

Contents lists available at ScienceDirect

Gynecologic Oncology Reports



journal homepage: www.elsevier.com/locate/gynor

The doc versus the bot: A pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology

Mary Katherine Anastasio^{a,*}, Pamela Peters^a, Jonathan Foote^b, Alexander Melamed^c, Susan C. Modesitt^d, Fernanda Musa^e, Emma Rossi^a, Benjamin B. Albright^f, Laura J. Havrilesky^a, Haley A. Moss^a

^a Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC, USA

^b Commonwealth Gynecologic Oncology, Bon Secours Health, Richmond, VA, USA

^c Division of Gynecologic Oncology, Vincent Department of Obstetrics & Gynecology, Massachusetts General Hospital, Boston, MA, USA

^d Division of Gynecologic Oncology, Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, GA, USA

^e Swedish Cancer Institute, Seattle, WA, USA

^f Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of North Carolina Chapel Hill, Chapel Hill, NC, USA

ARTICLE INFO

Keywords: Artificial intelligence Gynecologic oncology Patient education

ABSTRACT

Artificial intelligence (AI) applications to medical care are currently under investigation. We aimed to evaluate and compare the quality and accuracy of physician and chatbot responses to common clinical questions in gynecologic oncology. In this cross-sectional pilot study, ten questions about the knowledge and management of gynecologic cancers were selected. Each question was answered by a recruited gynecologic oncologist, ChatGPT (Generative Pretreated Transformer) AI platform, and Bard by Google AI platform. Five recruited gynecologic oncologists who were blinded to the study design were allowed 15 min to respond to each of two questions. Chatbot responses were generated by inserting the question into a fresh session in September 2023. Qualifiers and language identifying the response source were removed. Three gynecologic oncology providers who were blinded to the response source independently reviewed and rated response quality using a 5-point Likert scale, evaluated each response for accuracy, and selected the best response for each question. Overall, physician responses were judged to be best in 76.7 % of evaluations versus ChatGPT (10.0 %) and Bard (13.3 %; p < 0.001). The average quality of responses was 4.2/5.0 for physicians, 3.0/5.0 for ChatGPT and 2.8/5.0 for Bard (t-test for both and ANOVA p < 0.001). Physicians provided a higher proportion of accurate responses (86.7 %) compared to ChatGPT (60 %) and Bard (43 %; p < 0.001 for both). Physicians provided higher quality responses to gynecologic oncology clinical questions compared to chatbots. Patients should be cautioned against non-validated AI platforms for medical advice; larger studies on the use of AI for medical advice are needed.

1. Introduction

The application of artificial intelligence (AI) methodologies in medicine is growing significantly, particularly within the subfield of natural language processing (NLP) (Ramesh et al., 2004). NLP serves as an interface between computers and natural human language, and recent advancements in this field have enabled computers to write complex yet comprehensible responses (Lytinen, 2005). In November 2022, Open AI released a publicly available NLP tool known as ChatGPT (Generative Pretrained Transformer) (https://openai.com/blog/ chatgpt/). ChatGPT is a chatbot designed to respond to prompts or questions across a broad range of topics similar to human conversation. Soon after its introduction, ChatGPT became the fastest-growing consumer application with over 100 million users by January 2023 (Gitnux [Internet], 2023). It has been used to perform language-based tasks such as answering clinical questions, taking medical licensing examinations, and has even been proposed as a solution to the growing demands of electronic medical documentation (Gilson et al., 2023; Sallam, 2023;

https://doi.org/10.1016/j.gore.2024.101477

Received 3 May 2024; Received in revised form 3 August 2024; Accepted 6 August 2024 Available online 8 August 2024

2352-5789/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

^{*} Corresponding author at: Department of Obstetrics and Gynecology, Duke University, 2608 Erwin Rd, STE 210, Durham, NC 27705, USA. *E-mail address:* mm765@duke.edu (M.K. Anastasio).

Table 1

Study questions on management of gynecologic cancers and mean quality ratings for physician and chatbot responses.

Question	Mean quality rating*		
	Physician	ChatGPT	Bard
Does PARP inhibitor therapy following chemotherapy lead to improved survival outcomes for patients with advanced stage ovarian cancer?	5.0	2.7	3.3
At the time of diagnosis of advanced stage ovarian cancer, is it better to begin chemotherapy or have surgery to remove the cancer?	4.0	3.3	2.7
Can some forms of ovarian cancer be treated without chemotherapy?	3.7	2.0	2.0
Is it safe to have a laparoscopic hysterectomy for cervical cancer?		3.0	1.0
When are PET-CT scans indicated for surveillance in gynecologic cancers?	4.7	2.0	2.7
What are the options for ovarian cancer screening and prevention in patients with a germline BRCA1 or BRCA2 mutation?	3.7	2.7	3.3
When would a patient be recommended radiation therapy after surgery for early-stage endometrial cancer?	4.3	2.7	2.3
Is neuropathy from taxane-induced neuropathy curable?		4.0	3.3
Which patients with vulvar cancer are eligible for sentinel lymph node mapping?	2.7	3.7	4.3
What is the role of biomarkers in treatment for endometrial cancer?	5.0	3.7	2.7

*Quality scale was a 5-point Likert scale with 1 indicating very poor quality, 2: poor quality, 3: acceptable quality, 4: good quality, 5: very good quality. PARP: poly (ADP-ribose) polymerase; PET-CT: Positron emission tomography–computed tomography.

Sarraju et al., 2023). A recent study on the clinical potential of ChatGPT in obstetrics and gynecology demonstrated its ability to provide preliminary information on a range of topics in the specialty (Grunebaum et al., 2023).

In March 2023, Google announced a new and competing model known as Bard (Racing to Catch Up With ChatGPT). Bard has some unique differences from ChatGPT, which were met with both criticism and praise. Bard exhibits more caution compared to ChatGPT and often avoids providing medical, legal, or financial advice which may contain false information. For example, Bard responds with statements such as "I'm designed solely to process and generate text, so I'm unable to assist you with that" when prompted on recommendations for medical treatment. Given its novelty, there are very few publications on the utility of Bard in answering clinical questions.

Patients are increasingly relying on search engines and artificial intelligence chatbots for medical information, which leads to less reliance on medical professionals for healthcare advice. However, the potential of publicly available chatbots as clinical tools has been met with several valid concerns. First, while chatbots may provide casual, convincing responses to complex medical questions, AI models are known to "hallucinate" or provide false information in a way that often sounds honest and accurate. Because false information could be harmful to patients, caution should be used when considering the application of AI technology in medical practice. Additionally, AI tools do not cite the source of the information provided in a response, raising the issue of scientific misconduct. AI technologies are not directly connected to the internet and thus rely on training data updates. Thus, they may not provide the most up-to-date information depending on when the last training data update occurred (Sallam, 2023; Shen et al., 2023). Despite these limitations, some studies have shown that chatbots generate highquality and empathetic responses to patient clinical questions (Ayers et al., 2023; Goodman et al., 2023). The purpose of this study was to evaluate and compare the quality and accuracy of physician and chatbot responses from two publicly available chatbots; ChatGPT and Bard, to common clinical questions in gynecologic oncology.

2. Methods

In this cross-sectional pilot study, ten clinical questions about the knowledge and management of gynecologic oncology were generated by three study team members (MKA, HAM, LJH; Table 1). Five practicing gynecologic oncology attendings (two private, three academic) were then recruited for the study. Each clinical question was answered by 1) a recruited gynecologic oncologist, 2) ChatGPT AI platform, and 3) Bard by Google AI platform. Each practicing gynecologic oncologist was allowed 15 min to respond to each of two questions without using references to obtain answers during the response time. Gynecologic oncologists were blinded to the study design and unaware that their responses would be compared to chatbot responses. ChatGPT and Bard

responses were generated by inserting the question into a fresh session in September 2023. Qualifiers identifying the response source as an individual or referring the reader to a medical professional were removed to reduce bias. Responses were reviewed and rated independently by two board-certified gynecologic oncologists (HAM, LJH) and one gynecologic oncology fellow (PP) who were blinded to the response source. Each response (ChatGPT, Bard, and physician) to ten questions was reviewed for quality and accuracy by each of the three reviewers for a total of thirty reviews. The quality of the information was rated using a 5-point Likert scale (1-very poor, 2-poor, 3-acceptable, 4-good, 5-very good), and each response was evaluated for accuracy of medical information (accurate or inaccurate). Reviewers selected the best response of the three responses (gynecologic oncologist, ChatGPT, or Bard) for each question.

Continuous variables are reported using mean and standard deviation. Associations between categorical and continuous variables were assessed using *t*-test and analysis of variance when appropriate. Interrater reliability was evaluated using the kappa statistic. A p-value of < 0.05 was considered statistically significant. STATA version 18.0 was used for statistical analysis (Statacorp, College Station, TX, USA).

3. Results

Average physician responses (252.6 words) were shorter than ChatGPT (307.8 words) or Bard (307.4 words) chatbot responses. Overall, 6.7 % (2/30) of physician responses were rated by reviewers as poor quality, 13.3 % (4/30) acceptable, 33.3 % (10/30) good, and 46.7 % (14/30) very good. No physician responses were rated very poor quality by reviewers. Of ChatGPT responses, 6.7 % (2/30) were rated very poor quality, 26.7 % (8/30) poor, 30 % (9/30) acceptable, 36.7 % (11/30) good, and no ChatGPT responses were rated very good quality by reviewers (Table 2). Of Bard responses, 13.3 % (4/30) were rated very poor quality, 23.3 % (7/30) poor, 40 % (12/30) acceptable, and 20 % (6/30) good. No Bard responses were rated very good quality. The average quality of physician responses was 4.2/5.0 (SD=0.92) vs 3.0/ 5.0 (SD=0.96) for ChatGPT and 2.8/5.0 (SD=1.04) for Bard (t-test for both and ANOVA p < 0.001; Fig. 1). The proportion of responses rated good or very good was higher for physician responses (80 %) compared to ChatGPT and Bard (36.7 % and 23.3 %; p < 0.001 for both).

Physicians provided a higher proportion of accurate responses (86.7 %) compared to ChatGPT and Bard (60 % and 43 %, respectively; p < 0.001 for both). At least one rater reported inaccuracies in physician responses for 3 of 10 questions versus 8 of 10 questions for both ChatGPT and Bard responses. An example of an inaccurate response from Bard for "Can some forms of ovarian cancer be treated without chemotherapy?" was "Fallopian tube cancer is often diagnosed early, and surgery alone may be curative." Bard responded to "Is it safe to have a laparoscopic hysterectomy for cervical cancer?" with "Studies have shown that laparoscopic hysterectomy is just as effective as open

Table 2

Example of physician and chatbot responses to a question on the knowledge and management of cervical cancer with mean quality ratings for each response.

Question	Physician response	ChatGPT response	Bard response
Is it safe to have a laparoscopic hysterectomy for cervical cancer?	Laparoscopic hysterectomy can be a safe and effective option for treating cervical cancer in certain cases, but whether it is appropriate for a specific individual depends on various factors. The suitability of laparoscopic hysterectomy for cervical cancer depends on several factors, including the stage of cancer, the size and location of the tumor, the patient's overall health, and the surgeon's expertise. In some cases, laparoscopic hysterectomy may not be recommended, and other surgical or treatment options may be more appropriate. Traditional open surgery or a combination of treatments such as radiation therapy and chemotherapy may be recommended for more advanced cases. Laparoscopic hysterectomy has the advantage of being a minimally invasive procedure, which typically results in shorter hospital stays, faster recovery times, and smaller incisions compared to open surgery. However, it may not be suitable for every cervical cancer patient. It's essential to have a thorough discussion to understand the risks and benefits of different treatment options and to make an informed decision tailored to your specific condition. Your medical history, cancer stage, and individual circumstances may determine the most appropriate treatment plan for you. Mean quality rating: 4.7/5.0	For the earliest stage of cervical cancer, when the tumor is microscopic, superficial, and has not invaded into blood vessels or lymphatic channels, minimally invasive hysterectomy may be safe. However for more advanced cancers, when a radical hysterectomy has traditionally been considered to be needed, the best evidence suggests that minimally invasive surgery may increase the risk of cancer recurrence and death. For this reason, open surgery is considered standard in these cases. While there are several ongoing studies investigating if there may be a safe way to employ minimally invasive surgery in these more advanced cases of cervical cancer, they are not offered at all institutions.	Yes, it is safe to have a laparoscopic hysterectomy for cervical cancer. In fact, laparoscopic hysterectomy is now the preferred surgical approach for most women with early- stage cervical cancer. Laparoscopic hysterectomy is a minimally invasive surgery that is performed through several small incisions in the abdomen. The surgeon uses a laparoscope, a thin, tube-like camera, to view the inside of the abdomen and guide the surgery. Compared to traditional open hysterectomy, laparoscopic hysterectomy has a number of advantages, including: less blood loss, shorter hospital stay, faster recovery time, fewer complications. <u>Studies have shown that laparoscopic</u> <u>hysterectomy is just as effective as open hysterectomy is not appropriate for all women with cervical cancer. It may not be a good option for women with advanced-stage cancer, large tumors, or other medical conditions. Here are some of the potential risks of laparoscopic hysterectomy: bleeding, infection, damage to nearby organs, conversion to open hysterectomy if complications arise. It is important to note that these risks are rare, and laparoscopic hysterectomy is generally a safe procedure.</u>

Mean quality rating: 3.0/5.0

Mean quality rating: 1.0/5.0

Incorrect statement underlined.

hysterectomy in removing cervical cancer and preventing recurrence." ChatGPT provided an inaccurate response for "When are PET-CT scans indicated for surveillance in gynecologic cancers?" with "Patients who have completed surgery and chemotherapy or radiation therapy may undergo surveillance imaging with PET-CT to monitor any signs of cancer recurrence." In response to "What are the options for ovarian cancer screening and prevention in patients with a germline BRCA1 or BRCA2 mutation?", ChatGPT provided "bilateral salpingectomy" as an option and "returning to surgery at menopause to remove the ovaries." Overall, across 10 questions, the physician response was judged to be best in 76.7 % of the 30 total evaluations versus ChatGPT in 10 %, and Bard in 13.3 % (p < 0.001).

Evaluator ratings did not vary significantly across the 10 questions. There was high agreement in inter-rater reliability for best response of physician versus AI responses (kappa = 0.6273) and moderate agreement for high versus low quality ratings (kappa = 0.5536).

4. Discussion

In this cross-sectional pilot study evaluating physician and chatbot responses to common clinical questions in gynecologic oncology, physicians provided higher quality and more accurate responses compared to chatbots. Physician responses were judged to be superior to chatbot responses in the majority of evaluations. Together, these findings demonstrate the importance of relying on providers rather than chatbots for answers to complex clinical decision-making questions for patients with gynecologic cancer.

While the low quality of chatbot responses in our study is notable, the inaccuracies of one or both chatbot responses to almost every clinical question are more significant. For example, a Bard response incorrectly stated that ovarian cancers are often diagnosed early; nearly 75 % of

ovarian cancers are diagnosed at advanced stage (https://seer.cancer. gov/statfacts/html/common.html). Bard advised that laparoscopic hysterectomy is just as effective as open hysterectomy in removing cervical cancer and preventing recurrence; however, studies have shown inferior survival outcomes with laparoscopic surgery compared to open surgery for some patients with cervical cancer (Ramirez et al., 2018). ChatGPT advised bilateral salpingectomy and return for surgery at menopause for oophorectomy as an option for patients with germline mutations, while current recommendations are removal of bilateral fallopian tubes and ovaries between 35 and 45 depending on the specific mutation (Berek et al., 2010). Finally, ChatGPT also counseled on the use of PET-CT for surveillance imaging of gynecologic cancers; however, PET-CT is not routinely recommended for gynecologic cancer surveillance and is in fact discouraged unless patients report symptoms or have abnormal physical exam findings (Wisely, 2023). Misinformation from chatbots can be dangerous as it may lead to serious medical consequences and distrust between a patient and their provider.

In contrast to our findings, a prior study compared physician and AI chatbot responses to patient questions found on a public online forum and found that chatbots provided higher quality and more empathetic responses (Ayers et al., 2023). The prior study used a wide range of primary care topics and may have utilized fewer complex questions as the questions were generated by patients rather than physicians. In our study, we used complex questions generated by physicians about a subspecialized field with the expectation of nuanced answers, which may also explain the inaccuracies noted in our study. A second study on the accuracy and comprehensiveness of chatbot responses to physician-developed questions reported significant inaccuracies in responses and did not advise the use of chatbots for dissemination of medical information. However, the accuracy of chatbot responses improved over time (Goodman et al., 2023). This highlights the importance of regular



Fig. 1. Distribution of quality ratings of physician, ChatGPT, and Bard responses to gynecologic oncology clinical questions.

training data updates and ongoing improvements to current AI technologies to understand complex questions and provide more refined responses. While we do not advise the use of current AI chatbot versions by patients for clinical questions in gynecologic oncology, these tools may eventually become a valuable resource for gaining medical information with further research and validation studies.

A strength of this study is the investigation of use of the understudied but rapidly expanding field of artificial intelligence technology. Limitations include the limited number of questions utilized, physicians recruited to provide responses, and evaluators who rated responses, which all could introduce bias to the study. However, we found agreement in inter-rater reliability for best response and high versus low quality ratings. This serves as a pilot study, and future studies should compare physician versus AI chatbot responses to other types of clinical questions within gynecologic oncology. Lastly, the questions used as prompts for this study were intentionally complex, and subtleties within responses were anticipated. Future studies investigating the use of artificial intelligence technologies for simpler questions with straightforward responses are needed.

5. Conclusion

Physicians provided higher quality responses to common clinical questions in gynecologic oncology compared to ChatGPT and Bard chatbots. Chatbots provided longer, less accurate responses. When counseling patients on the knowledge and management of gynecologic cancers, one should consider the potential for patients to gather false information from AI technologies. Patients should be cautioned against non-approved/validated AI platforms for medical advice, and larger studies on the use of AI for medical advice are needed.

CRediT authorship contribution statement

Mary Katherine Anastasio: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. Pamela Peters: Writing – review & editing, Data curation, Conceptualization. Jonathan Foote: Writing – review & editing, Data curation. Alexander Melamed: Writing – review & editing, Data curation. Susan C. Modesitt: Writing – review & editing, Data curation. Fernanda Musa: Writing – review & editing, Data curation. Fernanda Musa: Writing – review & editing, Data curation. Fernanda Musa: Writing – review & editing, Data curation. Emma Rossi: Writing – review & editing, Data curation. Benjamin B. Albright: Writing – review & editing, Software, Formal analysis, Data curation. Laura J. Havrilesky: Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization. Haley A. Moss: Writing – review & editing, Supervision, Methodology, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., et al., 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern. Med. 183 (6), 589–596.
- Berek, J.S., Chalas, E., Edelson, M., Moore, D.H., Burke, W.M., Cliby, W.A., et al., 2010. Prophylactic and risk-reducing bilateral salpingo-oophorectomy: recommendations based on risk of ovarian cancer. Obstet Gynecol. 116 (3), 733–743.
- Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., et al., 2023. 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. 9, e45312.
- Gitnux [Internet]. [04/07/2023]. Available from: https://blog.gitnux.com/chat-gptstatistics/.
- Goodman, R.S., Patrinely, J.R., Stone Jr., C.A., Zimmerman, E., Donald, R.R., Chang, S.S., et al., 2023. Accuracy and Reliability of Chatbot Responses to Physician Questions. J. Am. Med. Assoc.netw Open. 6 (10), e2336483.
- Grunebaum, A., Chervenak, J., Pollet, S.L., Katz, A., Chervenak, F.A., 2023. The Exciting Potential for ChatGPT in Obstetrics and Gynecology. Am. J. Obstet. Gynecol.
- Lytinen, S.L., 2005. Artificial intelligence: Natural language processing. Van Nostrand's Scientific Encyclopedia.
- OpenAI. ChatGPT: optimizing language models for dialogue [Available from: https:// openai.com/blog/chatgpt/.
- Racing to Catch Up With ChatGPT, Google Plans Release of Its Own Chatbot. The New York Times.
- Ramesh, A.N., Kambhampati, C., Monson, J.R., Drew, P.J., 2004. Artificial intelligence in medicine. Ann. R Coll Surg. Engl. 86 (5), 334–338.
- Ramirez, P.T., Frumovitz, M., Pareja, R., Lopez, A., Vieira, M., Ribeiro, R., et al., 2018. Minimally Invasive versus Abdominal Radical Hysterectomy for Cervical Cancer. N Engl. J. Med. 379 (20), 1895–1904.
- Sallam, M., 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (basel). 11 (6).
- Sarraju, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F., Laffin, L., 2023. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. J. Am. Med. Assoc. 329 (10), 842–844.
- SEER Cancer Statistics Factsheets: Common Cancer Sites. National Cancer Institute Bethesda, MD [Available from: https://seer.cancer.gov/statfacts/html/common. html.
- Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., et al., 2023. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 307 (2), e230163.
- Choosing Wisely: Five tips for a meaningful conversation between patients and providers: Society of Gynecologic Oncology; 2023 [Available from: https://www.sgo.org/resources/choosing-wisely-five-tips-for-a-meaningful-conversation-between-patients-and-providers-2/.