Article

Consistency and Quality of ChatGPT Responses Compared to Clinical Guidelines for Ovarian Cancer: A Delphi Approach.

Dario Piazza , Federica Martorana , Annabella Curaba , Daniela Sambataro , Maria Rosaria Valerio , Alberto Firenze , Basilio Pecorino , Paolo Scollo , Vito Chiantera , Giuseppe Scibilia , Paolo Vigneri , Vittorio Gebbia * , Giuseppa Scandurra

Posted Date: 7 March 2024

doi: 10.20944/preprints202403.0385.v1

Keywords: artificial intelligence; ChatGPT; ovarian carcinoma; guidelines

Article

Consistency and Quality of ChatGPT Responses Compared to Clinical Guidelines for Ovarian Cancer: A Delphi Approach

Dario Piazza [1], Federica Martorana [2], Annabella Curaba [1], Daniela Sambataro [3], Maria Rosaria Valerio [4], Alberto Firenze [5], Basilio Pecorino [6,7], Paolo Scollo [6,7], Vito Chiantera [8], Giuseppe Scibilia [9], Paolo Vigneri [10,11], Vittorio Gebbia [1,12,*] and Giuseppina Scandurra [13]

[1] Medical Oncology Unit, Casa di Cura Torina, Palermo, Italy
[2] Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy
[3] Medical Oncology Unit, Ospedale Umberto I, Enna, Italy
[4] Medical Oncology Unit, Policlinico P. Giaccone, University of Palermo, Palermo, Italy
[5] Occupational Health Section, Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialties, University of Palermo, Palermo, Italy
[6] Gynecology Unit, Ospedale Cannizzaro, Catania, Italy
[7] Gynecology, Faculty of Medicine and Surgery, University of Enna Kore, Enna Italy
[8] Gynecology, University of Palermo, Palermo, Italy
[9] Gynecology Unit, Ospedale Paternò Arezzo, Ragusa, Italy
[10] Medical Oncology, University of Catania, Catania, Italy
[11] Medical Oncology, Istituto Clinico Humanitas, Catania, Italy.
[12] Medical Oncology, Faculty of Medicine and Surgery, University of Enna Kore, Enna Italy
[13] Medical Oncology Unit, Ospedale Cannizzaro, Catania, Italy
* Correspondence: vittorio.gebbia@unikore.it; Tel.: +39-330696205

Abstract: Introduction: In recent years, generative Artificial Intelligence models, such as ChatGPT, have been increasingly utilized in healthcare. Despite acknowledging the high potential of AI models in terms of quick access to sources and formulating a response to a clinical question, the results obtained using these models still require validation through comparison with established clinical guidelines. This study compares the responses of the AI model to eight clinical questions with the Italian Association of Medical Oncology (AIOM) guidelines for ovarian cancer. **Materials and Methods:** The authors used the Delphi method to evaluate responses from ChatGPT and the AIOM guidelines. An expert panel of healthcare professionals assessed responses based on clarity, consistency, comprehensiveness, usability, and quality using a 5-point Likert scale. The GRADE methodology assessed the evidence quality and the recommendations' strength. **Results:** A survey involving 14 physicians revealed that the AIOM guidelines consistently scored higher averages compared to the AI models with a statistically significant difference. Post-hoc tests showed that AIOM guidelines significantly differed from all AI models, with no significant difference among the AI models. **Conclusions:** While AI models can provide rapid responses, they must match established clinical guidelines regarding clarity, consistency, comprehensiveness, usability, and quality. These findings underscore the importance of relying on expert-developed guidelines in clinical decision-making and highlight potential areas for AI model improvement.

Keywords: artificial intelligence; ChatGPT; ovarian carcinoma; guidelines

Introduction

Ovarian cancer (OC) is a significant worldwide health concern, with high mortality rates and few therapeutic options. OC is the fifth most common malignancy, ranking fourth among cancer-related deaths in women in the USA, and is the leading cause of gynecologic cancer-related death in

 doi:10.20944/preprints202403.0385.v1

the Western world [1]. In Italy, OC ranks tenth among all female cancers (3%), with approximately 5,200 new diagnoses in 2020, 3,200 deaths in 2021, and a 5-year net survival rate of 43% from the time of diagnosis [2].

International and national guidelines have been developed to endow evidence-based recommendations for the diagnosis, treatment, and follow-up of OC cancer patients. The National Comprehensive Cancer Network (NCCN) and the European Society of Medical Oncology (ESMO) guidelines elaborate and constantly update evidence-based recommendations for managing OC [3,4]. The Italian Association of Medical Oncology (AIOM) has also developed guidelines to provide evidence-based recommendations for OC patients' diagnosis, treatment, and follow-up [5].

ChatGPT (Generative Pre-trained Transformer) is a natural language artificial intelligence model developed by OpenAI based on the transformer architecture [6,7]. The first version i.e., GPT-3.5, was a potent model capable of understanding context and generating highly accurate responses. However, with the introduction of GPT-4, the model's capabilities have been significantly enhanced. GPT-4 has substantially increased model size and the number of parameters, making it more accurate in understanding context and capable of generating creative and coherent responses [8]. Moreover, thanks to improved training and the algorithm, GPT-4 has become more efficient in handling user queries, providing better natural language interpretation, even in complex situations. Despite being based on the same architecture as its predecessor, GPT-4 represents a significant step forward in artificial intelligence and natural language processing [9]. Given its capabilities, ChatGTP may have significant applications in several medical fields, including oncology. It could provide immediate responses to frequently asked questions, freeing time for medical professionals to focus on more complex tasks [10]. In oncology, GPT-4 could interpret patient data, helping doctors understand symptom patterns and trends or treatment responses [10,11].

Furthermore, GPT-4 could assist health professionals in providing personalized reports on medical status, treatment options, and potential side effects to patients [10–13]. This tool could enhance patient understanding and decision-making, promoting patient-centered care [14]. However, using AI in patient care should always be coupled with appropriate ethical considerations, including privacy, accuracy, and transparency [14,15]. Additionally, using such tools in the medical field raises doubts and concerns about the accuracy and reliability of the information provided.

We conducted a study to investigate the consistency and quality of responses generated by OpenAI's language model – ChatGPT - to clinical queries concerning OC, comparing results to the Italian guidelines. The evaluations focused on the clarity of recommendations, relevance of evidence presented, comprehensiveness of the information, and applicability in clinical practice. The study provides a comparison of AI-generated clinical advice with established oncology guidelines, thereby assessing the utility and validity of AI in facilitating healthcare.

Materials and Methods

Study design. In this study, we employed a rigorous approach to evaluate the consistency and quality of responses generated by OpenAI's ChatGPT to clinical queries related to OC treatment, compared to the guidelines published by the Italian Association of Medical Oncology (AIOM). The latter guidelines offered responses to eight clinical queries, and these identical queries were posed to two versions of the ChatGPT model, 3.5 and 4. An additional set of queries was presented to ChatGPT model 4, with an optimally constructed prompt designed to elicit structured responses. Three rounds of questioning were conducted for each model and query type, replicating the real-world variability in question presentation (Table 1). The responses from these models were then compared with those outlined in the AIOM guidelines. These comparisons were carried out quantitatively by comparing the direct similarities and differences in the given advice and qualitatively by assessing the clarity, consistency, comprehensiveness, and usability of the information provided by the AI models (Figure 1). To perform this evaluation, we applied the Delphi method, which involves a panel of experts participating in iterative rounds of evaluation until a consensus is reached [16]. Our expert panel comprised diverse healthcare professionals and researchers, including oncologists, gynecologists,

 doi:10.20944/preprints202403.0385.v1

pathologists, radiologists, and evidence-based medicine experts. The experts assessed the AI responses using a 5-point Likert scale based on predefined criteria.

Furthermore, we used the GRADE methodology to assess the quality of evidence and the strength of the recommendations given by the AI models. GRADE is a systematic approach that helps to assess the quality of evidence in studies and the strength of health care recommendations [17]. This methodology was used to assess both the responses given by ChatGPT and the responses provided by the AIOM guidelines. Due to the nature of the study approval by the Ethics Committee or Informed Consent Statement was waived according to the Italian law.

Statistics. One-way ANOVA test to compare the mean scores of results. A Tukey post-hoc test was carried out to identify which groups significantly differed.

Table 1. The format of how questions are proposed concerning the model used.

Model	Prompt
ChatGPT-3.5	[Clinical Question #]* (as proposed from source document)
ChatGPT-4	[Clinical Question #]* (as proposed from source document)
ChatGPT-4	Act as an Italian multidisciplinary oncology group. We ask a question according to the PICO method. Reply extensively based on national and international guidelines and current evidence, indicate the limitations of the evidence, and indicate the ratio of benefits to harms. Also, provide answers with a formal GRADE approach indicating the overall quality of evidence and strength of recommendation.§ [Clinical Question #]*

*Questions asked in the same language as in the source document. § prompt structured and proposed in the same language as in source document.

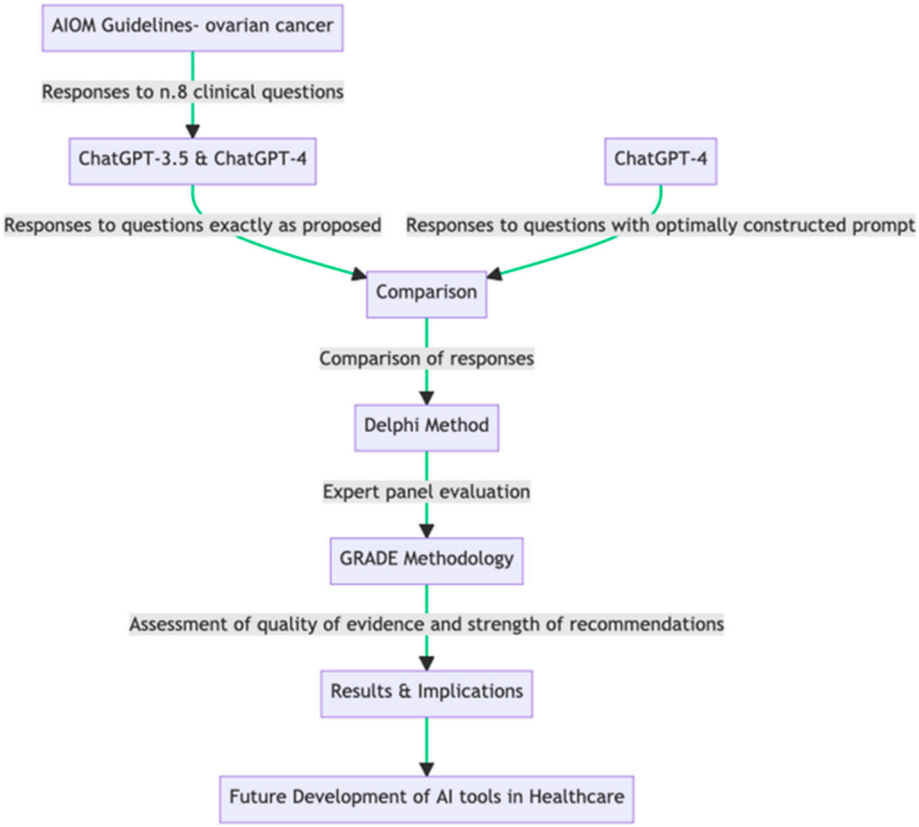

Figure 1. Flowchart of the study design.

 doi:10.20944/preprints202403.0385.v1

Results

The survey was conducted among 14 physicians, seven oncologists, and seven gynecologists, who thoroughly evaluated the responses to the eight clinical questions in five main domains: clarity, consistency, comprehensiveness, usability, and quality. Table 2 shows the evaluation questions grouped by domains and their average values. Figure 2 shows the average scores assigned in each domain for each model. The AIOM guidelines consistently scored higher averages compared to the artificial intelligence models.

We performed a one-way ANOVA test to compare the mean scores across the AI models and the guidelines. The test showed a significant difference between groups ($F = 21.66$, $p < 0.00001$), suggesting that at least one of the groups differed significantly from the others. Following the ANOVA results, a Tukey post-hoc test was carried out to identify which groups significantly differed. The test showed that the AIOM guidelines significantly differed from all other groups (ChatGPT-3.5, ChatGPT-4, and ChatGPT-4 with a prompt), with an adjusted p-value for multiple comparisons below 0.05. Among the artificial intelligence models, there was no significant difference between ChatGPT-3.5 and ChatGPT-4 or between ChatGPT-4 and ChatGPT-4 with a prompt (Table 3).

Table 2. Survey assessment questions and average results.

Domains	Questions	Mean	CI (± 95%)
clarity	How do you think the guideline expresses its recommendations?	4.28	0.14
	How does the ChatGPT-3.5 model's response to the clinical question express its recommendations?	1.23	0.12
	How does the ChatGPT-4 model's response to the clinical question express its recommendations?	2.23	0.21
	How does the prompted ChatGPT-4 model's response to the clinical question express its recommendations?	3.31	0.21
relevance	How relevant is the evidence in the guideline for the recommendations?	4.35	0.15
	How relevant is the evidence presented in the ChatGPT-3.5 model's response to the clinical question for the recommendations made?	1.36	0.09
	How relevant is the evidence presented in the ChatGPT-4 model's response to the clinical question for the recommendations made?	2.25	0.24
	How relevant is the evidence presented in the prompted ChatGPT-4 model's response to the clinical question for the recommendations made?	3.15	0.24
comprehensiveness	How comprehensive are the guidelines in addressing the topic?	4.53	0.13
	How comprehensive is the ChatGPT-3.5 model's response to the clinical question in addressing the topic?	1.11	0.06
	How comprehensively does the ChatGPT-4 model's response to the clinical question is in addressing the topic?	2.13	0.22
	How comprehensive is the prompted ChatGPT-4 model's response to the clinical question in addressing the topic?	2.95	0.23
applicability	How applicable is the guide to clinical practice?	4.28	0.14
	How applicable is the ChatGPT-3.5 model's response to the clinical question to clinical practice?	1.23	0.12
	How applicable is the ChatGPT-4 model's response to the clinical question to clinical practice?	2.26	0.23
	How applicable is the prompted ChatGPT-4 model's response to the clinical question to clinical practice?	2.82	0.27
quality	According to the GRADE approach, how would you rate the strength of the recommendations and the quality of the evidence presented in the guideline? 2.3 0.16 According to the GRADE approach		

 doi:10.20944/preprints202403.0385.v1

	How would you rate the recommendations' strength and the evidence's quality presented in the ChatGPT-3.5 model's response? According to the GRADE approach	1.88	0.12
	How would you rate the recommendations' strength and the evidence's quality presented in the ChatGPT-4 model's response? According to the GRADE approach	2.49	0.26
	How would you rate the recommendations' strength and the evidence's quality presented in the prompted ChatGPT-4 model's response?	2.38	0.26

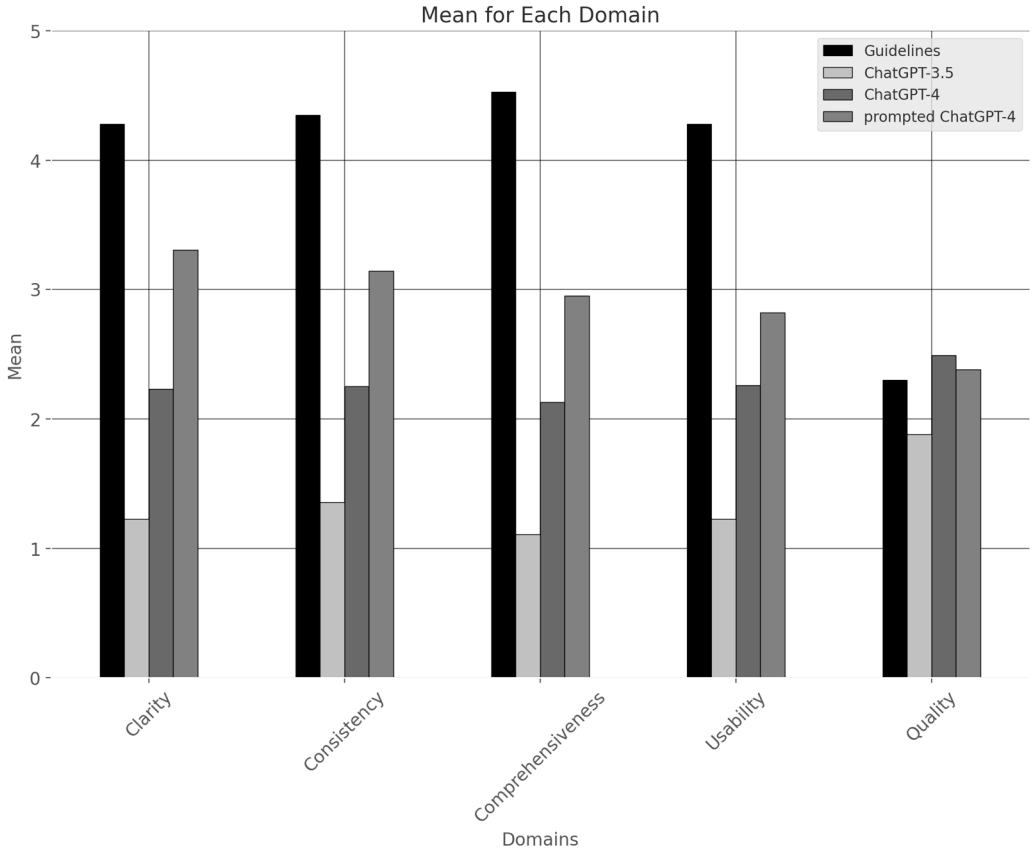

Figure 2. Average scores assigned in each domain for each model.

Table 3. Tukey Post-hoc Test Results.

Domain 1	Domain 2	Mean Difference	Adjusted p-value	Lower Bound	Upper Bound	Reject Null Hypothesis
ChatGPT-3.5	ChatGPT-4	0.91	0.0618	-0.037	1.857	False
ChatGPT-3.5	Guidelines	2.586	0.001	1.639	3.533	True
ChatGPT-3.5	Prompted ChatGPT-4	1.56	0.0012	0.613	2.507	True
ChatGPT-4	Guidelines	1.676	0.001	0.729	2.623	True
ChatGPT-4	Prompted ChatGPT-4	0.65	0.242	-0.297	1.597	False
Guidelines	Prompted ChatGPT-4	-1.026	0.0314	-1.973	-0.079	True

 doi:10.20944/preprints202403.0385.v1

Discussion

Recently, there has been increasing interest in incorporating AI into healthcare education, research, and clinical practice. One AI-based tool that has gained traction is ChatGPT, a large language model that can provide professional support to patients, medical professionals, researchers, and educators. Several studies investigated the potential applications and limitations of ChatGPT in medicine. Yeo et al. assessed the performance of ChatGPT in answering queries concerning cirrhosis and hepatocellular carcinoma (HCC). Their study showed that ChatGPT regurgitated extensive knowledge of cirrhosis and HCC, but only small proportions were labeled comprehensive [18]. Similarly, another study evaluated the feasibility of ChatGPT in healthcare and analyzed several clinical and research scenarios [19]. Results indicated that while AI-based language models like ChatGPT have impressive capabilities, they may perform poorly in real-world settings, especially medicine, where high-level and complex thinking is necessary.

Recently, the scientific community has raised ethical concerns about using ChatGPT to write scientific articles and other scientific outputs. A recent systematic review was conducted to investigate the utility of ChatGPT in healthcare [20]. They retrieved 60 records that examined ChatGPT in the context of healthcare education, research, or practice. Their findings highlighted the benefits of ChatGPT, which included improved scientific writing, enhanced research equity and versatility, utility in healthcare research, and saving time to focus on experimental design and downstream analysis. However, the authors also emphasized the need to address valid concerns associated with ChatGPT in healthcare, such as data protection and the potential negative impacts on physician-patient relationships. Kim et al. discussed the current acceptability of ChatGPT and large language model (LLM) chatbots in academic medicine and proposed guidelines for their utilization [21]. They identified the potential benefits of using ChatGPT and LLM chatbots, such as increased access to healthcare information and support. They also highlighted the challenges that need to be addressed, such as data privacy and the impact on medical professionalism.

The use of ChatGPT in oncology care has gained considerable attention in recent months. In an observational study, ChatGPT was evaluated for its ability to identify guideline-based treatments for advanced solid tumors [22]. The study demonstrated that ChatGPT can elaborate appropriate therapeutic choices for new diagnoses of advanced solid malignancies through standardized prompts. The valid therapy quotient (VTQ) was introduced as a ratio of medications listed by ChatGPT to those suggested in the NCCN guidelines, revealing that ChatGPT correctly identified guideline-based treatments in about 70% of cases. In a recent editorial, Kothari revealed that ChatGPT had quickly attracted many active users due to its extraordinary ability to understand and generate human-like language [23]. In addition, ChatGPT has generated various types of content, including scholarly work, exam questions, and discharge summaries. Hamilton et al. evaluated the clinical relevance and accuracy of ChatGPT-generated Next-generation sequencing (NGS) reports with first-line treatment recommendations for NSCLC patients with targetable driver oncogenes [24]. The study concluded that ChatGPT-generated reports were contextually accurate and clinically relevant.

Although the potential benefits of ChatGPT in healthcare are significant, researchers continue to investigate the technology's integration and effectiveness across diverse fields. Cheng et al. discussed how the integration of ChatGPT can enable a new era of surgical oncology [25], while Ebrahimi et al. evaluated whether a natural language processing tool like ChatGPT would be trustworthy for radiation oncology use [26]. A study by Haemmerli et al. evaluated the ChatGPT recommendations for glioma management by a panel of CNS tumor experts [27]. CNS tumor board experts assessed ChatGPT and found that while it performed severely in diagnosing glioma kinds, it performed well in recommending adjuvant treatments. Despite its inability to match the accuracy of expert judgment, ChatGPT shows promise as an additional tool when used in conjunction with a human in the loop. Huang et al. assessed the potential of ChatGPT-4 for AI-assisted medical education and decision-making in radiation oncology [28]. While noting ChatGPT-4's limits in some areas, the study showed the technology's potential for clinical decision support and medical education of the public and cancer patients. However, because of the possibility of hallucinations, confirming the authenticity of the content produced by models like ChatGPT is crucial.

 doi:10.20944/preprints202403.0385.v1

This paper is the first report comparing ChatGPT outputs to clinical guideline recommendations in oncology. Our study assessed the responses to eight clinical questions provided by the AIOM guidelines on ovarian cancer and three generative artificial intelligence models, ChatGPT-3.5, ChatGPT-4, and ChatGPT-4, with a structured prompt. A multidisciplinary team evaluated the responses across five main domains: clarity, consistency, comprehensiveness, usability, and quality, using a 5-point Likert scale. Results scores across the domains indicate that the AIOM guidelines consistently achieved higher mean scores than the generative artificial intelligence models. This report suggests that the physicians surveyed found the responses provided by the AIOM guidelines to be more precise, relevant, comprehensive, applicable, and of higher quality than those provided by the AI models. Medical experts developed medical-scientific guidelines based on extensive research and consensus among the medical community. At the same time, AI models, despite their advanced capabilities, may still need more subtlety and depth of understanding inherent in human expertise. The results of the one-way ANOVA test further support this observation, revealing a significant difference between the groups. These data suggest a statistically significant variation in the mean scores between at least one group pair, reinforcing the conclusion that the AIOM guidelines were evaluated more favorably. The Tukey post-hoc test, conducted to identify which specific groups differed significantly, indicated that the AIOM guidelines significantly differed from all other groups. Interestingly, there were no significant differences among the artificial intelligence models, suggesting that adding a structured prompt in ChatGPT-4 did not significantly enhance its performance in this context.

Limitations

While ChatGPT and other AI-based tools hold promise in healthcare education, research, and practice, it is essential to recognize and address their limitations and potential ethical concerns. Correct information and users' education on the appropriate use and potential pitfalls of AI-based language models are crucial to ensure that they are used to optimize their benefits while minimizing any potential harm.

Conclusions

The future of new generative artificial intelligence tools in the medical field is promising, potentially improving the quality and consistency of medical information provided to patients. However, assuring that the information provided is accurate and reliable is essential. Using international and national guidelines can help ensure the quality and consistency of medical information provided to patients. Further research is needed to evaluate the effectiveness of new generative artificial intelligence tools in improving patient outcomes and to address concerns about the accuracy and reliability of the information provided. In conclusion, ovarian cancer is a significant health concern worldwide, with high mortality rates and limited treatment options. New generative artificial intelligence tools, such as OpenAI's ChatGPT, have emerged as a potential solution to improve the quality and consistency of medical information provided to patients. However, using such tools in the medical field generates concerns about the accuracy and reliability of the information provided. International and national guidelines have been developed to ensure the quality and consistency of medical information provided to patients. The future of new generative artificial intelligence tools in the medical field is promising, but further research is needed to evaluate their effectiveness and address concerns about their accuracy and reliability. In conclusion, while AI models such as ChatGPT can provide rapid responses to clinical questions, our study suggests they must match up to established clinical guidelines regarding clarity, relevance, comprehensiveness, applicability, and quality, as oncologists and gynecologists perceive. These observations underscore the importance of relying on expert-developed guidelines in clinical decision-making while highlighting potential areas for improvement in AI models for clinical use. Tracking how these comparisons may change over time will be interesting as AI evolves.

Author Contributions: Conceptualization, D.P., G.S. and V.G.; Data curation, F.M., M.R.V., A.C., B.P. and G.S.; Formal analysis, G.S., P.V. and D.P.; Methodology, D.P., P.S., and V.G; Resources, F.M., D.P. AC; Supervision, D.P., V.C., and G.S.; Validation, All Authors.; Writing, D.P. and V.G.; Visualization, All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement, Ethics Committee approval, and informed Consent Statement: Not necessary.

Data Availability Statement: The data presented in this study are available from the corresponding author upon request.

Acknowledgments: The authors thank all health professionals who voluntarily participated to the study.

Conflicts of Interest: All authors declare no conflicts of interest.

References

1. Armstrong, D.K.; Alvarez, R.D.; Bakkum-Gamez, J.N.; Barroilhet, L.; Behbakht, K.; Berchuck, A.; Chen, L.; Cristea, M.; DeRosa, M.; Eisenhauer, E.L.; et al. Ovarian Cancer, Version 2.2020, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network* **2021**, *19*, 191–226, doi:10.6004/jnccn.2021.0007.
2. I Numeri Del Cancro 2023 | Associazione Italiana Registri Tumori Available online: https://www.registri-tumori.it/cms/notizie/i-numeri-del-cancro-2023 (accessed on 26 February 2024).
3. National Comprehensive Cancer Network - Home Available online: https://www.nccn.org (accessed on 7 February 2024).
4. Colombo, N.; Sessa, C.; Du Bois, A.; Ledermann, J.; McCluggage, W.G.; McNeish, I.; Morice, P.; Pignata, S.; Ray-Coquard, I.; Vergote, I.; et al. ESMO–ESGO Consensus Conference Recommendations on Ovarian Cancer: Pathology and Molecular Biology, Early and Advanced Stages, Borderline Tumours and Recurrent Disease. *Annals of Oncology* **2019**, *30*, 672–705, doi:10.1093/annonc/mdz062.
5. LINEE GUIDA CARCINOMA DELL'OVAIO Available online: https://www.aiom.it/linee-guida-aiom-2021-carcinoma-dellovaio/ (accessed on 7 February 2024).
6. OpenAI Available online: https://openai.com/ (accessed on 7 February 2024).
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2017; Vol. 30.
8. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2020; Vol. 33, pp. 1877–1901.
9. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners.
10. Xu, L.; Sanders, L.; Li, K.; Chow, J.C.L. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. *JMIR Cancer* **2021**, *7*, e27850, doi:10.2196/27850.
11. Papachristou, N.; Kotronoulas, G.; Dikaios, N.; Allison, S.J.; Eleftherochorinou, H.; Rai, T.; Kunz, H.; Barnaghi, P.; Miaskowski, C.; Bamidis, P.D. Digital Transformation of Cancer Care in the Era of Big Data, Artificial Intelligence and Data-Driven Interventions: Navigating the Field. *Seminars in Oncology Nursing* **2023**, *39*, 151433, doi:10.1016/j.soncn.2023.151433.
12. Taber, P.; Armin, J.S.; Orozco, G.; Del Fiol, G.; Erdrich, J.; Kawamoto, K.; Israni, S.T. Artificial Intelligence and Cancer Control: Toward Prioritizing Justice, Equity, Diversity, and Inclusion (JEDI) in Emerging Decision Support Technologies. *Curr Oncol Rep* **2023**, *25*, 387–424, doi:10.1007/s11912-023-01376-7.
13. Tawfik, E.; Ghallab, E.; Moustafa, A. A Nurse versus a Chatbot – the Effect of an Empowerment Program on Chemotherapy-Related Side Effects and the Self-Care Behaviors of Women Living with Breast Cancer: A Randomized Controlled Trial. *BMC Nurs* **2023**, *22*, 102, doi:10.1186/s12912-023-01243-7.
14. Xue, V.W.; Lei, P.; Cho, W.C. The Potential Impact of ChatGPT in Clinical and Translational Medicine. *Clinical & Translational Med* **2023**, *13*, e1216, doi:10.1002/ctm2.1216.
15. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595, doi:10.3389/frai.2023.1169595.
16. Taylor, E. We Agree, Don't We? The Delphi Method for Health Environments Research. *HERD* **2020**, *13*, 11–23, doi:10.1177/1937586719887709.
17. Guyatt, G.H.; Oxman, A.D.; Vist, G.E.; Kunz, R.; Falck-Ytter, Y.; Alonso-Coello, P.; Schünemann, H.J. GRADE: An Emerging Consensus on Rating Quality of Evidence and Strength of Recommendations. *BMJ* **2008**, *336*, 924–926, doi:10.1136/bmj.39489.470347.AD.

18. Yeo, Y.H.; Samaan, J.S.; Ng, W.H.; Ting, P.-S.; Trivedi, H.; Vipani, A.; Ayoub, W.; Yang, J.D.; Liran, O.; Spiegel, B.; et al. Assessing the Performance of ChatGPT in Answering Questions Regarding Cirrhosis and Hepatocellular Carcinoma. *Clin Mol Hepatol* **2023**, doi:10.3350/cmh.2023.0089.
19. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst* **2023**, *47*, 33, doi:10.1007/s10916-023-01925-4.
20. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887, doi:10.3390/healthcare11060887.
21. Kim, J.K.; Chua, M.; Rickard, M.; Lorenzo, A. ChatGPT and Large Language Model (LLM) Chatbots: The Current State of Acceptability and a Proposal for Guidelines on Utilization in Academic Medicine. *Journal of Pediatric Urology* **2023**, S1477513123002243, doi:10.1016/j.jpurol.2023.05.018.
22. Schulte, B. Capacity of ChatGPT to Identify Guideline-Based Treatments for Advanced Solid Tumors. *Cureus* **2023**, doi:10.7759/cureus.37938.
23. Kothari, A.N. ChatGPT, Large Language Models, and Generative AI as Future Augments of Surgical Cancer Care. *Ann Surg Oncol* **2023**, *30*, 3174–3176, doi:10.1245/s10434-023-13442-2.
24. Hamilton, Z.; Naffakh, N.; Reizine, N.M.; Weinberg, F.; Jain, S.; Gadi, V.K.; Bun, C.; Nguyen, R.H.-T. Relevance and Accuracy of ChatGPT-Generated NGS Reports with Treatment Recommendations for Oncogene-Driven NSCLC. *JCO* **2023**, *41*, 1555–1555, doi:10.1200/JCO.2023.41.16_suppl.1555.
25. Cheng, K.; Wu, H.; Li, C. ChatGPT/GPT-4: Enabling a New Era of Surgical Oncology. *International Journal of Surgery* **2023**, *Publish Ahead of Print*, doi:10.1097/JS9.0000000000000451.
26. Ebrahimi, B.; Howard, A.; Carlson, D.J.; Al-Hallaq, H. ChatGPT: Can a Natural Language Processing Tool Be Trusted for Radiation Oncology Use? *International Journal of Radiation Oncology*Biology*Physics* **2023**, S0360301623003541, doi:10.1016/j.ijrobp.2023.03.075.
27. Haemmerli, J.; Sveikata, L.; Nouri, A.; May, A.; Egervari, K.; Freyschlag, C.; Lobrinus, J.A.; Migliorini, D.; Momjian, S.; Sanda, N.; et al. *ChatGPT in Glioma Patient Adjuvant Therapy Decision Making: Ready to Assume the Role of a Doctor in the Tumour Board?*; Neurology, 2023;
28. Huang, Y.; Gomaa, A.; Semrau, S.; Haderlein, M.; Lettmaier, S.; Weissmann, T.; Grigo, J.; Tkhayat, H.B.; Frey, B.; Gaipl, U.; et al. Benchmarking ChatGPT-4 on a Radiation Oncology in-Training Exam and Red Journal Gray Zone Cases: Potentials and Challenges for Ai-Assisted Medical Education and Decision Making in Radiation Oncology. *Front. Oncol.* **2023**, *13*, 1265024, doi:10.3389/fonc.2023.1265024.