Assessing the Accuracy of ChatGPT's Answers to Common Uterine and Cervical Cancer Questions

Aqdas Malik[1]*, Rashida Suleiman[2], Sarya Bala[2], Shima Ibrahim[3],
Moza Al Kalbani[2], Hilal Al-Busaidi[2] and Ikram A Burney[2]

[1]Department of Information Systems, Sultan Qaboos University, Muscat, Oman

[2]Sultan Qaboos Comprehensive Cancer Care & Research Centre

[3]Department of Gynecological Oncology, Bezmialem Vakıf University

Received: 13 March 2025

Accepted: 2 December 2025

*Corresponding author: a.malik@squ.edu.om

DOI 10.5001/omj.2025.106

Abstract

***Background**:* Emerging digital technologies have the capacity to provide access to the general public about various aspects of cancer management, including prevention strategies, recommendations for screening, and management modalities. The use of the Artificial Intelligence (AI) platforms, such as ChatGPT has grown exponentially and is being used increasingly to seek information by the public and medical professionals alike.

***Objectives**:* To evaluate the accuracy of responses provided by ChatGPT on the prevention, screening, treatment, and risk factors of common gynecological cancers. The assessment primarily focuses on the usage of ChatGPT by the primary care providers and public with limited subject-specific knowledge.

Methods: We studied the reliability of ChatGPT in answering questions about two of the most common forms of gynecological cancers. ChatGPT version 3.5 was posed with 20 questions each about prevention, screening, and treatment of cancer of the uterus and the uterine cervix. The responses were reviewed and rated as accurate, inadequate, or inaccurate by 5 independent physicians and gynecologists with experience of 18±3 years working in gynecological oncology. The specialists provided reasons for marking the response inadequate or inaccurate.

Results: Overall, 20 out of 40 (50%) responses by ChatGPT were regarded as either inaccurate or inadequate. The inaccurate and inadequate responses were related to questions regarding the treatment of the two cancers. However, responses related to questions about p

Conclusion: ChatGPT may provide accurate information about prevention of gynecological cancers, but public and medical practitioners should not rely on the responses to make medical decisions, as most of the responses in this domain were inadequate, or even inaccurate. Qualified physicians should be consulted to obtain reliable information and to make individualized decisions. However, if ChatGPT were to be used as a source of information, it could be integrated with clinician oversight to improve accuracy.

Keywords: Cancer, Uterus, Cervix Uteri, Cancer prevention, Cancer Screening, Oncology Artificial Intelligence, Large Language Models , ChatGPT

Introduction

Gynecological cancers constitute the 5th most common cause of cancer worldwide[18]. There is a growing public awareness about prevention, early detection, early diagnosis, and treatment of these cancers. High-quality information for patients and their relatives is widely available in print and digital media[1]. Several organizations, such as the World Health organization (WHO) publish information about prevention, screening, and management of these cancers; however, the information is generic[2].

The advent of various emerging digital technologies plays a vital role in cancer management. Websites such as YouTube, Google search, and WebMD have provided access to the public about various aspects of management of cancer, including prevention strategies, recommendations for screening and early detection, and management modalities[3]. Through these technologies, patients can enhance self-management skills, access disease-related information remotely, and improve survivorship knowledge[19]. However, there are concerns about the quality of information obtained, especially in terms of accuracy and reliability.

More recent technological advancements including Artificial Intelligence (AI) has shown a strong potential for patient education, cancer screening, symptom monitoring, and survivorship care delivery that ultimately leads to improve the quality of care provided to the cancer patients[4] ChatGPT is one of the leading conversational AI platforms being used by billions of active users across the globe. Today more and more people use ChatGPT to access information related to various medical conditions, including various aspects of the management of cancer. In clinical scenarios, ChatGPT has performed better than or on a par with other AI systems and even human physicians. The model, however, performs inconsistently and varies by usage. Perera Molligoda Arachchige et al. (2023) have highlighted ChatGPT's strengths and limitations as a source of medical information[5]. For instance, ChatGPT enhanced medical workflows through report generation, increased efficiency, and reduced clinician workload, but it has the propensity to produce false positives[5]. A number of other studies have also reported varied accuracy levels in AI-generated cancer management information through this tool. Though ChatGPT has been promising in retrieving key oncological information[6], its results are not always aligned with established guidelines[7].For instance, in the medical oncology settings, the platform has shown potential in reviewing and extracting key contents from cancer patient records, interpreting next-generation sequencing reports and suggesting clinical trial options[6]. ChatGPT has also demonstrated accuracy in providing information about common cancer myths and misconceptions that align closely with the answers provided by the National Cancer Institute[8.]

On the contrary, AI technologies such as ChatGPT have limitations that are imperative to recognize particularly within the healthcare context. Inaccurate or inadequate information may not only produce emotional and psychological distress, but also may have a negative impact on the decision-making process and the subsequent health outcomes. The platform has struggled to reliably and robustly provide cancer treatment recommendations that align with National Comprehensive Cancer Network (NCCN) guidelines[7]. For example, the accuracy and reliability of responses generated by the platform related to the patient, surveillance, and diagnosis of liver, prostate and colon cancers were also deemed not optimal by the physicians[9] [10] [11] [12].

In the current study, we report the accuracy and reliability of ChatGPT in providing information about prevention, early detection and screening, diagnosis, and stage-specific treatment choices of the two most common forms of gynecological cancers, the uterine cancer and the cervical cancer. The primary objective is to assess how accurate the response might be, if the public or the primary care practitioners with limited knowledge about the subject of management of common gynecological cancers were to consult the freely available version of ChatGPT.

Methods

Two authors separately created twenty questions related to the management of endometrial and cervical cancers. Both authors have been engaged in providing continuing medical education (CME) to primary care practitioners, keeping them current with the best practices and most up-to-date information in gynecologic oncology. As part of this study, the authors systematically reviewed the most frequent questions for primary care providers and compiled and analyzed common issues regarding cervical and uterine cancer. In addition to gathering these repeated questions, the authors applied their extensive clinical practice to frame additional questions that reflect actual real-life issues faced in practice. Through this systematic approach, the derived questions precisely address the most common patient queries, rendering the evaluation of ChatGPT's responses highly relevant to real-life actual medical dialogue. A third author reviewed and refined the questions. The questions were then submitted to ChatGPT 3.5 on March 19, 2024. ChatGPT answers were independently assessed by 5 physicians working in academic centers in two different countries (a medical oncologist and 4 gynecological oncologists). Adopting the methodology incorporated by similar studies[13] [14] [15] all reviewer categorized each answer as accurate (all information is correct and relevant); inadequate (information is correct, but either incomplete, or irrelevant); or inaccurate (not correct information). A score of (+1) was assigned for accurate response, (0) for inadequate response, and (-1) for incorrect response. For each answer, the mean reviewer score was calculated. The responses were considered accurate if the mean score was greater than 0.5, inaccurate if less than -0.5; and inadequate if between 0.5 and -0.5. The reviewers were asked to provide explanations if in their opinion the answers provided by ChatGPT were inadequate or inaccurate. This study did not require institutional ethics committee approval because no human subject research was conducted.

Results

The median age of the five reviewers was 45 (range: 40-61) years, with a mean of 18±3 years' experience in gynecological oncology. Table 1 shows questions, mean scores, and the reviewer's explanation for inadequate or inaccurate responses by the ChatGPT for cervical cancer, and table 2 shows the same information for uterine cancer.

Table 1: Cervical Cancer related questions, score, and comments.

Question No.	**Question**	**Score**	**If response is inadequate, or inaccurate, comments**
1.	What are the three most common symptoms of cervical cancer?	1	
2.	Is there anything I can do to prevent cervical cancer?	1	
3.	What is the relationship between human papilloma virus and cervical cancer?	1	
4.	How is cervical cancer treated?	0.4	Trachelectomy was not mentioned as an option for early stage and as fertility-sparing techniques. Conversely, radical hysterectomy was stated as a treatment for advanced disease.

5.	Who is eligible for a cervical cancer screening?	0.8	People with a cervix who are not sexually active and LGBTQ class can also choose to screen
6.	What are the options for cervical cancer screening?	0.4	HPV is not mentioned as the gold standard, and VILI and VIA are used in low-resource-setting only.
7.	Are the results of cervical screening reliable?	0.4	The sensitivity and specificity were not mentioned
8.	At what age should screening for cervical cancer start?	0.8	Conditions for screening regardless of sexual activity are not mentioned
9.	How frequently should a person undergo cervical cancer screening?	1	
10.	Dose pregnancy increase the risk of developing cancer of cervix?	1	
11.	What is the effect of treatment of cervical cancer on subsequent pregnancy?	0.2	Chemoradiation causes infertility, and this is not mentioned
12.	What would happen if the screening for cervical cancer showed an abnormal result?	0.4	Simple hysterectomy is an option for women who have completed their family
13.	What are the different stages of premalignant lesion of the cervix?	1	
14.	What are the treatment options for premalignant lesion of the cervix?	0.2	Hysterectomy is a treatment option for CIN3 who have completed their family, and this was not mentioned.
15.	What are the treatment options of early-stage cervical cancer?	-0.8	Early-stage cervical cancer is typically classified as either stage IA (confined to the cervix) or stage IB (spread beyond the cervix but not to nearby organs). The treatment of choice is surgery. Chemoradiation is not often used for early-stage disease.
16.	What are the treatment options for locally advanced cervical cancer?	-0.6	Locally advanced disease refers to stage IIB, Stage III, or stage IVA disease. Chemoradiation is the treatment of choice. Mention of surgery, targeted therapy or clinical trial may be misleading. Immunotherapy has recently been tested but not approved for use in this setting.
17.	What are the treatment options for metastatic and recurrent cervical cancer	0.2	Preferred options and indications for immune checkpoint inhibitor, such as CPS score are not mentioned
18.	What is the risk involved in undergoing surgery for cervical cancer	0.2	DVT, ureteric, bladder and bowel injury not mentioned
19.	What are the side effects of radiotherapy for cervical cancer?	0.8	General information is provided but the chances of a particular side effect or chances of severe side effects are not mentioned
20.	What are the side effect of chemotherapy for cervical cancer?	0.6	General information is provided but the chances of a specific side effect or chances of severe side effects not

mentioned e.g. peripheral neuropathy, cutaneous toxicity, etc.

Table 2: Uterine Cancer related questions, score, and comments.

Question No.	Question	Score	If response is inadequate, or inaccurate, comments
1.	What are the risk factors for uterine cancer?	0.4	Missing risk factors such as: early *menarche and late menopause,* genetic mutation (other than MMR Gene): and medical conditions such as Liver cirrhosis due to impairment of estrogen degradation
2.	What common genetic mutations are associated with uterine cancer?	0.8	Chances of developing uterine cancer in case of genetic mutations was not provided
3.	Who should have genetic testing for uterine cancer?	0.6	Early onset of bilateral cancers is a risk relevant to ovarian cancer, not uterine cancer
4.	What is the standard screening test for endometrial cancer?	0.6	TVUS and biopsy are diagnostic, not screening tests for the general population.
5.	What is the risk of developing endometrial cancer in patients who take tamoxifen for treatment of breast cancer?	0	The numerical value of risk is not provided. Tamoxifen associated endometrial cancer is often strongly associated with poor prognosis, such as type 2 carcinoma, and sarcoma, rather than endometroid type
6.	What is the lifetime risk of endometrial cancer in people known to have Cowden syndrome?	0.6	The lifetime risk is mentioned as high as 51%. The recorded risk of endometrial cancer is 1 in 4 women (around 25-28%)
7.	Could a patient with endometrial cancer receive fertility sparing treatment?	0.8	
8.	What options are available for fertility-sparing in patients with endometrial cancer?	0.4	Ovarian transposition is an option for cervical cancer, not for endometrial cancer
9.	Which progesterone is best for the treatment of endometrial cancer?	1	
10.	What are predictors of response to progesterone for treatment of endometrial cancer?	0.8	
11.	What is the exact duration of therapeutic benefit from progesterone therapy in patients with endometrial cancer wishing to preserve fertility?	0.2	The answer highlighted factors related to treatment response and treatment process, rather than answering the estimate duration by evidence)
12.	Is pregnancy possible after receiving progesterone therapy for endometrial cancer?	0.8	
13.	What are the risks of fertility preservation in endometrial cancer?	0.4	Lack of a tumor specimen (histology) may limit detection of Lynch syndrome. Risk of an undetected synchronous ovarian cancer Both are not mentioned

14.	What is the prognosis of endometrial cancer?	0.4	Molecular classification and risk groups are not included in the response
15.	What is the standard primary surgical management for endometrial cancer?	0.4	Management is based on stage and grade. e.g, omentectomy is done in certain histological types, even if there is no apparent disease on the omentum, and this is not mentioned
16.	What is the best surgical approach for management of endometrial cancer?	1	
17.	What is the role of surgery in patients with advance stage endometrial cancer?	0.6	
18.	I am recently diagnosed with metastatic endometrial cancer and my doctor is planning for surgery. Do I need any further treatment after surgery?	0	Adjuvant chemotherapy is administered if the disease were localized, and not when it was metastatic to start with. In that case, palliative chemotherapy is given. Radiotherapy is not part of the management of metastatic disease
19.	I am recently diagnosed with metastatic endometrial cancer and my doctor decided that I'm not fit for surgery. What are the treatment options?	-0.2	Radiotherapy, especially brachytherapy is not part of treatment of metastatic endometrial cancer
20.	What are the treatment options for recurrence of endometrial cancer?	0.2	No mention of medications to be used. Palliative chemotherapy with or without immune checkpoint inhibitors should be first option . -Inhibitors of PI3K/AKT/mTOR pathway are not approved for the management of endometrial cancer

For questions related to cervical cancer, 2 responses were regarded as inaccurate (both related to treatment options for various stages of cervical cancer), and 8 options were regarded as inadequate. The reliability rate (accurate response = overall score >0.5) was 50% (10/20 questions). Similarly, for questions related to endometrial cancer, 10 responses were regarded as inadequate (mainly related to treatment, including fertility sparing options and duration of treatment, prognosis, and specifics of surgery, and chemotherapy. Toxic treatments, such as radiotherapy and brachytherapy were proposed as options of treatment for metastatic endometrial cancer). The reliability rate for responses related to endometrial cancer (accurate response = overall score >0.5) was also 50% (10/20 questions).

For assessing the inter-coder reliability of the physician's ratings for the responses of ChatGPT to cervical and uterine cancer questions, we used Krippendorff's Alpha. As a robust statistical measure, Krippendorff's Alpha evaluates the agreement between multiple raters when dealing with ordinal data. This measure is also deemed a suitable choice for this research setting as Krippendorff's Alpha is less sensitive to the number of raters or categories. Krippendorff's Alpha yielded a value of 0.85 for cervical cancer questions that indicate a high inter-coder reliability. Likewise, the Alpha value was 0.78 for uterine cancer questions, reflecting substantial agreement among raters. Results from this statistical measure suggest that ratings of physicians were reliable and consistent among both the sets of questions on cervical and uterine cancer.

Discussion

A total of 20 out of 40 (50%) responses by the Chat GPT were regarded as either 'inaccurate' (not correct information) or 'inadequate' (information is correct, but either incomplete, or irrelevant). The inaccurate and inadequate responses were related to questions regarding treatment (Q4, 11,12, and 14-18 for cervical cancer, and Q8, 11,13, and 14, 15, and 18-20 for endometrial cancer) or screening (Q6,7 for cervical cancer), and risk factors (Q1 for endometrial cancer). However, responses related to questions about prevention, eligibility for screening, premalignant lesions of the cervix, and the side effects of treatment of cervical cancer were regarded as 'accurate'. Similarly, responses related to questions about genetic mutations, genetic testing, lifetime risk of developing uterine cancer, and non-surgical treatment of endometrial cancer were regarded as 'accurate'. However, only 6/20 responses (30%) for cervical cancer, and 2/20 responses (10%) for endometrial cancer received a maximal mean score of 1. More than 85% of responses were deemed appropriate and consistent by at least 2 independent experts. Generally, mean score was higher for questions related to prevention and risk factors for cervical cancer, and this pattern is consistent with the report from a similar study about HCC[9]. Similar inferences were drawn about ChatGPT as a patient information source for prostate cancer[10]. However, our method was different to the method used in a similar study about the appropriateness of answers of ChatGPT to common questions on colon cancer, where responses were judged in relation to the American Society of Colon and Rectal Surgeon handouts. Moreover, the authors used a different scoring system in that study.

It was observed that the majority of inaccurate or inadequate responses were related to specific treatment options, especially related to clinical stage. Although AI surpassed the USMLE pass-mark[16] and made correct diagnoses in the Emergency room[17] more often than the human doctors (97% vs 87%), yet a significant number of responses related to stage-specific treatment recommendations were inadequate. These included either irrelevant information, such as, surgery, immunotherapy and targeted therapy in locally advanced cervical cancer, or incomplete information, such as, the failure of ChatGPT to mention hysterectomy as a treatment option for CIN3 if the woman has completed family life, or failure to mention risks involved in surgery for cervical cancer, or more specifically, the lack of mention of combined positive score (CPS) to select patients for immune checkpoint inhibitor in metastatic or recurrent cancer of cervix.

There are several limitations of the study. We only evaluated ChatGPT 3.5. Although ChatGPT is not designed to provide accurate medical information, it is assumed that patients and the public will use ChatGPT to obtain information, and this is precisely what we wish to highlight. At the moment, ChatGPT may not provide 'medically accurate' answers. Its reliability in providing accurate answers may improve over time with more data integration. As an input to ChatGPT, questions asked may be repeated with varied phrasing, at different times, and across diverse cultural contexts worldwide. Since there is rapidly increasing awareness about ChatGPT, even this function of ChatGPT may improve over time and it may recognize different phrases. Secondly, the five reviewers from one geographical area, and this to some extent may limit the generalizability of our results and inferences. However, it should be noted that there was a medical oncologist, and gynecological oncologists, working in at least two different countries. Furthermore, they responded to ChatGPT answers independently. However, it is recognized that similar studies should be carried out in other regions of the world and in different languages. Finally, we used three scores: +1, 0 or -1. This lends to subjectivity in categorizing an answer as inadequate, however, we asked the reviewers to provide explanation if the answers were regarded inadequate or inaccurate.

In summary, ChatGPT may provide accurate information about prevention of gynecological cancers, but people should not rely on the responses to make medical decisions, as most of the responses in this domain were inadequate. Patients and caregivers must continue to consult qualified physicians to obtain reliable and accurate information and to make individualized decisions. Also, public and primary care physicians should be made aware that if ChatGPT were to be integrated with clinician oversight, the accuracy of information would increase considerably.

Conclusion

This study aims to assess the accuracy of information generated by ChatGPT about prevention, early detection and screening, diagnosis, and stage-specific treatment choices of cervical and uterine cancers. Our analysis highlights that while ChatGPT provides accurate information concerning prevention and screening, responses related to the treatment of the two cancers were often inadequate or inaccurate. These findings represent an evident limitation for the use of AI-generated models in clinical decision-making and emphasize a clear need for cautious practice of these tools, with human oversight to ensure alignment and accuracy with clinical guidelines. Future studies should compare the performance (including accuracy and consistency) of newer iterations of ChatGPT with other available AI tools (Claude, Gemini, Deepseek, and Qwen etc.). Finally, researchers in this domain should develop comprehensive frameworks for integrating AI technologies and tools into various clinical workflows and settings.

Disclosure

The authors report no conflicts of interest with the present work.

References

1. Godfrey, K., Agatha, T., & Nankumbi, J. (2016). Breast cancer knowledge and breast self-examination practices among female university students in Kampala, Uganda: a descriptive study. Oman Medical Journal, 31(2), 129.

2. Kakotkin VV, Semina EV, Zadorkina TG, Agapov MA. Prevention Strategies and Early Diagnosis of Cervical Cancer: Current State and Prospects. Diagnostics (Basel). 2023 Feb 7;13(4):610. doi: 10.3390/diagnostics13040610.

3. Shaffer KM, Turner KL, Siwik C, Gonzalez BD, Upasani R, Glazer JV, Ferguson RJ, Joshua C, Low CA. Digital health and telehealth in cancer care: a scoping review of reviews. The Lancet Digital Health. 2023 May 1;5(5):e316-27.

4. Kashoub, M., Al Abdali, M., Al Shibli, E., Al Hamrashdi, H., Al Busaidi, S., Al Rawahi, M., ... & Al Alawi, A. (2023). Artificial Intelligence in Medicine: A Double-edged Sword or a Pandora's Box?. Oman Medical Journal, 38(5), e542.

5. Perera Molligoda Arachchige AS. Role of ChatGPT 3.5 in emergency radiology, with a focus on cardiothoracic emergencies: Proof with examples. iRADIOLOGY. 2024.

6. Uprety D, Zhu D, West H. ChatGPT—a promising generative AI tool and its implications for cancer care. Cancer. 2023 Aug 1;129(15):2284-9.

7. Chen S, Kann BH, Foote MB, Aerts HJ, Savova GK, Mak RH, Bitterman DS. The utility of ChatGPT for cancer treatment information. MedrXiv. 2023 Mar 23:2023-03.

8. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI cancer spectrum. 2023 Apr 1;7(2):pkad015.

9. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, Kamaya A, Tse JR. Accuracy of Information Provided by ChatGPT Regarding Liver Cancer Surveillance and Diagnosis. AJR Am J Roentgenol. 2023 Oct;221(4):556-559. doi: 10.2214/AJR.23.29493.

10. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer?, Urology, Volume 180, 2023, Pages 35-58, ISSN 0090-4295, https://doi.org/10.1016/j.urology.2023.05.040.

11. Emile SH, Horesh N, Freund M, Pellino G, Oliveira L, Wignakumar A, Wexner SD. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? Surgery. 2023 Nov;174(5):1273-1275. doi: 10.1016/j.surg.2023.06.005.

12. Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: An assessment of ChatGPT responses. Am J Otolaryngol. 2024 Jan-Feb;45(1):104085. doi: 10.1016/j.amjoto.2023.104085.

13. Alasker A, Alsalamah S, Alshathri N, Almansour N, Alsalamah F, Alghafees M, AlKhamees M, Alsaikhan B. Performance of large language models (LLMs) in providing prostate cancer information. BMC urology. 2024 Aug 23;24(1):177.

14. Freire Y, Laorden AS, Pérez JO, Sánchez MG, García VD, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. The Journal of Prosthetic Dentistry. 2024 Apr 1;131(4):659-e1.

15. Olszewski R, Brzezinski J, Watros K, Manczak M, Owoc J, Jeziorski K. Exploring the role of AI-driven chatbots in patient care: a critical evaluation amidst healthcare staff shortages. European Heart Journal. 2024 Oct;45(Supplement_1):ehae666-3495.

16. Singhal K, Azizi S, Tu T et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2

17. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, O'Connor RD, van Ginneken B, Kurstjens S. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. Ann Emerg Med. 2024 Jan;83(1):83-86. doi: 10.1016/j.annemergmed.2023.08.003

18. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-249. doi:10.3322/caac.21660

19. Senbekov M, Saliev T, Bukeyeva Z, Almabayeva A, Zhanaliyeva M, Aitenova N, Toishibekov Y, Fakhradiyev I. The recent progress and applications of digital technologies in healthcare: a review. International journal of telemedicine and applications. 2020;2020(1):8830200.

20. Introducing ChatGPT. https://openai.com/blog/chatgpt/. Accessed 16 June, 2024.