

Exploring Chatbot Applications in Pancreatic Disease Treatment: Potential and Pitfalls

Alberto Balduzzi^{1,2}, Matteo De Pastena², Susanna Tondato², Federico Gronchi², Tommaso Dall'Olio², Giuseppe Malleo², Antonio Pea², Salvatore Paiella², Roberto Salvia²

¹General Surgery Unit, Pederzoli Hospital, Peschiera del Garda, Verona, Italy.

²Department of General and Pancreatic Surgery, The Pancreas Institute, University of Verona Hospital Trust, Verona, Italy.

Correspondence to: Alberto Balduzz, General Surgery Unit, Pederzoli Hospital, Peschiera del Garda, Verona, Italy General and Pancreatic Surgery Department, Pancreas Institute, University and Hospital Trust of Verona, Verona, Italy. E-mail: alberto.balduzzi.bo@gmail.com

Received: 11 June 2025 | Approved: 13 June 2025 | Online: 13 June 2025

Abstract

Background: The use of chatbots to respond across various domains is becoming more integrated into daily life, potentially replacing traditional search engines. The study aimed to investigate the performance of different Large Language Models (LLMs) in providing recommendations regarding pancreatic cancer (PC) to surgeons.

Methods: Standardized prompts were engineered to query four freely accessible LLMs (ChatGPT-4, Personal Intelligence by Inflection AI, Anthropic Claude 3 Haiku Version 3.5, Perplexity AI) on October 9th, 2024. Fourteen questions included the incidence, diagnosis, and treatment for radiologically resectable, borderline resectable, locally advanced, and metastatic PC. Three different investigators queried the LLMS



© The Author(s) 2025. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or

format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

DOI: 10.20517/scierxiv202506.0366.v2

<https://www.scierxiv.com/>

simultaneously. The reliability and accuracy of the responses were evaluated using a 4-point Likert scale and then compared to the international guidelines. Descriptive statistics were used to report outcomes as counts and percentages.

Results: Overall, 72% of the responses were deemed correct (scored 3 or 4). Claude provided the most accurate responses (32%), followed by ChatGPT (28%). ChatGPT-4 and Anthropic Claude 3 Haiku Version 3.5 achieved the overall highest score rate (4-point) at 50% and 52%, respectively. Regarding the quality and accuracy of the responses, ChatGPT cited guidelines most frequently (29%). However, only 19% of all evaluated responses included guideline citations.

Conclusion: The LLMs are still not suitable for safe, standalone use in the medical field, their rapid learning capabilities suggest they may become indispensable tools for medical professionals in the future.

Keywords: LLM, Artificial intelligence, pancreas, PDAC, guidelines.

INTRODUCTION

Over recent decades, advancements in artificial intelligence (AI) have revolutionized various domains, including the development of conversational agents known as chatbots. These systems utilize AI and natural language processing (NLP) to interpret user queries and generate automated responses, mimicking human dialogue.^[1] Chatbots are primarily built on two components: a general-purpose AI framework and a user-friendly chat interface.^[2] By leveraging large language models (LLMs), chatbots demonstrate conversational reasoning, producing contextually appropriate responses.^[3]

The public release of ChatGPT (Chat Generative Pretrained Transformer) in November 2022 marked a significant milestone, accelerating the proliferation of similar AI-driven tools across industries.^[4] However, most LLMs, including ChatGPT, are trained predominantly on non-medical datasets.^[5]

This limitation has sparked widespread discussions regarding the use of such tools in sensitive fields like medicine, raising ethical concerns about bias reinforcement, potential plagiarism, misinformation, and the reliability of AI-generated content.^[6,7]

Despite these challenges, researchers have begun investigating the quality and reliability of chatbot-generated medical information.^[8] Preliminary studies have compared the accuracy and readability of responses from various chatbots across general medical topics and specific specialties.^[9]

Recent efforts include evaluations of chatbots' performance on topics such as cancer care, ophthalmology, and other specialized domains^[10-14]

Despite extensive research on AI-generated information in medical fields, there is a lack of data evaluating the performance of Large Language Models (LLMs) specifically addressing pancreatic cancer treatment. This study aims to bridge that gap by evaluating the accuracy and reliability of chatbot-generated information on pancreatic cancer comparing multiple LLMs under standardized clinical conditions. Specifically, we analyzed chatbot responses to queries posed by specialized clinicians on topics including risk factors, diagnosis, and management of resectable, borderline resectable, locally advanced, and metastatic pancreatic cancer. Our primary objectives were to assess the accuracy of AI-provided medical information, evaluate its concordance with international guidelines, and determine the comprehensiveness of the content delivered.

MATERIALS AND METHODS

LLM Selection and Objectives

The study team selected four free online chatbots (ChatGPT-4, Personal Intelligence by Inflection AI, Anthropic Claude 3 Haiku Version 3.5, Perplexity AI) based on their widespread usage among the general population and the experience of the surgeons involved in the study. The primary objective of the study was to determine the accuracy of chatbots in providing clinicians with recommendations on pancreatic cancer diagnosis and treatments. This was achieved by querying the 4 chatbots and comparing the responses to guidelines.

Ethical approval was not required for this study as it involved non-sensitive survey responses, in line with the Verona Pancreas Institute ethical policy.

Questionnaire Selection & Testing

A panel of pancreatic surgeons meticulously developed the questionnaire. The questionnaire consisted of 14 questions evenly distributed across seven topics: epidemiology (n=2), diagnosis (n=2), treatment options for radiologically resectable disease (n=2), borderline resectable disease (n=2), locally advanced disease (n=2), metastatic pancreatic cancer (n=2), and best supportive care (n=2) (Figure 1). The process began with a thorough review of the most widely used clinical guidelines for pancreatic cancer:

- Clinical Practice Guidelines in Oncology (NCCN guidelines®) for Pancreatic Adenocarcinoma^[15]
- ESMO Clinical Practice Guideline for diagnosis, treatment, and follow-up for pancreatic cancer^[16]
- Pancreatic cancer in adults: diagnosis and management from National Institute for Health and Care Excellence^[17]
- Clinical Practice Guidelines for Pancreatic Cancer 2022 from the Japan Pancreas Society: a synopsis^[18]

Each question was derived directly from the clinical guidelines. The phrasing was constructed to reflect conventional medical terminology at a surgeon level. Before querying, all chatbots were prompted with the same introductory statement: "I am a pancreatic surgeon, and I'd like to know more about pancreatic cancer,". This was designed to stimulate the perspective of a specialist and trigger the chatbot to provide a precise recommendation rather than generic or broad information related to pancreatic cancer. Follow-up prompts were not used.

Figure 1: Questionnaire performed by chatbots

- 1) What are the known risk factors for sporadic Pancreatic cancer?
- 2) Is new onset of diabetes mellitus considered as risk factor for pancreatic cancer?
- 3) What is the correct diagnostic pathway for pancreatic cancer diagnosis?
- 4) Is positron emission tomography (PET) recommended as a diagnostic method in patients with suspected pancreatic cancer?
- 5) When is surgical approach indicated as first line therapy for pancreatic cancer?
- 6) Is prophylactic extended lymph node and nerve plexus dissection recommended in patients with resectable pancreatic cancer undergoing surgery?
- 7) What is the definition of border-line resectable pancreatic cancer?
- 8) Is combined arterial resection recommended in patients with BR pancreatic cancer?
- 9) What is the first-line treatment recommended for patients with locally advanced, unresectable pancreatic cancer?
- 10) What is the chemoradiotherapy regimen recommended for patients with locally advanced, unresectable pancreatic cancer?
- 11) What is the first-line chemotherapy recommended for pancreatic cancer patients with distant metastases?
- 12) Is surgical resection recommended for pancreatic cancer patients with postoperative metastases/recurrences?
- 13) Which of the two types of stents—plastic stents or metallic stents—recommended in resectable or borderline resectable pancreatic cancer patients with obstructive jaundice?
- 14) Is anticoagulant therapy for preventing venous thromboembolism recommended in patients with unresectable pancreatic cancer undergoing chemotherapy?

Query strategy

ChatGPT-4, Personal Intelligence by Inflection AI, Anthropic Claude 3 Haiku Version 3.5 and Perplexity AI were queried on October 9th, 2024, by three study authors.

These co-investigators performed the queries in Verona, Italy. Each AI chatbot was queried through its respective user interfaces. The questions were posed simultaneously under identical conditions to ensure comparability of the data collected. Outputs along with their citations were extracted without modification. We did not make post-hoc changes to study prompts. The study team queried the chatbots in English.

Performance Evaluation

The study team applied a standardized data collection form to collate response data (Supplementary data). We defined accurate performance by the LLM as alignment with clinical practice guideline recommendations. When a response contained a guideline citation, its accuracy was verified against the most recent version of the respective guideline. When there was no clear reference to guidelines, the response was assessed based on the normal clinical practice of the reference Pancreatic Institute (Verona, Italy). The quality and accuracy of each response were evaluated using a composite measure of factual correctness, guideline adherence, and comprehensiveness, as determined by a

four-point Likert scale, scoring as follows: 4) correct and comprehensive, 3) correct but not comprehensive, 2) some correct with disinformation, and 1) completely incorrect. Two general surgery residents independently scored the responses. After the initial evaluation of the responses, the two team members jointly re-evaluated each response and provided an overall score. The variability in chatbot responses to identical questions was assessed for consistency. Responses that received a score of 3 or 4 were considered correct. Responses were considered misaligned with society guidelines if they conflicted with recommendations or failed to provide meaningful answers.

The Flesch-Kincaid reading grade level score was used to assess readability. Notably, the National Institutes of Health recommends an eighth-grade reading level for materials intended for the general population.^[19]

Statistical Analysis

Descriptive statistics were performed. One member of the study team performed data analysis using descriptive statistics to report dichotomous outcomes including counts and percentages. The statistical analysis was performed using the statistical software GraphPad Prism 10 (San Diego, USA). ANOVA identified differences in rater evaluation and readability, with significance determined at $p < 0.05$. Data are presented as mean and standard deviation. The chatbot Assessment Reporting Tool is currently under development.^[20]

RESULTS

A total of 168 questions were posed to the chatbots. All chatbots answered every question. Table 1 summarizes the distribution of correct answers, with 71.5% of responses scored as either 3 or 4. The most accurate chatbot was Anthropic Claude 3 Haiku Version 3.5, which provided 38/120 (31.6%) correct answers, followed by ChatGPT-4 with 33/120 (27.5%). Of the 54 answers receiving the highest score of 4, Anthropic Claude 3 Haiku Version 3.5 provided the majority (22/54, 40.7%), followed closely by ChatGPT-4 (21/54, 38.9%).

Table 2 highlights the best of three answers to each question. ChatGPT-4 achieved the highest number of correct and comprehensive responses (11/14, 78.6%), followed by Anthropic Claude 3 Haiku Version 3.5 (10/14, 71.4%) and Perplexity AI (6/14, 42.9%).

However, ChatGPT-4 exhibited the greatest variability in responses, with 10/14 (71.4%) answers rated inconsistently by evaluators. Surgical treatment options and systemic therapies received the lowest average scores, with most responses scoring below 3. ChatGPT-4 produced the longest responses on average, while Anthropic Claude 3 Haiku Version 3.5 offered the highest proportion of precise recommendations.

ChatGPT-4 provided the highest number of guideline-cited answers (12/42, 28.6%), followed by Anthropic Claude 3 Haiku Version 3.5 (10/42, 23.8%) and Personal Intelligence by Inflection AI (7/42, 16.7%). Overall, only 18.5% of responses referenced guidelines, indicating room for improvement in evidence-based citations.

All LLMs delivered responses exceeding the Flesch-Kincaid reading level of 6 recommended by the National Institute of Health (NIH) and the average American reading level of 8.^[21] The reading levels of the LLMs' answers ranged from 9,4 (Claude, most accessible) to 20 (PiAI, least accessible). On average, all four LLMs demonstrated an average reading score exceeding 12, indicative of advanced (college-level) reading proficiency. The most accessible LLM was Perplexity (Flesch-Kinkaid grade level 12,3), while the least accessible was PIAI (15,6, $p<0,007$). (Table 3)

Table 1: Distribution of correct answers in the four chatbots

	chatGPT	Claude	PiAI	Perplexity	All 4
Likert score 4	21	22	0	11	54
Total Answers	42	42	42	42	168
%	50	52,3809524	0	26,1904762	32%
Likert score 3+4	33	38	23	26	120
Total answers	42	42	42	42	168
%	78,5714286	90,4761905	54,7619048	61,9047619	71,50

Table 2: Distribution of correct answers when considering only the best of three answers

	chatGPT	Claude	piAI	Perplexity
likert score 4	11	10	0	6
tot	14	14	14	14

%	78,57143	71,42857	0	42,85714
---	----------	----------	---	----------

Table 3: Measure of accuracy and readability of each LLM. Flesch-Kincaid scores represent readability. “References produced by search” quantifies the number of references cited in the average output for each model

	ChatGPT (SD)	Claude (SD)	PiAI (SD)	Perplexity (SD)	
Likert score	3,3 (0,76)	3,45 (0,56)	2,40 (0,43)	2,78 (0,89)	P<0,0007
Flesch-Kincaid readability score	13,94 (1,36)	12,94 (2,47)	15,61 (2,87)	12,31 (1,20)	P<0,0007
References Produced by Search	12	10	7	2	

DISCUSSION

The rapid adoption of chatbots worldwide has positioned them as potentially transformative tools across various domains, including medicine. Their user-friendly interfaces and ability to synthesize vast amounts of knowledge have generated significant interest in their application for research and clinical decision-making. However, as their role in healthcare expands, concerns about their accuracy, reliability, and safety become increasingly pertinent. This study aimed to evaluate the safety and reliability of chatbot-generated responses in addressing pancreatic cancer, a highly specialized and critical medical field.

Consistent with prior research, our findings indicate that while chatbots frequently provide accurate and reliable responses, a significant proportion of their outputs are

coarse, incomplete, or inaccurate. These limitations render chatbots unsuitable for independent medical consultation without expert oversight. For example, our results align with studies such as Hermann et al., which demonstrated ChatGPT's ability to accurately discuss cervical cancer prevention but highlighted shortcomings in areas requiring nuanced understanding, such as diagnosis and treatment.^[22] Similar trends have been observed in other medical specialties, including head and neck surgery, bariatric surgery, and oncology, further underscoring the variability in chatbot performance across different domains.^[23-25]

A notable observation in this study was the variability in chatbot responses, even when identical queries were posed. This inconsistency was particularly evident in ChatGPT, likely due to differences in the specific versions used. Although all researchers employed the same underlying model (LLM Model 4), disparities were observed between the paid version, the free 4.0 version, and the free 4.0 Mini version. The paid version consistently outperformed its free counterparts, providing superior responses in 10 out of 14 cases. However, overlapping response quality occurred in 8 of these cases, demonstrating that free versions occasionally match the performance of paid versions in certain contexts.

The readability scores highlighted that LLMs use advanced language, well above the comprehension level of the general population. In this context, the responses may be easily understood by medical personnel, such as the pancreatic surgeons asking the questions, but could be challenging for patients seeking information about pancreatic cancer. The complexity of the language may, however, be influenced by the initial context, as the chatbot was consistently informed at the start of the questionnaire that the user was a pancreatic surgeon. Therefore, further evaluations are needed to assess the feasibility of using chatbots as an information source for the general population.

Another key finding was the inconsistency in the sources utilized by chatbots. Despite having access to publicly available guidelines, such data was incorporated in only 18.5% of responses. Another critical finding was the inconsistency in the sources utilized by chatbots. Despite having access to publicly available guidelines, such sources were explicitly referenced in only 18.5% of responses. Moreover, chatbots occasionally incorporate information from less authoritative platforms, such as patient

blogs or general online content, which lack the rigor of peer-reviewed scientific literature. This limitation underscores a key challenge for current-generation large language models (LLMs): their inability to transparently assess or prioritize the reliability of their training data. Such shortcomings raise concerns about their suitability for specialized medical applications, where adherence to evidence-based guidelines and credible sources is paramount.^[26]

Previous studies have documented inaccuracies and fabrications in chatbot-generated responses, a phenomenon referred to as "hallucination"^[27] This issue arises when chatbots generate plausible-sounding but factually incorrect responses, making errors less apparent to users. Our study found this issue to be more pronounced in responses to highly specialized queries, where chatbots deviated from established guidelines and relied on less authoritative sources. These compromises in source selection and accuracy can pose significant risks, particularly in high-stakes fields like oncology care. In clinical settings, such errors could lead to misinformation with serious consequences. Therefore, we suggest future work should focus on domain-specific fine-tuning and integration with up-to-date clinical databases to mitigate hallucination risk.

Our findings emphasize the need for caution in integrating chatbots into clinical practice. While the potential of these tools to revolutionize healthcare is undeniable, substantial advancements are required to ensure their safety and reliability. Training LLMs with domain-specific datasets curated by medical experts could significantly enhance their performance and address existing limitations. Although current LLMs are not yet ready for autonomous clinical decision-making, they could be integrated as decision-support tools. With expert oversight, these systems might assist in synthesizing clinical data, guiding diagnostic pathways, and recommending treatment strategies according to the most updated guidelines and resources. Such improvements would enable chatbots to serve as valuable tools for disseminating medical knowledge, supporting healthcare professionals, and enhancing patient education.^[20] Soon, LLMs might play a crucial role in enhancing, rather than replacing, the clinician's expertise.^[28]

This study is not without limitations. First, inter-reviewer variability may have influenced the evaluation of chatbot responses, as subjective judgment was inherent to the assessment process. Second, in prompting, different user identities (e.g., general

physicians or patients) might elicit different responses that would benefit from further research. Third, the main use of free chatbot versions, which are known to have reduced data processing capabilities compared to paid versions, may have impacted our findings. Future research should include a broader range of chatbot models and focus on developing objective evaluation criteria to provide more robust and reproducible insights.

While the rapid evolution of artificial intelligence presents exciting possibilities for healthcare, our findings underscore the need for continued scrutiny and refinement of chatbot technologies. Ensuring their safety, reliability, and adherence to evidence-based practices is essential before their widespread adoption in clinical settings.

CONCLUSION

Chatbots are emerging as increasingly efficient tools for answering queries rapidly, offering a convenient alternative to the time-consuming process of searching through traditional literature. However, the extensive and often non-curated datasets used by LLMs pose significant risks, as unchecked outputs can lead to misinformation and errors.

The rapid advancements in LLM technology are promising, but their successful integration into clinical practice will require significant improvements in reliability, transparency, and adherence to evidence-based guidelines. If these challenges are addressed, LLMs could transform from supplementary tools into essential assets in the surgical field.

DECLARATION

Acknowledgment

The research leading to these results has received funding from the European Union—NextGenerationEU through the Italian Ministry of University and Research under PNRR—M4C2-I1.3 Project PE 00000019 "HEAL ITALIA" to the authors CUP B33C22001030006. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them

Author Contributions

Conceptualization & Methodology: A. Balduzzi, M. De Pastena; Statistical Analysis: S. Tondato; Writing – Original Draft: S. Tondato, F. Gronchi; Writing – Review & Editing: A. Balduzzi, M. De Pastena, S. Tondato.

Availability of Data and Materials

All data and materials pertinent to this study are available from the corresponding author upon reasonable request.

Financial Support and Sponsorship

No financial support or external sponsorship was received for the conduct of this research or the preparation of this manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Ethical Approval and Consent to Participate

Ethical approval was not required for this study as it involved non-sensitive survey responses, in line with the Verona Pancreas Institute ethical policy.

Consent for Publication

All authors agree and consent to the publication of this manuscript.

Copyright

© The Author(s) 2025.

REFERENCE

1. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* 2023;15:192.[DOI:10.3390/fi15060192]
2. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med* 2023;388:1233-9.[DOI:10.1056/nejmsr2214184]
3. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology,

- and psoriasis: A comparative study. *Int J Med Inform* 2024;190:105562.[PMID:39059084 DOI:10.1016/j.ijmedinf.2024.105562]
4. Nirala KK, Singh NK, Purani VS. A survey on providing customer and public administration based services using AI: chatbot. *Multimed Tools Appl* 2022;81:22215-46.[PMID:35002470 DOI:10.1007/s11042-021-11458-y PMCID:PMC8721490]
 5. Gallifant J, Fiske A, Levites Strekalova YA, et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health* 2024;3:e0000417.[PMID:38236824 DOI:10.1371/journal.pdig.0000417 PMCID:PMC10795998]
 6. Liebrezn M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health* 2023;5:e105-6.[PMID:36754725 DOI:10.1016/s2589-7500(23)00019-5]
 7. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health* 2023;5:e102.[PMID:36754723 DOI:10.1016/s2589-7500(23)00023-7]
 8. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *N Engl J Med* 2023;388:1201-8.[DOI:10.1056/nejmra2302038]
 9. Wagner MW, Ertl-Wagner BB. Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information. *Can Assoc Radiol J* 2024;75:69-73.[PMID:37078489 DOI:10.1177/08465371231171125]
 10. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res* 2023;25:e47479.[PMID:37389908 DOI:10.2196/47479 PMCID:PMC10365578]
 11. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq* 2023. Available from: <https://www.researchsquare.com/article/rs-2566942/v1> [Last accessed on 10 Jun 2025]
 12. Lee JW, Yoo IS, Kim JH, et al. Development of AI-generated medical responses using the ChatGPT for cancer patients. *Comput Methods Programs Biomed* 2024;254:108302.[DOI:10.1016/j.cmpb.2024.108302]
 13. Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery* 2023;174:1273-5.[DOI:10.1016/j.surg.2023.06.005]

14. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol* 2023;141:589-97.[DOI:DOI:10.1001/jamaophthalmol.2023.1144]
15. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Pancreatic Adenocarcinoma version 3.2024. Available from: <https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1455> [Last accessed on 10 Jun 2025]
16. Conroy T, Pfeiffer P, Vilgrain V, et al; ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Pancreatic cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol* 2023;34:987-1002.[DOI:10.1016/j.annonc.2023.08.009]
17. Pancreatic cancer in adults: diagnosis and management. Final. London: National Institute for Health and Care Excellence; 2018. Available from: <https://www.nice.org.uk/guidance/ng85> [Last accessed on 10 Jun 2025]
18. Okusaka T, Nakamura M, Yoshida M, et al; Committee for Revision of Clinical Guidelines for Pancreatic Cancer of the Japan Pancreas Society. Clinical Practice Guidelines for Pancreatic Cancer 2022 from the Japan Pancreas Society: a synopsis. *Int J Clin Oncol* 2023;28:493-511.[DOI:10.1007/s10147-023-02317-x]
19. Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *J Patient Exp* 2021;8:2374373521998847.[PMID:34179407 DOI:10.1177/2374373521998847 PMID:PMCID:PMC8205335]
20. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 2023;29:2988.[PMID:37957381 DOI:10.1038/s41591-023-02656-2]
21. Cotugna N, Vickery CE, Carpenter-Haeefe KM. Evaluation of literacy level of patient education pages in health-related journals. *J Community Health* 2005;30:213-9.[PMID:15847246 DOI:10.1007/s10900-004-1959-x]
22. Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. *Gynecol Oncol* 2023;179:164-8.[PMID:37988948 DOI:10.1016/j.ygyno.2023.11.008]
23. Lee Y, Tessier L, Brar K, et al; ASMBS Artificial Intelligence and Digital Surgery Taskforce. Performance of artificial intelligence in bariatric surgery: comparative

- analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions. *Surg Obes Relat Dis* 2024;20:609-13.[DOI:10.1016/j.soard.2024.04.014]
24. Carl N, Schramm F, Haggenmüller S, et al. Large language model use in clinical oncology. *NPJ Precis Oncol* 2024;8:240.[PMID:39443582 DOI:10.1038/s41698-024-00733-4 PMCID:PMC11499929]
25. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol* 2023;13:1256459.[PMID:38107064 DOI:10.3389/fonc.2023.1256459 PMCID:PMC10722294]
26. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw Open* 2023;6:e2336483.[PMID:37782499 DOI:10.1001/jamanetworkopen.2023.36483 PMCID:PMC10546234]
27. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* 2023;15:e39238. Available from: <https://www.cureus.com/articles/158289-high-rates-of-fabricated-and-inaccurate-references-in-chatgpt-generated-medical-content> [Last accessed on 02 Apr 2025]
28. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025;25:117. [DOI: 10.1186/s12911-025-02954-4]