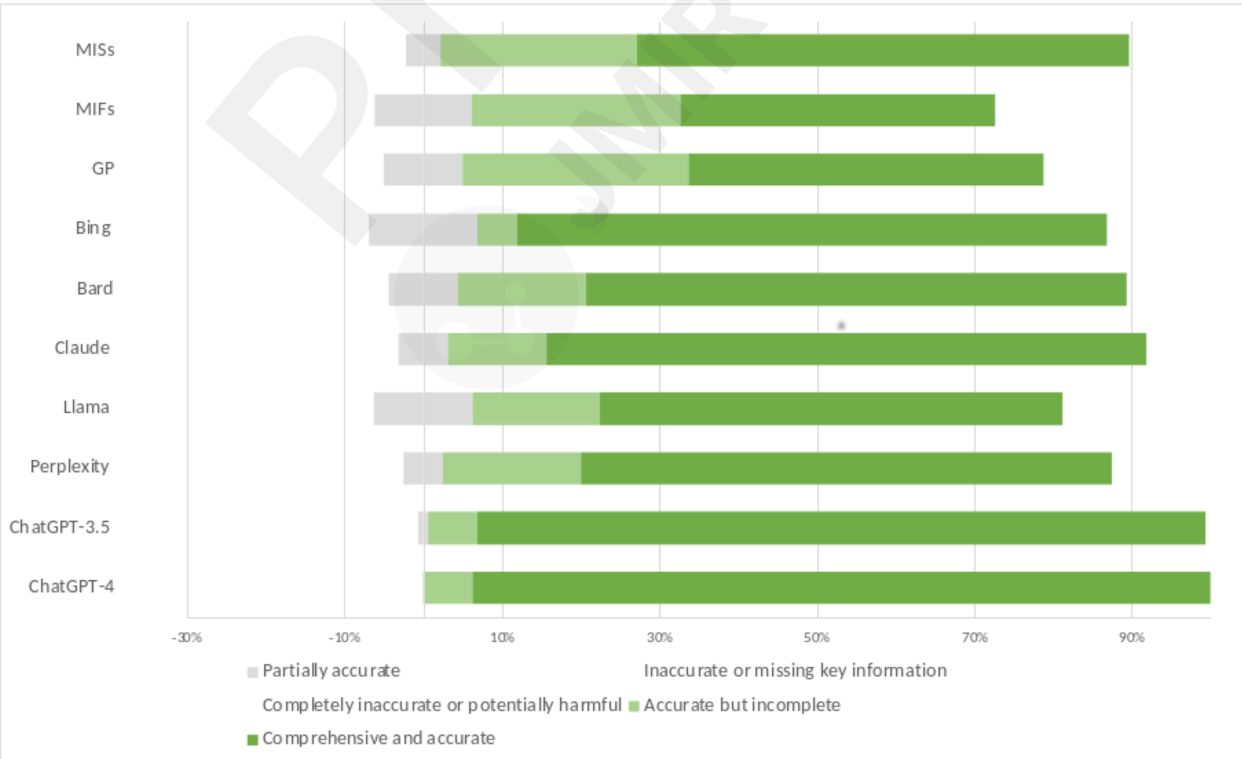


Figure 1. Responses accuracy of each group on all questions. MISs; Minimally invasive surgery fellows, MIFs; General practitioners, GPs.

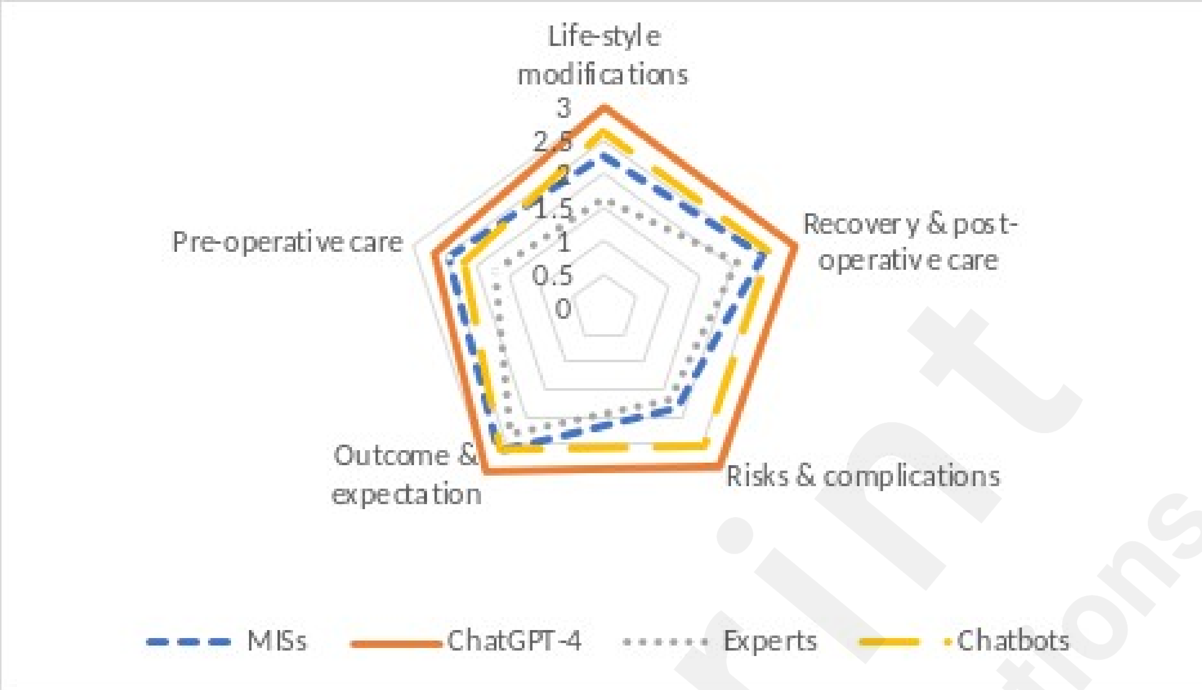


Figure 1. Comparison of performance score between the top-performing chatbot (ChatGPT-4), the top-performing expert group (MISs), all chatbots (Chatbots), and all expert groups (Experts) across various categories. Minimally invasive surgeons, MISs.

Incorrect and misleading answers

In this study, responses rated as '0' (Inaccurate or missing key information) or '-1' (Completely inaccurate, potentially harmful, or misleading) on our scoring system were considered "incorrect." Out of 640 expert responses, 111 (17.34%) were incorrect. Among incorrect responses, 9.84% (63/111) were classified as potentially harmful or misleading. Most of the wrong or misleading answers belonged to the lifestyle modification category (39/144, 27.08%), and 43.59% (17/39) of them were potentially harmful or misleading.

In contrast, chatbots demonstrated a lower rate of incorrect responses, accounting for 5.71% (32/560) of their total responses, of which 50% (16/32) were marked as potentially harmful or misleading. Among chatbots, Llama had the highest percentage of incorrect answers, 12.5% (10/80), of which 50% (5/10) were potentially harmful or misleading. None of the responses from ChatGPT-3.5 and ChatGPT-4 were marked as incorrect.

Readability

The overall Flesch-Kincaid Reading Ease score for all the chatbots was 36.52 (SD 5.79), translating into a range from 11th grade to college level, based on Flesch grade (Table 3). ChatGPT-3.5 showed the lowest readability, with a reading score of 29.63 associated with "very confusing," and Bard showed the highest readability, with a reading score of 45.91 associated with "difficult." Other chatbots scored within the relatively "difficult" range, with reading scores between 31 and 45.91. Further analysis suggests their language requires the reading skills expected of 11th-grade to college-graduate students. Figure 3 shows the Flecsh-Kinkade Reading Ease versus the performance scores.

Table 3. Readability analysis for chatbots

	Bard	Llama	Claude-2	Bing	Perplexity	Chatgpt-4	Chatgp-3.5	Overall (mean ± sd)
FKGL[a]	11	13	12	11	13	13	13	-
FRE[b]	45.91	33.90	37.07	41.84	36.28	31.01	29.64	36.52 ± 5.79

FRC[c]	Difficult	Difficult	Difficult	Difficult	Difficult	Difficult	Very Difficult	-
FSL[d]	College	College	College	College	College	College	College Graduate	-
FGI[e]	13.89	15.72	15.06	14.22	15.96	15.71	16.65	15.32 ± 0.98
CLI[f]	12.83	14.55	14.63	12.88	12.53	15.25	15.53	14.03 ± 1.25
ARI-S[g]	11.23	13.15	12.27	10.85	11.80	12.77	13.9	12.28 ± 1.08
ARI-A[h]	17-18	24+	18-24	16-17	17-18	18- 24	24+	-

[a]Flesch Kincaid Grade Level

[b]Flesch-Kincaid Reading Ease

[c]Flesch Readability Category

[d]Flesch School Level

[e]Gunning Fog Score

[f]Coleman-Liau Index

[g]Automated Readability Index Score

[h]Automated Readability Index Age.

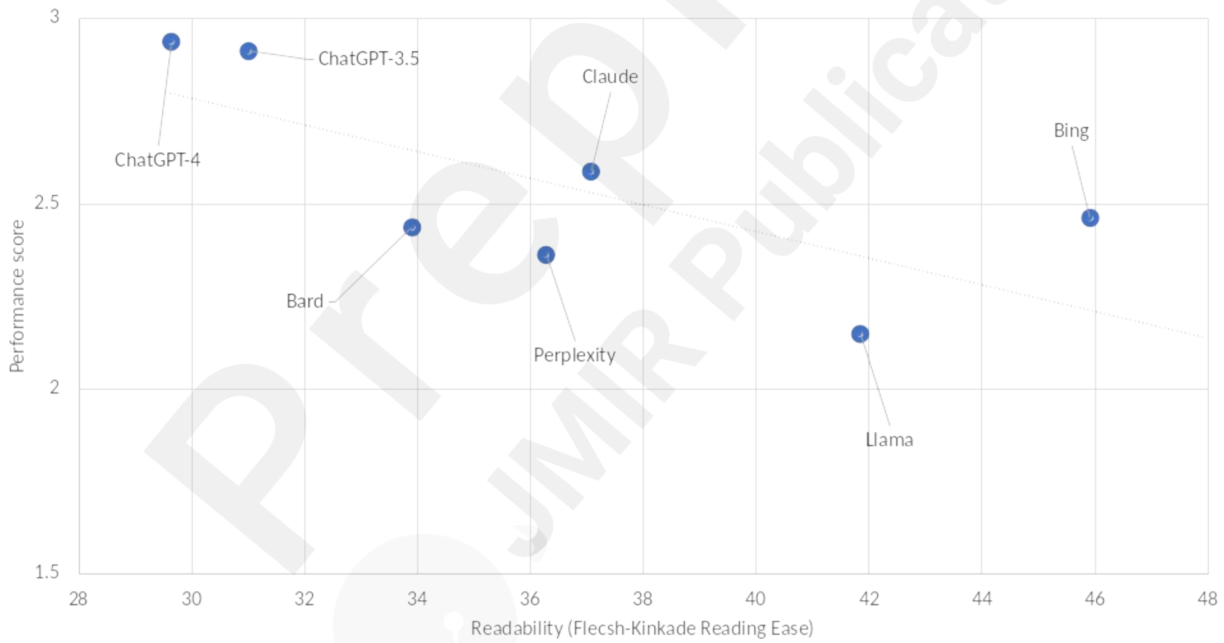


Figure 3. The inverse relationship between performance score and readability (measured by the Flesch-Kinkade Reading Ease) for various chatbots, with a downward trend indicating that higher accuracy correlates with lower readability.

Reproducibility and reliability of responses

The results showed good to poor reliability, with ICC ranging from 0.23 to 0.61. ChatGPT-3.5 had the highest reproducibility, achieving 95% of the same scores for 40 questions (ICC = 0.586, *P* = .003). This was closely followed by ChatGPT-4 with 92% of identical scores for answers (ICC = 0.541, *P* = .008) (Table 4).

Table 4. Reproducibility and reliability of chatbot responses in rest-retest analysis

	Identical scores for answers, n (%)	ICC[a]	P-value	Reliability
Bard	20 (50%)	0.212	.228	Poor

Llama	19 (47.5%)	0.281	.151	Poor
Claude	25 (62.5%)	0.362	.081	Poor
Perplexity	25 (62.5%)	0.540	.008	Fair
Chat-GPT4	37 (92%)	0.541	.008	Fair
Bing	25 (62.5%)	0.574	.004	Fair
Chat-GPT3.5	38 (95%)	0.586	.003	Fair

[a]Intraclass Correlation Coefficient

Discussion

Principal finding

Recently, AI-based chatbots have gained popularity and are now commonly used for various purposes, including looking for medical and health information [37]. We conducted this study to evaluate chatbots' educational value and limitations. This study found that AI-based chatbots can provide accurate and comprehensive answers with fair reliability to frequently asked questions about LSG compared to bariatric surgery experts. However, chatbot answers require a minimum level of education to be readable by users.

Accuracy and comprehensiveness

Differing from previous studies [38–40], we selected questions that chatbots are realistically expected to answer correctly. Chatbots are trained using widely available information from the internet and do not undergo specialized medical training [41]. They are more suited for providing general information to patients rather than answering complex medical queries. Therefore, We focused on general patient questions that may not been addressed by surgeons or other members of a bariatric multidisciplinary team. These are typically the questions that patients turn to online resources to find answers to. However, using search engines for this purpose has drawbacks, such as providing biased information for marketing purposes [19].

Chatbots' responses were more accurate and comprehensive than those of the expert group, with a higher percentage of accurate and comprehensive answers and a lower rate of misleading and potentially harmful answers. However, MISs (the best-performing expert group) were surpassed in overall performance and domain-based performance by the top-performing chatbots (ChatGPT-4 and ChatGPT-3.5).

Our findings align with the study conducted by Samaan et al., which showed that ChatGPT-3.5 responded comprehensively to 131 out of 151 patient questions (86.8%). Similarly, in our study, 426 out of 560 (76.07%) of chatbots' responses scored the highest (comprehensive and accurate). A study by Haver et al. found that ChatGPT provided appropriate responses for 88% of common questions related to breast cancer prevention and screening, as evaluated by fellowship-trained breast radiologists [42]. In a study by Ye et al., rheumatologists and patients evaluated the responses generated by ChatGPT-4 and rheumatologists using a survey focused on patient questions about rheumatology. Patients' ratings exhibited a parallel trend, with no statistically significant difference between physician and chatbot-generated answers. Conversely, rheumatologists rated the physician-generated responses superior in comprehensiveness and accuracy [39]. This discrepancy in findings may stem from the methodology employed, particularly in selecting and curating questions for the study.

Incorrect responses and errors

The best-performing chatbots (ChatGPT-4 and ChatGPT-3.5) did not receive 0 or -1 scores for any of their answers (meaning Inaccurate or missing key information and completely inaccurate, potentially harmful, or misleading, respectively). However, 32 of 560 (5.71%) chatbots' answers received either a -1 or 0 score. Discrepancies arise when chatbots encounter questions that demand responses based on heterogeneous online information.

Q: How long until I can start exercising after laparoscopic sleeve gastrectomy? Range of weeks

Bard: Light exercises […] can be resumed within the first week after surgery. Moderate exercises […] can be resumed within the second week after surgery. Strenuous exercises: Running, heavy lifting, and contact sports can be resumed within the fourth week after surgery.

(Accessed dated December 23, 2023)

The issue of LLMs assigning weights to online resources during the learning process needs to be clarified. It lacks clarity, particularly regarding how different resources with varying levels of reliability and quality should be prioritized. There is currently no consensus on the ideal time to begin exercising after bariatric surgery, with protocols ranging from 7 days to 6 months or more post-surgery [43]. The protocol for post-operative exercise in our center prohibits patients from starting strenuous activity and heavy lifting for at least eight weeks. Additionally, relevant guidelines recommend avoiding lifting weights of more than 15 lbs during the first six weeks after surgery [44]. To minimize the risk of AI-based chatbot critical errors, educating the public about AI chatbot limitations, such as AI hallucinations and reasoning constraints, is essential to address negative sentiments and promote effective usage techniques [45]. For example, self-checking (asking the question multiple times) to see if any divergent or contradictions happen could help find and avoid chatbot errors [46]. Also, some chatbots like Bing and Perplexity cite relevant references when responding to queries. Despite their relatively lower performance compared to ChatGPT and Claude in our study, the ability to fact-check answers against corresponding references provides a valuable opportunity to verify the accuracy and reliability of the information provided.

Readability

Chatbots with better performance tend to produce responses of greater complexity, characterized by lower Flesch reading ease scores falling within the "very confusing" range. Conversely, models such as Bard and Bing (with lower performance scores) tend to generate relatively more easily understandable answers. However, our findings indicate a potential literacy gap, with most chatbots mandating individuals from an 11th to 14th-grade education level for comprehension by the average patient. Notably, ChatGPT-4 exhibits enhanced comprehension of questions, adeptly providing only the relevant information requested without unnecessary detail. This sensitivity to phrases such as "yes or no" or "in three sentences" is less pronounced in ChatGPT-3.5 and Claude. However, they provide sufficient and relevant information as requested by the questions.

Q: Can laparoscopic sleeve gastrectomy lead to long-term nutritional deficiencies? Yes or No

ChatGPT-4: Yes.
(Accessed dated December 28, 2023)

Q: Can laparoscopic sleeve gastrectomy lead to long-term nutritional deficiencies? Yes or No

Calude-2: Yes, laparoscopic sleeve gastrectomy can lead to potential nutritional deficiencies in the long term. Because a portion of the stomach is removed during the procedure, the body's ability to absorb some vitamins and minerals like iron, vitamin B12, folate, thiamine, and calcium can be reduced over time. Patients must

> *closely follow their bariatric team's nutritional guidelines after surgery, which include taking prescribed vitamin and mineral supplements daily, following up with blood work, and meeting protein intake recommendations. Most sleeve gastrectomy patients can avoid significant vitamin or mineral deficiencies after surgery with proper supplementation and nutritional counseling. However, non-compliance with follow-up care can increase the risks over the long run.*
>
> *(Accessed dated December 21, 2023)*

Chatbot responses are generally easier to read and understand than online materials [47], and chatbots can always be asked to explain matters in a simpler and easy-to-understand language. It is also worth mentioning that in the study by Ye et al., patients rated chatbot responses as easy to read as physician-written responses to questions [39].

Reproducibility and reliability

The reproducibility and reliability of answers varied significantly among different chatbots. The top-performing chatbots, such as ChatGPT-3.5 and ChatGPT-4, showed superior reliability, achieving predominantly similar performance scores. Generally, chatbots demonstrated varying degrees of consistency, which underscores the importance of evaluating their performance in clinical contexts. This finding aligns with a previous study by Saman et al., which reported similar reproducibility for ChatGPT; however, their analysis did not include other chatbots that demonstrated poor reliability and reproducibility [38]in our study. Furthermore, a study by Kochanek et al. found that factors such as the day of interaction do not influence the quality of answers provided by chatbots. It is important to note that, currently, most chatbots lack the capability to retain the memory of previous queries, limiting their ability to offer personalized responses. However, recent developments indicate that this memory feature is being integrated into chatbots to enhance user experience and provide more tailored information. Future research should also evaluate the effect of retaining memory on the accuracy and helpfulness of chatbots for patients.

Future of chatbots in patient education

The performance differences among chatbots can be significant and may be influenced by variability in architecture, training methodology, and safety practices. Further research is needed to identify the factors contributing to chatbot performance and develop standardized guidelines for development and evaluation in healthcare settings. While chatbots show promise in transforming patient education by providing accessible and often comprehensive information on medical inquiries, there are notable areas for improvement, such as empathy [48].

Artificial intelligence's growing prominence, especially in LLM and multimodal large language models (MLLM) healthcare, has fueled discussions surrounding surpassing physicians in diverse facets of patient care. AI tools capable of managing patients' entire journey, from providing information to making treatment decisions and facilitating personalized medicine and follow-ups, are becoming increasingly relevant. However, the feasibility of AI overtaking physicians hinges on several factors, including the current capabilities of AI technology and the multifaceted nature of medical tasks.

Reflecting on past experiences, notably the "artificial intelligence winter" periods, we recognize the danger of unrealistic expectations surrounding AI's capabilities. While the aspiration for AI tools with 100% accuracy in diagnosis and treatment, coupled with high levels of empathy, is commendable, achieving such a feat in the near future is unrealistic. Even if such a tool were developed, transitioning to and integrating it into healthcare systems would require time and consideration.

In this context, focusing on using AI for patient education is a more achievable and practical goal. AI offers advantages such as round-the-clock availability and acceptable knowledge, which can complement physicians' efforts in informing and educating patients. However, to employ AI's full potential in this capacity, it is essential to establish specialized and standardized tests to evaluate and benchmark AI tools. For example, Barletta et al. have proposed a method of clinical chatbot evaluation based on the “quality in use” of ISO/ IEC 25010. These approaches will ensure that AI's contributions to patient care are effective and reliable.

Limitations

While we included prominent chatbots, our research does not include all available AI-based platforms. However, the head-to-head comparison methodology offers valuable insights into AI-assisted bariatric surgery consultation. This study does not include that chatbots can help simplify and rephrase their responses and be asked to explain the matter further. Also, a readability analysis was not conducted because the expert responses were in Persian. In this study, the questions were phrased by experts. However, how the patients phrase their questions could significantly impact the answers provided by chatbots. Patients' education and literacy also influence their competence in drafting questions for chatbots. Given the rapid advancements in this area, larger-scale, multi-site evaluations comparing evolving AI solutions to established standards of care are crucial moving forward.

Conclusion

This study found that chatbots’ responses to patients’ questions are generally more accurate and comprehensive than those of bariatric surgery experts. AI-based chatbots have the potential to enhance patient education and support with better availability. Ongoing efforts to improve chatbot readability, reliability, and empathy are crucial for maximizing their effectiveness and integration into the healthcare system.

Acknowledgment

We thank Negin Eissazade, Sina Dibaei, Benyamin Kazemi, and Fatemeh Abbasi for their invaluable contributions to this study. This study did not receive any funding. AI tools, including ChatGPT (OpenAI, 2024), were used solely for enhancing the readability and for grammar checks of this manuscript. The intellectual content, study design, data analysis, and conclusions are entirely the result of human input and effort. All final edits and approvals were made by the authors.
Amirreza Izadi: Conceptualization, Literature Review, Methodology, Data Collection, Data Analysis, Writing Draft, Review & Editing; Hesam Mosavari: Conceptualization, Methodology, Data Collection, Writing Draft, Review & Editing; Ali Hosseininasab: Data Collection, Writing Draft, Review & Editing; Ali Jaliliyan: Data Collection, Data Analysis, Review & Editing; Arzhange Jafari: Literature Review, Data Collection, Writing Draft; Aghil Rostami: Data Analysis, Review & Editing; Maziar Moradi-Lakeh: Data Analysis, Review & Editing, Project Supervision; Foolad Eghbali: Conceptualization, Data Collection, Writing Draft, Review & Editing, Project Supervision

Competing interests

The authors declare that they have no competing interests.

Abbreviations

ASMBS: American Society for Metabolic and Bariatric Surgery
AI: Artificial Intelligence
ARI-A: Automated Readability Index Age
ARI-S: Automated Readability Index Score

CLI: Coleman-Liau Index
FKGL: Flesch Kincaid Grade Level
FRC: Flesch Readability Category
FSL: Flesch School Level
FRE: Flesch-Kincaid Reading Ease
FAQs: Frequently Asked Questions
GPs: General Practitioners
FGI: Gunning Fog Score
IRB: Institutional Review Board
IFSO: International Federation for the Surgery of Obesity and Metabolic Disorders
ICC: Intraclass Correlation Coefficient
ISMBS: Iran Society for Metabolic and Bariatric Surgery
LSG: Laparoscopic Sleeve Gastrectomy
LLMs: Large Language Models
MBS: Metabolic and Bariatric Surgery
MISs: Minimally Invasive Surgeons
MIFs: Minimally Invasive Surgery Fellows
MLLM: Multimodal Large Language Models
SAGES: Society of American Gastrointestinal and Endoscopic Surgeons
SD: Standard Deviations

References

1. World Obesity Day 2022 – Accelerating action to stop obesity. World Obesity Day 2022 – Accelerating action to stop obesity. 2024. Available from: https://www.who.int/news/item/04-03-2022-world-obesity-day-2022-accelerating-action-to-stop-obesity [accessed Apr 25, 2024]

2. Ansari S, Haboubi H, Haboubi N. Adult obesity complications: challenges and clinical impact. Ther Adv Endocrinol Metab 2020 Jun 22;11:2042018820934955. PMID:32612803

3. Abdelaal M, le Roux CW, Docherty NG. Morbidity and mortality associated with obesity. Ann Transl Med 2017 Apr;5(7):161. PMID:28480197

4. Xia Q, Campbell JA, Ahmad H, Si L, De Graaff B, Palmer AJ. Bariatric surgery is a cost-saving treatment for obesity—A comprehensive meta-analysis and updated systematic review of health economic evaluations of bariatric surgery. Obesity Reviews 2020 Jan;21(1):e12932. doi: 10.1111/obr.12932

5. Okunogbe A, Nugent R, Spencer G, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ Global Health BMJ Specialist Journals; 2021 Oct 1;6(10):e006351. PMID:34737167

6. Hecker J, Freijer K, Hiligsmann M, Evers SMAA. Burden of disease study of overweight and obesity; the societal impact in terms of cost-of-illness and health-related quality of life. BMC Public Health 2022 Jan 7;22(1):46. doi: 10.1186/s12889-021-12449-2

7. Nguyen NT, Brethauer SA, Morton JM, Ponce J, Rosenthal RJ, editors. The ASMBS Textbook of Bariatric Surgery. Cham: Springer International Publishing; 2020. doi: 10.1007/978-3-030-27021-6ISBN:978-3-030-27020-9

8. Yarigholi F, Bahardoust M, Mosavari H, Tehrani FM, Gholizadeh H, Shahmiri SS, Rezvani H, Oshidari B, Garakani K, Eghbali F. Predictors of Weight Regain and Insufficient Weight Loss According to Different Definitions After Sleeve Gastrectomy: a Retrospective Analytical Study. OBES SURG 2022 Dec;32(12):4040–4046. doi: 10.1007/s11695-022-06322-3

9. Xu C, Yan T, Liu H, Mao R, Peng Y, Liu Y. Comparative Safety and Effectiveness of Roux-en-Y Gastric Bypass and Sleeve Gastrectomy in Obese Elder Patients: a Systematic Review and Meta-analysis. OBES SURG 2020 Sep 1;30(9):3408–3416. doi: 10.1007/s11695-020-04577-2

10. Welbourn R, Hollyman M, Kinsman R, Dixon J, Liem R, Ottosson J, Ramos A, Våge V, Al-Sabah S, Brown W, Cohen R, Walton P, Himpens J. Bariatric Surgery Worldwide: Baseline Demographic Description and One-Year Outcomes from the Fourth IFSO Global Registry Report 2018. OBES SURG 2019 Mar 1;29(3):782–795. doi: 10.1007/s11695-018-3593-1

11. Nudotor RD, Canner JK, Haut ER, Prokopowicz GP, Steele KE. Comparing remission and recurrence of hypertension after bariatric surgery: vertical sleeve gastrectomy versus Roux-en-Y gastric bypass. Surg Obes Relat Dis 2021 Feb;17(2):308–318. PMID:33189600

12. Correia JC, Waqas A, Assal J-P, Davies MJ, Somers F, Golay A, Pataky Z. Effectiveness of therapeutic patient education interventions for chronic diseases: A systematic review and meta-analyses of randomized controlled trials. Front Med (Lausanne) 2023 Jan 25;9:996528. PMID:36760883

13. Kim HJ, Madan A, Fenton-Lee D. Does Patient Compliance with Follow-up Influence Weight Loss After Gastric Bypass Surgery? A Systematic Review and Meta-Analysis. OBES SURG 2014 Apr 1;24(4):647–651. doi: 10.1007/s11695-014-1178-1

14. Brown WA, Burton PR, Shaw K, Smith B, Maffescioni S, Comitti B, Cowley MA, Laurie C, Way A, Nottle P. A Pre-Hospital Patient Education Program Improves Outcomes of Bariatric Surgery. OBES SURG 2016 Sep 1;26(9):2074–2081. doi: 10.1007/s11695-016-2075-6

15. Groller KD. Systematic review of patient education practices in weight loss surgery. Surgery for Obesity and Related Diseases 2017 Jun 1;13(6):1072–1085. doi: 10.1016/j.soard.2017.01.008

16. Tom K, Phang PT. Effectiveness of the video medium to supplement preoperative patient education: A systematic review of the literature. Patient Education and Counseling 2022 Jul 1;105(7):1878–1887. doi: 10.1016/j.pec.2022.01.013

17. Gualtieri LN. The doctor as the second opinion and the internet as the first. CHI '09 Extended Abstracts on Human Factors in Computing Systems New York, NY, USA: Association for Computing Machinery; 2009. p. 2489–2498. doi: 10.1145/1520340.1520352

18. Stevenson FA, Kerr C, Murray E, Nazareth I. Information from the Internet and the doctor-patient relationship: the patient perspective – a qualitative study. BMC Fam Pract 2007 Aug 16;8(1):47. doi: 10.1186/1471-2296-8-47

19. Cai HC, King LE, Dwyer JT. Using the Google™ Search Engine for Health Information: Is There a Problem? Case Study: Supplements for Cancer. Current Developments in Nutrition 2021 Feb;5(2):nzab002. doi: 10.1093/cdn/nzab002

20. Strzelecki A. Google Medical Update: Why Is the Search Engine Decreasing Visibility of Health and Medical Information Websites? IJERPH 2020 Feb 12;17(4):1160. doi: 10.3390/ijerph17041160

21. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science American Association for the Advancement of Science; 2018 Mar 9;359(6380):1146–1151. doi: 10.1126/science.aap9559

22. Waszak PM, Kasprzycka-Waszak W, Kubanek A. The spread of medical fake news in social media – The pilot quantitative study. Health Policy and Technology 2018 Jun 1;7(2):115–118. doi: 10.1016/j.hlpt.2018.03.002

23. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ 2023 Mar 14; PMID:36916887

24. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee S-S. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell 2023 Oct 31;6:1237704. PMID:38028668

25. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023 Feb 8;9:e45312. PMID:36753318

26. Kung TH, Cheatham M, ChatGPT, Medenilla A, Sillos C, Leon LD, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv; 2022. p. 2022.12.19.22283643. doi: 10.1101/2022.12.19.22283643

27. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, Berkowitz ST, Finn AP, Jahangir E, Scoville EA, Reese TS, Friedman DL, Bastarache JA, van der Heijden YF, Wright JJ, Ye F, Carter N, Alexander MR, Choe JH, Chastain CA, Zic JA, Horst SN, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan CM, Padmanabhan C, Bailey CE, Schlegel CE, Chambless LB, Gibson MK, Osterman TJ, Wheless LE, Johnson DB. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Network Open 2023 Oct 2;6(10):e2336483. doi: 10.1001/jamanetworkopen.2023.36483

28. expert reaction to study comparing physician and AI chatbot responses to patient questions | Science Media Centre. Available from: https://www.sciencemediacentre.org/expert-reaction-to-study-comparing-physician-and-ai-chatbot-responses-to-patient-questions/ [accessed Nov 12, 2023]

29. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res 2023 Jun 30;25:e47479. PMID:37389908

30. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. PMID:36981544

31. ChatGPT Needs to Go to College. Will OpenAI Pay? - The Washington Post. Available from: https://www.washingtonpost.com/business/2023/06/05/chatgpt-needs-better-training-data-will-openai-and-google-pay-up-for-it/f316828c-035d-11ee-b74a-5bdd335d4fa2_story.html [accessed Apr 25, 2024]

32. Reavis KM, Barrett AM, Kroh MD, editors. The SAGES Manual of Bariatric Surgery. Cham: Springer International Publishing; 2018. doi: 10.1007/978-3-319-71282-6ISBN:978-3-319-71281-9

33. Salminen P, Kow L, Aminian A, Kaplan LM, Nimeri A, Prager G, Behrens E, White KP, Shikora S, IFSO Experts Panel, Dayyeh BKA, Alfaris N, Al Qahtani A, Andersen B, Angrisani L, Bashir A, Batterham RL, Behrens E, Bhandari M, Bond D, Chevallier J-M, Cohen RV, Dicker D, Fox CK, Garneau P, Gawdat K, Haddad A, Himpens J, Inge T, Kurian M, Faria SL, Macedo G, Miras AD, Moize V, Pattou F, Poggi L, Ponce J, Ramos A, Rubino F, Sanchez-Pernaute A, Sarwer D, Sharma AM, Stier C, Thompson C, Vidal J, Petry TBZ. IFSO Consensus on Definitions and Clinical Practice Guidelines for Obesity Management—an International Delphi Study. OBES SURG 2024 Jan;34(1):30–42. doi: 10.1007/s11695-023-06913-8

34. Sharaiha RZ, Shikora S, White KP, Macedo G, Toouli J, Kow L. Summarizing Consensus Guidelines on Obesity Management: A Joint, Multidisciplinary Venture of the International Federation for the Surgery of Obesity & Metabolic Disorders (IFSO) and World Gastroenterology Organisation (WGO). Journal of Clinical Gastroenterology 2023 Nov;57(10):967–976. doi: 10.1097/MCG.0000000000001916

35. ASMBS Public Education Committee. FAQs of Bariatric Surgery. FAQs of Bariatric Surgery. 2020. Available from: https://asmbs.org/patients/faqs-of-bariatric-surgery/ [accessed Nov 2, 2023]

36. IFSO. FAQ on Obesity Surgery. FAQ on Obesity Surgery. 2023. Available from: https://www.ifso.com/faq-obesity-surgery/ [accessed Nov 2, 2023]

37. Choudhury A, Elkefi S, Tounsi A. Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: A cross sectional survey study. Fernandes TP, editor. PLoS ONE 2024 Mar 8;19(3):e0296151. doi: 10.1371/journal.pone.0296151

38. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Srinivasan N, Park J, Burch M, Watson R, Liran O, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes Surg 2023;33(6):1790–1796. PMID:37106269

39. Ye C, Zweck E, Ma Z, Smith J, Katz S. Doctor Versus Artificial Intelligence: Patient and Physician Evaluation of Large Language Model Responses to Rheumatology Patient Questions in a Cross-Sectional Study. Arthritis & Rheumatology 2024 Mar;76(3):479–484. doi: 10.1002/art.42737

40. Copeland-Halperin LR, O'Brien L, Copeland M. Evaluation of Artificial Intelligence-generated Responses to Common Plastic Surgery Questions. Plast Reconstr Surg Glob Open 2023 Aug;11(8):e5226. PMID:37654681

41. How ChatGPT and our language models are developed | OpenAI Help Center. Available from: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-

developed [accessed May 9, 2024]

42. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology 2023 May 1;307(4):e230424. doi: 10.1148/radiol.230424

43. Tabesh MR, Eghtesadi M, Abolhasani M, Maleklou F, Ejtehadi F, Alizadeh Z. Nutrition, Physical Activity, and Prescription of Supplements in Pre- and Post-bariatric Surgery Patients: An Updated Comprehensive Practical Guideline. OBES SURG 2023 Aug;33(8):2557–2572. doi: 10.1007/s11695-023-06703-2

44. Tabesh MR, Maleklou F, Ejtehadi F, Alizadeh Z. Nutrition, Physical Activity, and Prescription of Supplements in Pre- and Post-bariatric Surgery Patients: a Practical Guideline. OBES SURG 2019 Oct;29(10):3385–3400. doi: 10.1007/s11695-019-04112-y

45. Theophilou E, Koyutürk C, Yavari M, Bursic S, Donabauer G, Telari A, Testa A, Boiano R, Hernandez-Leo D, Ruskov M, Taibi D, Gabbiadini A, Ognibene D. Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study. In: Basili R, Lembo D, Limongelli C, Orlandini A, editors. AIxIA 2023 – Advances in Artificial Intelligence Cham: Springer Nature Switzerland; 2023. p. 481–496. doi: 10.1007/978-3-031-47546-7_33ISBN:978-3-031-47545-0

46. Ahmad MA, Yaramis I, Roy TD. Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. arXiv; 2023. Available from: http://arxiv.org/abs/2311.01463 [accessed May 11, 2024]

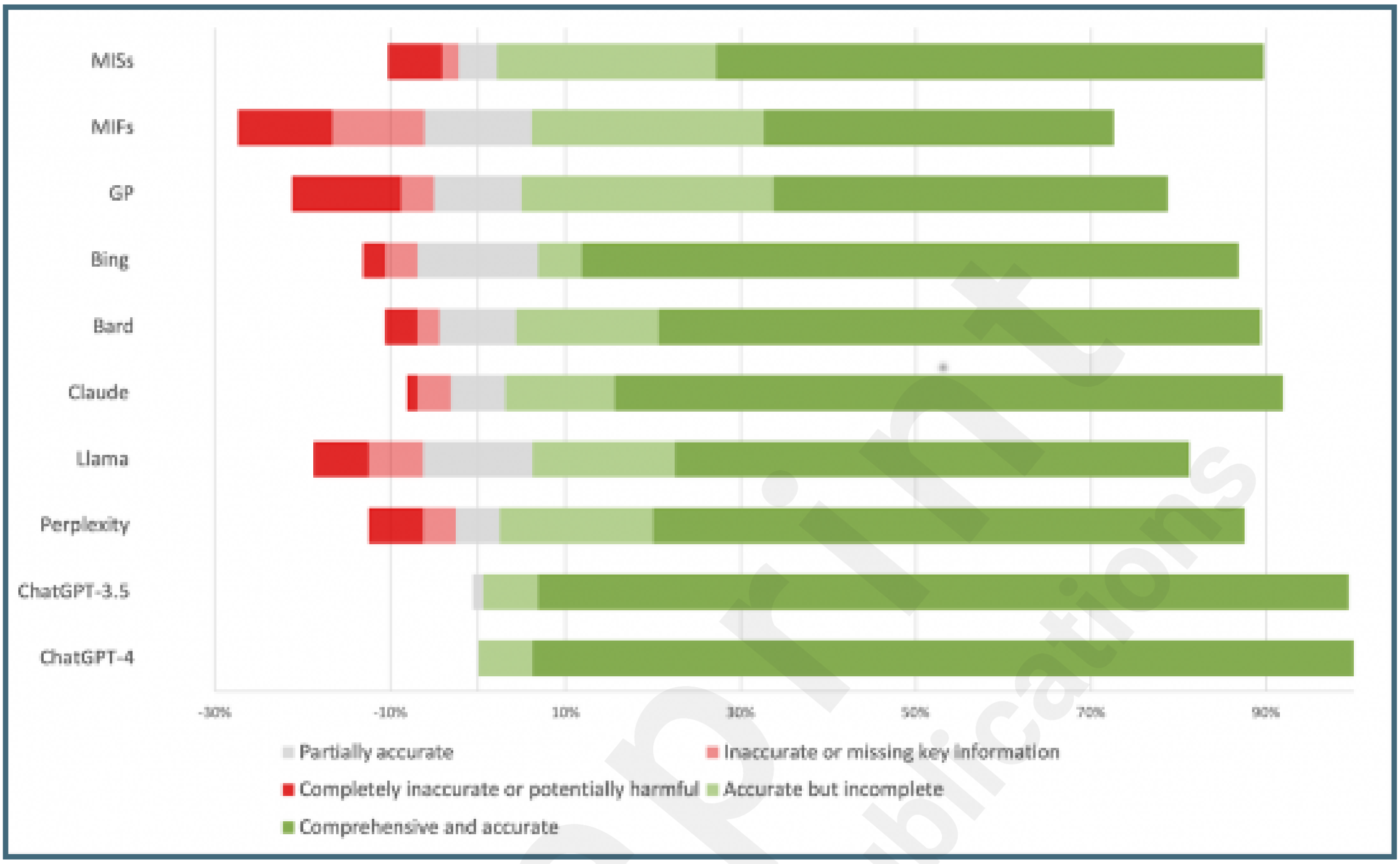
47. Meleo-Erwin Z, Basch C, Fera J, Ethan D, Garcia P. Readability of online patient-based information on bariatric surgery. Health Promot Perspect 2019 May 25;9(2):156–160. PMID:31249804

48. Seitz L. Artificial empathy in healthcare chatbots: Does it feel authentic? Computers in Human Behavior: Artificial Humans 2024 Jan;2(1):100067. doi: 10.1016/j.chbah.2024.100067

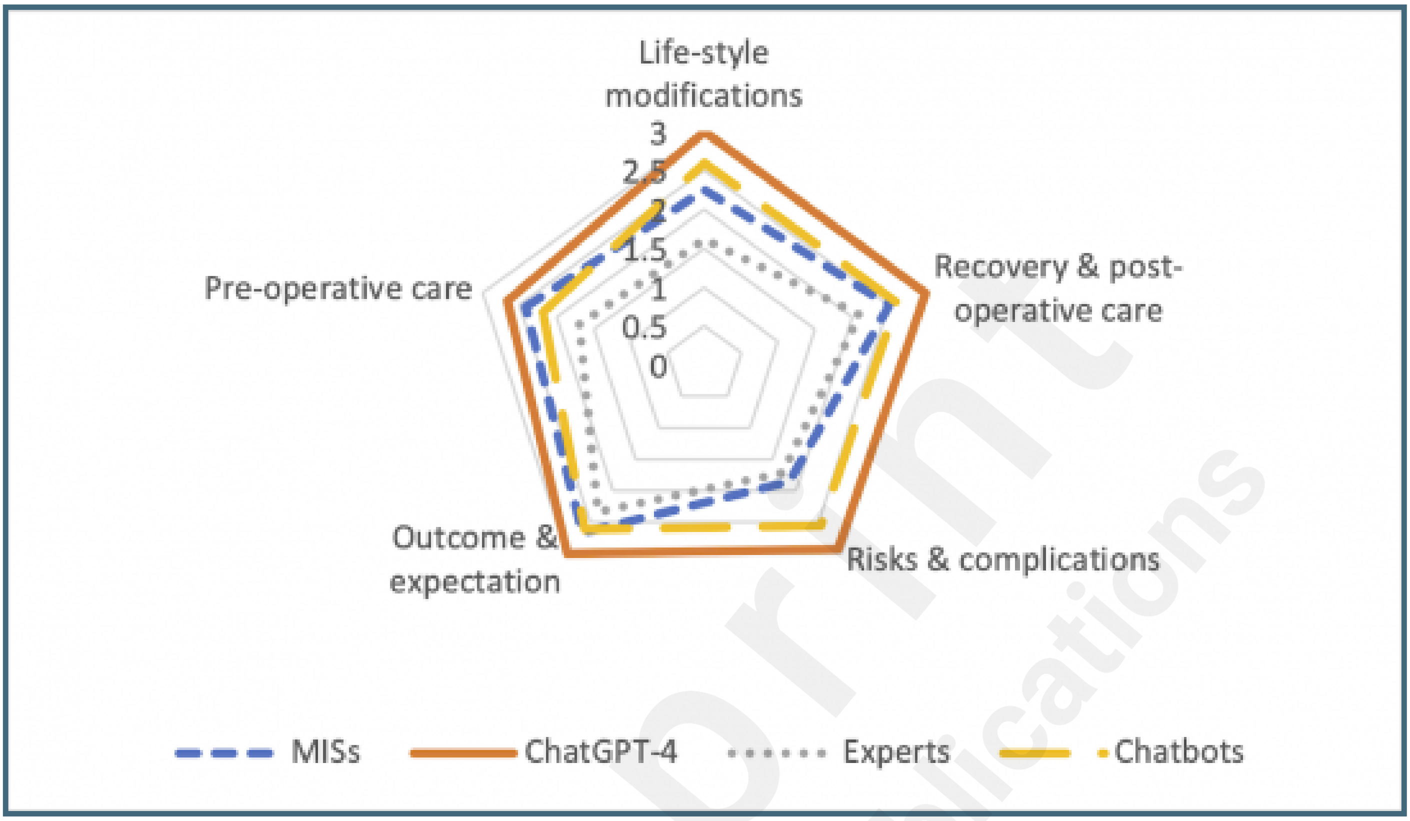
Supplementary Files

Figures

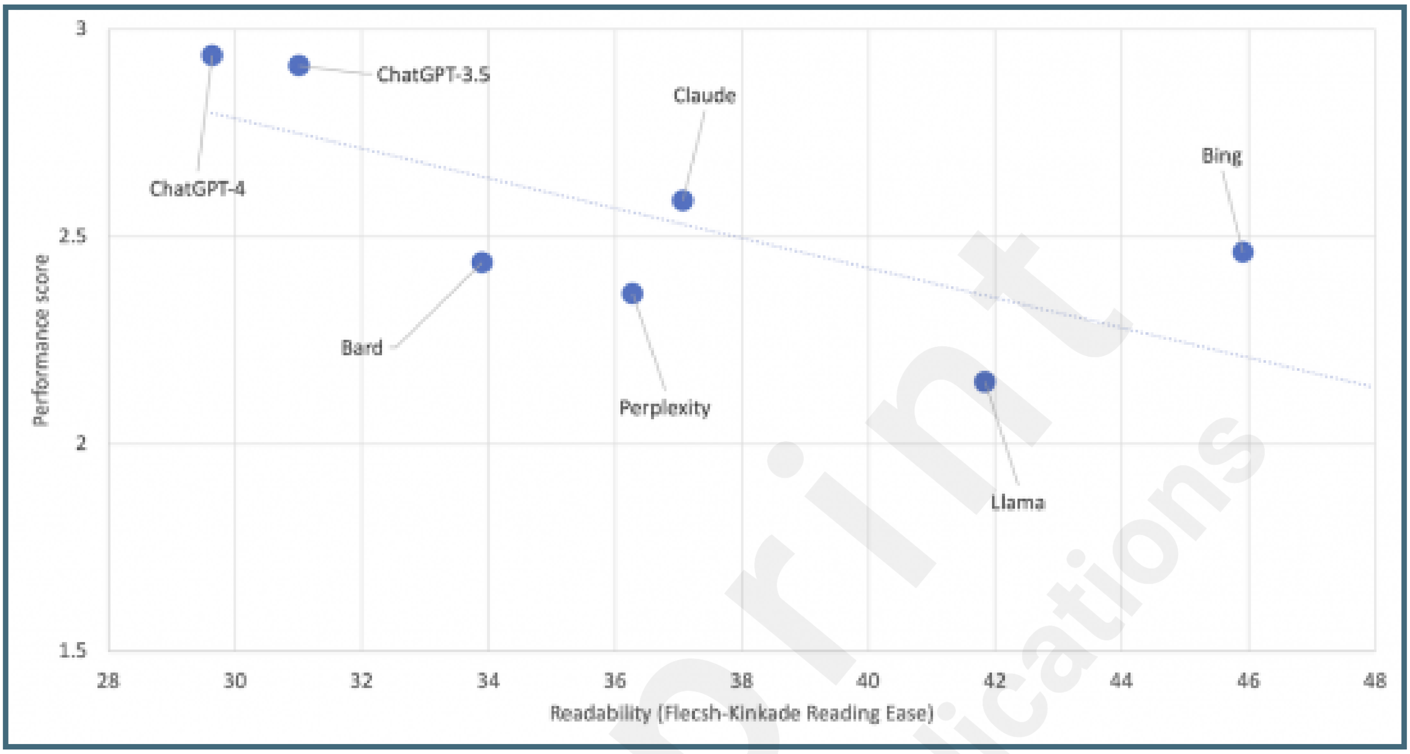
Responses accuracy of each group on all questions, categorized by the degree of accuracy and completeness.



Comparison of performance score between the top-performing chatbot (ChatGPT-4), the top-performing expert group (MISs), all chatbots (Chatbots), and all expert groups (Experts) across various categories.



The inverse relationship between performance score and readability (measured by the Flesch-Kinkade Reading Ease) for various chatbots, with a downward trend indicating that higher accuracy correlates with lower readability.



Multimedia Appendixes

Questions And Answers.
URL: http://asset.jmir.pub/assets/75f8f6fb5112fc895610e9bafcf345c3.docx

Performance Analysis Across Domains.
URL: http://asset.jmir.pub/assets/275f5fc4a8e6dffa0e6aed8e0faa1a4f.docx