

Assessment of ChatGPT and Google in delivering gynecologic cancer education content: A comparative evaluation of accuracy, completeness, and reference quality

Alexandra H. Smick, Martins Ayoola, Rajeshree Rajpara, Isabel Lazo, Dana M. Chase

Submitted to: JMIR Cancer
on: April 23, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5
Supplementary Files..... 17
 Multimedia Appendixes 18
 Multimedia Appendix 0..... 18



Assessment of ChatGPT and Google in delivering gynecologic cancer education content: A comparative evaluation of accuracy, completeness, and reference quality

Alexandra H. Smick¹ MD; Martins Ayoola¹ MD; Rajeshree Rajpara²; Isabel Lazo³ MD; Dana M. Chase¹ MD

¹ Gynecologic Oncology University of California, Los Angeles Los Angeles US

² Undergraduate College University of California, Los Angeles Los Angeles US

³ Gynecologic Oncology Olive View-UCLA Medical Center Sylmar US

Corresponding Author:

Alexandra H. Smick MD

Gynecologic Oncology
University of California, Los Angeles
100 Medical Plaza Suite 383
Los Angeles
US

Abstract

Background: Patients with newly diagnosed gynecologic (GYN) cancers often seek information online, but the quality of available resources may be inconsistent. While ChatGPT may offer an alternative to traditional internet search engines, its performance remains largely understudied in GYN oncology.

Objective: To compare the completeness, accuracy, and reference quality of patient education responses generated by ChatGPT versus Google in clinical scenarios involving a new diagnosis of GYN cancer.

Methods: Clinical scenarios representing early- and advanced-stage endometrial, ovarian, and cervical cancers were developed by GYN oncologists using publicly available patient education materials. Each included four standardized questions on etiology, prognosis, treatment, and treatment efficacy. ChatGPT-4 and Google were queried for each question, with new sessions for ChatGPT and private browsing for Google to minimize bias. Responses were independently rated by four blinded GYN oncology experts. Accuracy was scored on a 6-point Likert scale; completeness and reference quality were scored on 3-point scales. Reference quality was categorized as low (commercial), medium (institutional/government), or high (peer-reviewed). Reviewer comments were also collected. Descriptive statistics and Wilcoxon signed-rank tests were used for analysis.

Results: Across six clinical scenarios and 21 standardized questions (n=84 total responses), ChatGPT outperformed Google on all evaluated domains. The median accuracy score was 6 (IQR 5-6) for ChatGPT and 5 (IQR 4-6) for Google (P=.04). The median completeness score was 3 (IQR 3-3) for ChatGPT and 2 (IQR 1-3) for Google (P=.009). Reference quality was also higher for ChatGPT with a median score of 3 (IQR 3-3) compared to 2 (IQR 2-2) for Google (P=.009). Reviewer comments noted that ChatGPT provided more accurate, comprehensive, and personalized responses, while Google returned less detailed content from general consumer health websites rather than peer-reviewed sources.

Conclusions: ChatGPT may serve as a reliable and high-quality resource for patient education in GYN oncology. Compared to Google, it delivered more accurate, complete, and better-referenced content. Further research is needed to assess the readability and accessibility of ChatGPT-generated content across diverse patient populations.

(JMIR Preprints 23/04/2025:76471)

DOI: <https://doi.org/10.2196/preprints.76471>

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 visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>



Original Manuscript

ORIGINAL PAPER

Title: Assessment of ChatGPT and Google in delivering gynecologic cancer education content: A comparative evaluation of accuracy, completeness, and reference quality

Authors: Alexandra H Smick, Martins Ayoola, Rajeshree Rajpara, Isabel Lazo, Dana M Chase

Affiliations:

Department of Gynecologic Oncology, University of California Los Angeles

Department of Gynecologic Oncology, Olive View Medical Center

Undergraduate College, University of California Los Angeles

Corresponding Author:

Alexandra Smick

ahsmick@mednet.ucla.edu

10906 Fairbanks Way

Culver City, CA 90230

The manuscript was presented at the 2025 Winter SGO Meeting as a poster.

Abstract:

Background: Patients with newly diagnosed gynecologic (GYN) cancers often seek information online, but the quality of available resources may be inconsistent. While ChatGPT may offer an alternative to traditional internet search engines, its performance remains largely understudied in

GYN oncology.

Objective: To compare the completeness, accuracy, and reference quality of patient education responses generated by ChatGPT versus Google in clinical scenarios involving a new diagnosis of a GYN cancer.

Methods: Clinical scenarios representing early- and advanced-stage endometrial, ovarian, and cervical cancers were developed by GYN oncologists using publicly available patient education materials. Each included four standardized questions on etiology, prognosis, treatment, and treatment efficacy. ChatGPT-4 and Google were queried for each question, with new sessions for ChatGPT and private browsing for Google to minimize bias. Responses were independently rated by four blinded GYN oncology experts. Accuracy was scored on a 6-point Likert scale; completeness and reference quality were scored on 3-point scales. Reference quality was categorized as low (commercial), medium (institutional/government), or high (peer-reviewed). Reviewer comments were also collected. Descriptive statistics and Wilcoxon signed-rank tests were used for analysis.

Results: Across six clinical scenarios and 21 standardized questions (n=84 total responses), ChatGPT outperformed Google on all evaluated domains. The median accuracy score was 6 (IQR 5-6) for ChatGPT and 5 (IQR 4-6) for Google (P=.04). The median completeness score was 3 (IQR 3-3) for ChatGPT and 2 (IQR 1-3) for Google (P=.009). Reference quality was also higher for ChatGPT with a median score of 3 (IQR 3-3) compared to 2 (IQR 2-2) for Google (P=.009). Reviewer comments noted that ChatGPT provided more accurate, comprehensive, and personalized responses, while Google returned less detailed content from general consumer health websites rather than peer-reviewed sources.

Conclusions: ChatGPT may serve as a reliable and high-quality resource for patient education in GYN oncology. Compared to Google, it delivered more accurate, complete, and better-referenced content. Further research is needed to assess the readability and accessibility of ChatGPT-generated content across diverse patient populations.

Keywords: Artificial intelligence, gynecologic cancer, patient education

Introduction:

Approximately 100,000 individuals are diagnosed with a gynecologic (GYN) malignancy each year in the United States [1]. Following a new cancer diagnosis, patients frequently seek supplemental educational materials to better understand their condition and treatment options. Traditional resources, such as those from professional societies and academic institutions, are often comprehensive but presented in generalized formats, requiring patients to interpret the content in the context of their own diagnosis and clinical stage [2]. This underscores the need for more personalized, accessible approaches to patient education in GYN oncology.

Artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, has

gained significant attention in medicine due to its ability to generate rapid, detailed responses based on vast amounts of textual data [3,4]. Previous research has explored the utility of LLMs for clinical tasks such as staging lung cancer, summarizing histopathology reports, and supporting diagnostic decision-making [5,6]. While early findings are promising, concerns remain regarding the accuracy, reliability, and lack of citation transparency in AI-generated content, reinforcing the need for expert oversight.

As ChatGPT becomes increasingly accessible to the public, patients are likely to turn to it, alongside traditional internet search engines, for information about their diagnosis [2]. Prior studies in oncology have evaluated the quality and readability of ChatGPT's responses to patient questions, highlighting variability in content accuracy and source quality [7,8]. Other investigations have emphasized the potential of AI to improve health literacy by simplifying medical content, yet its role in reliably supporting patient education remains unknown [9]. Within GYN oncology specifically, research on AI-based educational tools is limited. Early studies have examined ChatGPT's ability to provide guideline-concordant recommendations for ovarian cancer diagnosis and management, but no prior work has systematically evaluated its performance in delivering educational content to newly diagnosed patients across multiple GYN cancers [10,11].

This study aimed to assess ChatGPT's effectiveness as a patient education tool in the context of newly diagnosed GYN cancers. We compared ChatGPT responses to those generated by Google search results across standardized clinical scenarios, focusing on three key domains: accuracy, completeness, and reference quality. Our findings contribute to the growing body of literature evaluating AI's role in digital health and its potential integration into patient-centered cancer education.

Methods:

In August 2024, ChatGPT 4.0 and a standardized internet search engine (Google) were evaluated using six clinical scenarios representing common presentations of GYN cancers, including early- and advanced-stage ovarian, endometrial, and cervical cancer. These cases were developed by a panel of GYN oncology providers to reflect frequently encountered clinical presentations. For each cancer type, seven patient-centered questions were designed based on commonly asked topics in clinical practice and information available from professional society websites, yielding a total of 21 unique questions [12] (*Table 1*). The number of questions was selected to capture key domains (etiology, prognosis, treatment, and efficacy) while maintaining feasibility for blinded review, consistent with approaches used in prior studies comparing AI-generated health content [13].

For ChatGPT, a new session was initiated prior to each question to prevent inter-question learning. For Google, searches were performed in private browsing mode with a cleared cache, and the top-ranked webpage from each query was recorded (*see supplementary data*). Responses from both platforms were evaluated for accuracy, completeness, and reference quality by four blinded GYN oncology reviewers. Reviewer comments were also collected to capture qualitative impressions. Accuracy, defined as concordance with NCCN guidelines and current literature, was scored using a 1-6 Likert scale. Completeness, defined as inclusion of major concepts relevant to the clinical question, was scored using a 1-3 Likert scale. Reference quality, defined as low-quality (commercial websites, not peer-reviewed), medium-quality (institutional or government resources, reputable but not peer-reviewed), or high-quality (peer-reviewed publications or randomized controlled trials), was scored using a 1-3 Likert scale (*Table 1*). While no validated assessment tool exists for this specific application, the scoring rubrics were adapted from prior literature evaluating online health information quality [14].

Statistical analysis was conducted using JMP Pro (Version 16.0.0; SAS Institute Inc., Cary, NC). Descriptive statistics were calculated for all outcomes and paired comparisons between platforms were conducted using Wilcoxon signed-rank tests.

Table 1. Clinical questions about diagnosis, treatment options, and survival outcomes for common gynecologic malignancies. Likert scales used for grading completeness, accuracy, and reference quality.

#	Clinical Question
1	What causes endometrial cancer?
2	How long will I live with stage IB endometrial cancer?
3	What is the recommended treatment for stage IB endometrial cancer?
4	How effective is first line treatment for stage IB endometrial cancer?
5	How long will I live with stage IIIC endometrial cancer?
6	What is the recommended treatment for stage IIIC endometrial cancer?
7	How effective is first line treatment for stage IIIC endometrial cancer?
8	What causes cervical cancer?
9	How long will I live with stage IA2 cervical cancer?
10	What is the recommended treatment for stage IA2 cervical cancer?
11	How effective is first line treatment for stage IA2 cervical cancer?
12	How long will I live with stage IIIC1 cervical cancer?
13	What is my recommended treatment for stage IIIC1 cervical cancer?
14	How effective is first line treatment for stage IIIC1 cervical cancer?
15	What causes ovarian cancer?
16	How long will I live with stage IC1 ovarian cancer?
17	What is the recommended treatment for stage IC1 ovarian cancer?
18	How effective is first line treatment for stage IC1 ovarian cancer?
19	How long will I live with stage IVB ovarian cancer?
20	What is the recommended treatment for stage IVB ovarian cancer?

21	How effective is first line treatment for stage IVB ovarian cancer?
----	---

Accuracy	1	2	3	4	5	6
	Completely incorrect	More incorrect than correct	Approximately equal incorrect and correct	More correct than incorrect	Nearly all correct	Correct

Completeness	1	2	3
	incomplete, addresses some aspects of the question, but significant parts are missing or incomplete	adequate, addresses all aspects of the question and provides the minimum amount of information required to be considered complete	comprehensive, addresses all aspects of the question and provides additional information or context beyond what was expected

Reference Quality	1	2	3
	Bottom tier: Any .com (webMD, etc)	Middle tier: Any .gov or .org (ACS, NIH, etc)	Top tier: Peer reviewed or RCT (NCCN, uptodate, etc.)

Results:

Accuracy was graded by four physicians on a Likert scale of 1-6 with *1-completely incorrect* and *6-completely correct*. The median score was 5.00 (IQR 4.00-6.00) for Google vs 6.00 (IQR 5.00-6.00) for ChatGPT ($P=.04$). For Google, the distribution by accuracy score from 1 to 6 was 6%, 2.4%, 8.3%, 16.7%, 8.3%, and 58.3% respectively. For ChatGPT, the distribution from 1 to 6 was 0%, 0%, 2.4%, 10.7%, 17.9%, 69.0% respectively. A score of 5 or 6 (the most accurate) accounted for 66.6% of responses for Google vs 86.9% for ChatGPT (*Table 2*).

Table 2. Comparison of accuracy scores for clinical scenarios between ChatGPT and Google.

Accuracy score	Google n=84	ChatGPT n=84	p-value
1; n (%)	5 (6.0%)	0 (0.0%)	
2; n (%)	2 (2.4%)	0 (0.0%)	
3; n (%)	7 (8.3%)	2 (2.4%)	
4; n (%)	14 (16.7%)	9 (10.7%)	
5; n (%)	7 (8.3%)	15 (17.9%)	
6; n (%)	49 (58.3%)	58 (69.0%)	
Mean $\pm$ SD	4.94 $\pm$ 1.51	5.54 $\pm$ 0.78	P=.04
Median (Q1-Q3)	5.00 (4.00-6.00)	6.00 (5.00-6.00)	

Completeness was then graded by four physicians on a Likert scale of 1-3 with *1-incomplete, addressing only some aspect of the questions* and *3-comprehensive, addressing all aspects of the question*. The median score was 2.00 (IQR 1.00-3.00) for Google vs 3.00 (IQR 3.00-3.00) for ChatGPT (P=.009). For Google, the distribution by completeness score from 1 to 3 was 36.9%, 28.6%, 34.5% respectively. For ChatGPT, the distribution from 1 to 3 was 1.2%, 17.8%, and 81.0% respectively. A score of 2 or 3 (mostly complete) accounted for 63.1% of responses for Google vs 98.8% for ChatGPT (Table 3).

Table 3. Comparison of completeness scores for clinical scenarios between ChatGPT and Google.

Completeness score	Google n=84	ChatGPT n=84	p-value
1; n (%)	31 (36.9%)	1 (1.2%)	
2; n (%)	24 (28.6%)	15 (17.8%)	
3; n (%)	29 (34.5%)	68 (81.0%)	
Mean $\pm$ SD	1.98 $\pm$ 0.85	2.80 $\pm$ 0.43	P=.009
Median (Q1-Q3)	2.00 (1.00-3.00)	3.00 (3.00-3.00)	

Finally, reference quality was graded by four physicians on a Likert scale of 1-3 with 1-*bottom tier (.com)* and 3-*top tier (peer-reviewed)*. The median score 2.00 (IQR 2.00-2.00) for Google vs 3.00 (3.00-3.00) for ChatGPT (P=.009). For Google, the distribution by reference quality score from 1 to 3 was 10.7%, 77.4%, 11.9% respectively. For ChatGPT, the distribution from 1 to 3 was 0.0%, 22.6%, and 77.4% respectively. A score of 2 or 3 (highest quality references) accounted for 89.3% of responses for Google vs 100% for ChatGPT (Table 4).

Table 4. Comparison of reference quality scores for clinical scenarios between ChatGPT and Google.

Reference score	Google n=84	ChatGPT n=84	p-value
1; n (%)	9 (10.7%)	0 (0.0%)	
2; n (%)	65 (77.4%)	19 (22.6%)	
3; n (%)	10 (11.9%)	65 (77.4%)	
Mean $\pm$ SD	2.01 $\pm$ 0.48	2.77 $\pm$ 0.42	P=.009
Median (Q1-Q3)	2.00 (2.00-2.00)	3.00 (3.00-3.00)	

Reviewers then provided comments on their perception of the accuracy, completeness, and reference quality of ChatGPT vs Google. Comments are summarized in Table 5. Overall, reviewers assessed that Google source answers had clear, precise language, but often did not completely answer the question and did not give enough information on prognosis or treatment options. For ChatGPT, reviewers generally commented that answers were very complete and comprehensive especially regarding stage and prognosis but might be more challenging for the patient to understand.

Table 5. Reviewer comments on completeness, accuracy, and reference quality of Google vs ChatGPT.

Source	Reviewer Comments
<u>Google</u>	Technically all the information was correct, but it didn't answer the specific question. They would likely need to answer this question with another source.
	The question is not answered, only treatments are mentioned. It also makes it sound like the treatments may not be that effective.
	The information looks very accurate, but there is a lot of set up with prognosis and you have to scroll down to get your answer.
	Having to weed through the information to get the answer, the information is not as clear and is a little confusing.
	Some information is outdated.
	Very general information, easy for a person with average literacy to understand.
	Answers about survival are given but there is no information about the stage.
	There is no mention of treatment effectiveness or immunotherapy.
<u>ChatGPT</u>	Accurate information is given that is straight to the point and concise.
	Complete and thorough answers were given with high quality resources.
	ChatGPT does a nice job of putting information into context for the patient.
	ChatGPT is comprehensive but doesn't address immunotherapy as an option.
	Specific data of effectiveness given with very thorough data on treatment options.
	There is very good information provided on surgery and chemotherapy (including details about bevacizumab and PARPi) and discusses palliative care.
	Information is concise and complete. It talked about what the stage meant.
	My only concern is the patient's reading level and ability to understand the language in CGPT versus the broad, simple language used by Google.
<u>Overall</u>	I thought I was going to come in biased against ChatGPT. As I looked through the answers, I think ChatGPT did much better answering questions than a random Google search with surprisingly high quality, nuanced answers.
	Google searches with answers from Mayo, COH, ACS were good for basic questions, like what is this cancer, what is the stage; although it gave you all the information on staging and you have to scroll to your area of interest.
	ChatGPT did a better job at individualizing the answer. It also did better at prognosis, effectiveness questions. It seemed to do it with nuance and "compassion." But it was also direct, and I think that can be helpful for physicians when it is difficult to give an accurate prognosis.

Discussion:

Principal Results

This study found that ChatGPT performed better than Google sources when providing patient education material for clinical scenarios with a new diagnosis of GYN cancer. Physician-reviewers rated ChatGPT significantly better in terms of accuracy, reliability, and reference quality. Reviewers were surprised by the quality of the ChatGPT answers and references and saw the potential to better address patient-specific questions regarding diagnosis, prognosis, and treatment options with both precise and comprehensive information. The largest educational deficit identified from the Google sources was too general information that was not specific to the patient's cancer diagnosis, stage, or treatment options and therefore difficult to individually interpret.

Comparison with Prior Work

Few studies have looked specifically at the role of ChatGPT for patient education in GYN oncology. Initial research on artificial intelligence in medicine focused on its diagnostic role and potential for clinical decision-making. For example, in otorhinolaryngology, ChatGPT was compared to UpToDate in terms of usefulness and reliability in diagnosing certain common clinical scenarios. UpToDate was found to be more accurate than ChatGPT 3.5 ($P=.009$) and the authors concluded that improvements needed to be made in the usefulness and reliability of ChatGPT's evidence-based knowledge [14]. Similarly, in various clinical cases of thyroid cancer, ChatGPT was found to be reasonably accurate (~76.7%) when providing information about thyroid cancer, but recommendations about evaluation, treatment, and follow-up were deemed general and inadequate [15].

As AI is increasingly explored for patient education, several prior studies have looked at generating patient education materials from ChatGPT. One study used various LLM platforms to generate patient education material for both common and rare dermatologic conditions and found that these platforms were able to create general material that was accessible and understandable at an appropriate reading level [16]. Another study focused on the ability of ChatGPT to provide patients information about screening mammography but documented concerns about the understandability and actionability of the answers, specifically from a health literacy standpoint [17]. While AI seems to offer an opportunity to enhance patient education, there continues to be concern about the possibility of misinformation and reference quality.

This study offers new insight into the evolving role of LLMs in cancer patient education. Compared to top-ranked Google search results, ChatGPT consistently outperformed in terms of accuracy, completeness, and reference quality. Reviewers noted that ChatGPT responses were generally more comprehensive and better aligned with the clinical context, often including relevant treatment information and prognostic details that were lacking in standard search results. While previous concerns have been raised about the citation quality of earlier ChatGPT models, our findings suggest that reference quality has significantly improved and now exceeds that of typical online search outputs. Overall, these results support the potential of ChatGPT to generate personalized, high-quality educational content that may enhance patient understanding, support shared decision-making, and improve access to reliable health information.

Strengths and Limitations

Strengths of this study include the novel design of investigating the role of artificial intelligence in patient education for GYN malignancies. By querying ChatGPT 4.0 and Google with multiple patient scenarios, results were more generalizable to patients with different types and stages of GYNcancers. The use of blinded GYN oncology providers as reviewers allowed for assessment of patient education accuracy by the physicians who are guiding patients through treatment. Evaluation of accuracy, completeness, and reference quality offered a comprehensive assessment of the content of patient education material.

Some limitations of this study include a relatively small number of reviewers, a subjective grading scale, and no direct assessment of the readability of the sources used. While the viewers were blinded to others' responses, there is the possibility of individual bias in terms of interpretation of completeness and accuracy and how these compare to evidence-based guidelines. Finally, health literacy is an important consideration when creating and evaluating patient education resources, as comprehension and treatment compliance is often correlated with a patient's understanding of their diagnosis.

Conclusion:

The results of this study suggest that ChatGPT may be a valuable resource for patient education in the field of GYN oncology. As AI continues to evolve and become more readily accessible in the healthcare setting, it may be preferable to refer patients LLMs like ChatGPT to gather patient-specific information about their diagnosis, prognosis, and treatment options. ChatGPT may be able to bridge the gap between patient-oriented education and evidence-based guidelines while providing answers on one easy-access platform rather than multiple sites. As the role of ChatGPT evolves in the patient education space, it will be important to consider the reading level of education materials and make any necessary adaptations to match patients' health literacy levels. Future research will need to address all these issues in order to continue to safely integrate artificial intelligence into both patient education and clinical practice.

Acknowledgements:

AHS contributed to conceptualization, data collection, statistical analysis, and manuscript writing. MA contributed to data collection, statistical analysis, and manuscript editing. IL contributed to conceptualization, data collection, and manuscript editing. RR contributed to data collection and manuscript writing. DMC contributed to conceptualization, data collection, and manuscript editing.

Conflicts of Interest: None declared

References:

1. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 8 Registries, Nov 2023 Sub (1975-2021) - Linked To County Attributes - Time Dependent (1990-2022). Income/Rurality, 1969-2022 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2024, based on the November 2023 submission.
2. Bhattad PB, Pacifico L. Empowering Patients: Promoting Patient Education and Health Literacy. *Cureus*. 2022;14(7):e27336. Published 2022 Jul 27. doi:10.7759/cureus.27336.

3. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. *Front Oncol*;2023(13):1268915. doi:10.3389/fonc.2023.1268915.
4. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*;2023(388):1201-1208. doi:10.1056/NEJMra2302038.
5. Tozuka R, Johnno H, Amakawa A, et al. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. *Jpn J Radiol*;2024. doi:10.1007/s11604-024-01705-1.
6. Hewitt KJ, Wiest IC, Carrero ZI, et al. Large language models as a diagnostic support tool in neuropathology. *J Pathol Clin Res*;2024(10):e70009. doi:10.1002/2056-4538.70009.
7. Anderson PB, Wanken ZJ, Perri JL, et al. Patient information sources when facing repair of abdominal aortic aneurysm. *J Vasc Surg*;2020(71):497-504. doi:10.1016/j.jvs.2019.04.460.
8. Łaskiewicz J, Krajewski W, Tomczak W, et al. Performance of ChatGPT in providing patient information about upper tract urothelial carcinoma. *Contemp Oncol Pozn*;2024(28):172-181. doi:10.5114/wo.2024.141567.
9. Gupta M, Gupta P, Ho C, Wood J, Guleria S, Virostko J. Can generative AI improve the readability of patient education materials at a radiology practice? *Clin Radiol*;2024(79):e1366-e1371. doi:10.1016/j.crad.2024.08.019.
10. Piazza D, Martorana F, Curaba A, et al. The consistency and quality of ChatGPT responses compared to clinical guidelines for ovarian cancer: a Delphi approach. *Curr Oncol Tor Ont*;2024(31):2796-2804. doi:10.3390/curroncol31050212.
11. Finch L, Broach V, Feinberg J, et al. ChatGPT compared to national guidelines for management of ovarian cancer: did ChatGPT get it right? - A Memorial Sloan Kettering Cancer Center Team Ovary study. *Gynecol Oncol*;2024(189):75-79. doi:10.1016/j.ygyno.2024.07.007.
12. Loeser C. Educational materials | Foundation for Women's Cancer. Foundation for Women's Cancer - Eradicate or Lessen the Impact of Gynecologic Cancer. Published February 11, 2025. <https://foundationforwomenscancer.org/resources/educational-materials/>.
13. Chou HH, Chen YH, Lin CT, et al. AI-driven patient support: Evaluating the effectiveness of ChatGPT-4 in addressing queries about ovarian cancer compared with healthcare professionals in gynecologic oncology. *Support Care Cancer*. 2025;33(4):337. Published 2025 Apr 1. doi:10.1007/s00520-025-09389-7.
14. Karimov Z, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: comparative study of usefulness and reliability of chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol*;2024(281):2145-2151. doi:10.1007/s00405-023-08423-w.
15. Cavnar Helvacı B, Hepsen S, Candemir B, et al. Assessing the accuracy and reliability of ChatGPT's medical responses about thyroid cancer. *Int J Med Inf*;2024(191):105593. doi:10.1016/j.ijmedinf.2024.105593.
16. Lambert R, Choo ZY, Gradwohl K, Schroedl L, Ruiz De Luzuriaga A. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. *JMIR Dermatol*;2024(7):e55898. doi:10.2196/55898.
17. Spuur K, Currie G, Al-Mousa D, Pape R. Suitability of ChatGPT as a source of patient information for screening mammography. *Health Promot Pract*;2024. doi:10.1177/15248399241285060.

Supplementary Files

Multimedia Appendixes

Supplemental data.

URL: <http://asset.jmir.pub/assets/8f21ff6c13875c8455ba52f3dcbccc33.docx>