Archives of Gynecology and Obstetrics (2024) 309:1543–1549
https://doi.org/10.1007/s00404-023-07272-6

GYNECOLOGIC ONCOLOGY

Will I soon be out of my job? Quality and guideline conformity of ChatGPT therapy suggestions to patient inquiries with gynecologic symptoms in a palliative setting

Eva-Marie Braun[1] · Ingolf Juhasz-Böss[2] · Erich-Franz Solomayer[3] · Daniel Truhn[4] · Christiane Keller[5] · Vanessa Heinrich[6] · Benedikt Johannes Braun[7]

Received: 13 August 2023 / Accepted: 15 October 2023 / Published online: 17 November 2023

Abstract

Purpose The market and application possibilities for artificial intelligence are currently growing at high speed and are increasingly finding their way into gynecology. While the medical side is highly represented in the current literature, the patient's perspective is still lagging behind. Therefore, the aim of this study was to evaluate the recommendations of ChatGPT regarding patient inquiries about the possible therapy of gynecological leading symptoms in a palliative situation by experts.

Methods Case vignettes were constructed for 10 common concomitant symptoms in gynecologic oncology tumors in a palliative setting, and patient queries regarding therapy of these symptoms were generated as prompts for ChatGPT. Five experts in palliative care and gynecologic oncology evaluated the responses with respect to guideline adherence and applicability and identified advantages and disadvantages.

Results The overall rating of ChatGPT responses averaged 4.1 (5 = strongly agree; 1 = strongly disagree). The experts saw an average guideline conformity of the therapy recommendations with a value of 4.0. ChatGPT sometimes omits relevant therapies and does not provide an individual assessment of the suggested therapies, but does indicate that a physician consultation is additionally necessary.

Conclusions Language models, such as ChatGPT, can provide valid and largely guideline-compliant therapy recommendations in their freely available and thus in principle accessible version for our patients. For a complete therapy recommendation, an evaluation of the therapies, their individual adjustment as well as a filtering of possible wrong recommendations, a medical expert's opinion remains indispensable.

Keywords Language model · Palliative care · Gynecologic oncology · Artificial intelligence

✉ Eva-Marie Braun
braun.em@web.de

[1] Center for Integrative Oncology, Die Filderklinik, Im Haberschlai 7, 70794 Filderstadt-Bonlanden, Germany

[2] Department of Gynecology, University Medical Center Freiburg, Hugstetter Straße 55, 79106 Freiburg, Germany

[3] Department of Gynecology, Obstetrics and Reproductive Medicine, Saarland University Hospital, Kirrberger Straße, Building 9, 66421 Homburg, Germany

[4] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Pauwelsstraße 30, 52074 Aachen, Germany

[5] Center for Palliative Medicine and Pediatric Pain Therapy, Saarland University Hospital, Kirrberger Straße, Building 69, 66421 Homburg, Germany

[6] Department of Radiation Oncology, University Hospital Tübingen, Crona Kliniken, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany

[7] Department of Trauma and Reconstructive Surgery at the Eberhard Karls University Tübingen, BG Unfallklinik Tübingen, Schnarrenbergstrasse 95, 72076 Tübingen, Germany

What does this study add to the clinical work:

Large Language Models can provide valid and largely guideline-compliant therapy recommendations. We show when, why and where a medical expert's evaluation and filtering is nonetheless indispensable.

Introduction

Artificial intelligence and large language models (LLMs) are increasingly being used both in general and medicine in particular. The LLM GPT (Generative Pre-Trained Transformer) in its online form, the chatbot ChatGPT, made publicly available by the developer OpenAI, is enjoying great popularity. Within 2 months of its launch, it had over 100 million users, with over 1 billion page views per month as of February 2023 [1]. ChatGPT gained greater attention in the general context of medicine when, applied to the U.S. medical school final exam, the software was able to answer multiple choice exam questions above a passing grade [2].

ChatGPT has also been able to demonstrate its high level of knowledge in gynecological testing situations in a virtual Objective Structured Clinical Examination (OSCE) [3]. Other fields of application in our field have been to demonstrate the capabilities and limitations of the system with respect to scientific writing [4, 5]. In the current medical literature, however, the essential aspect regarding the clinical potential of LLMs is their applicability to diagnostic and therapeutic problems [6]. In gynecology, too, the use of LLMs is increasingly being discussed [7]. It's ability to answer clinical questions, as well as its supportive use in a multidisciplinary tumor board with breast cancer patients are just 2 recent examples of potential use cases in our field [8, 9]. The authors of the article point out that a conscious use of these systems is necessary to exploit advantages appropriately and to avoid wrong answers. This is especially important when patients work with these systems without sufficient contextual knowledge to correctly interpret a ChatGPT answer.

Currently, an increasing number of female patients are seeking advice on disease symptoms through ChatGPT. The authors are not aware of any statistics on this, but their own experience shows that more and more patients are specifically using this option to obtain a clearly formulated answer to a specific question from an AI (artificial intelligence) system rather than a range of different information, as is the case with classic online searches. Although initial analyses show that, depending on the type of question, the subject area of the question and the queried symptomatology, the answers can give a correct overview, there is no structured survey of the quality of these systems, especially with regard to gynecologic oncology symptomatology in a palliative situation.

Aim of this work was to evaluate the recommendations of the freely accessible version of ChatGPT regarding constructed patient inquiries about the possible therapy of gynecologic oncology symptoms in patients in a palliative treatment situation by experts and to classify them against the background of the current guideline standards. In addition, advantages and disadvantages of the technique will be discussed, in particular to better understand the response patterns depending on the questions wording.

Materials and methods

Short case vignettes were constructed for 10 common accompanying symptoms in gynecologic oncology tumor patients in a palliative treatment situation (Table 1). From this, one prompt was formulated per case vignette, which was constructed according to the following pattern: "I am an (age) year old patient with a (tumor diagnosis) (with metastases) with a symptom in a palliative treatment situation. What therapy is available for my (symptom)?". The search history was cleared after every query. Chat GPT based on GPT 3.5 was used in the version dated March 23, 2023 and the query was performed on 04/16/2023. The prompts were entered in the above structure and the given answers were transferred to a Word document for the experts to assess. In total, the prompts and answer texts were submitted to 5 experts from the fields of gynecologic oncology ($n=3$) and palliative care ($n=2$), each with more than 10 years of professional experience, for evaluation. A general evaluation of the treatment proposal (Likert scale 5 = agree; 1 = disagree), the assessment of the evidence of the treatment proposal (Likert scale 5 = present; 1 = not present), and the applicability of the proposal (Likert scale 5 = completely applicable; 1 = not applicable at all) were queried. In addition, the evaluating experts were allowed to give free-text answers on the pros and cons of the treatment recommendations. The evaluation of the experts was processed as a numerical descriptive evaluation and the free text comments were included in the discussion (Table 1). No actual patient data were used for this work. All experts consented to the publication of their answers.

Results

The overall rating across all case vignettes averaged 4.1 (range 3–5). Guideline conformity of all responses was rated an average of 4.0 (range 2–5), while applicability was 3.3 (range 2–5). As part of the answering of all questions, ChatGPT pointed out that the answer was an overview of the basic therapy options and that a visit to a physician was necessary for actual treatment.

ChatGPT responses generally followed a schematic approach. Suitable therapeutics were named for therapy and drug substance groups with exemplary active ingredient names and their mechanisms of action understandable to laypersons were explained. In addition, information on non-drug and integrative therapy options were also provided in varying degrees of detail (Fig. 1). Experts agreed that some recommendations could have been more specific. For numerous patients, ChatGPT provided more general advice divided into therapeutic treatment groups rather than tailoring treatment recommendations to the specific disease or individual patient, limiting their direct clinical usefulness. In addition, all therapeutic options were reported as being of equal value, without any evaluation for the patient in light of her own condition. In addition, individual therapeutic procedures were omitted. Table 1 contains the detailed treatment recommendations for each patient as well as the ratings of the experts (Table 1). All responses of the PIs are available as an appendix to this article (Appendix 1).

Discussion

The present work shows the basic potential of large language models with regard to a general, medical consultation of our patients. ChatGPT was also to provide usable and also predominantly guideline-compliant answers to the patients' questions in the freely available version. At the same time, however, there is still a need for expert consultation, especially with regard to completeness, the weighting of the individual therapy suggestions, and their individual evaluation for the specific case of illness of the inquiring patient, which is also indicated by the responses of the AI.

The answer to the first question on dyspnea in pulmonary and hepatic metastatic breast cancer impressively demonstrates the approach of the language model. The leading symptom dyspnea is understood and various therapy options are given in an overview style. Especially the listing of opiates as palliative relief of dyspnea shows that the language model basically understood the patient's problem and situation. However, in addition to other correct answers with bronchodilators, therapeutics are listed that are rarely an applicable therapy in a palliative situation, but at the same time-specific oncological systemic therapy for symptom control is not listed. These response patterns are also known from other surveys on the therapeutic quality of language models [10]: although the replies of the AI are not obviously incorrect, the leading symptomatology, in this case dyspnea, is determinant and triggers corresponding therapy recommendations for various differential diagnoses of dyspnea, which may also lie outside the palliative context of the query situation.

Table 1 Case vignettes

No	Age (in years)	Main diagnosis	Metastasis	Symptom	General evaluation	Evidence of treatment proposal	Applicability
1	34	Mamma carcinoma	Pulmonary, hepatic	Dyspnea	3,8	3,8	3,2
2	62	Cervical carcinoma	Lymphogenic, osseous	Continuous pain	4,6	4,4	3,4
3	66	Mamma carcinoma	Ossareous, hepatic	Acute Pain	3,8	3,6	3
4	74	Ovarian Carcinoma	Peritoneal	Constipation	4,2	4	3,4
5	68	Endometrial carcinoma	Lymphogen	Nausea	4,6	4,2	3,6
6	82	Ovarian Carcinoma	Hepatic, peritoneal	Vomiting	4	4,4	3,2
7	58	Uterine Sarcoma	Local	Fear	4	3,2	2,6
8	43	Mamma carcinoma	Osseous, hepatic	Fatigue	4,2	4,4	3,4
9	85	Vulvar carcinoma	Lymphogenic, local	Trouble sleeping through the night	4,4	4	4
10	68	Mamma carcinoma	Ossaeous, cerebral	Exulcerating tumor with bleeding	3,6	3,6	3,4

The table shows clinical data of the constructed patients and the experts' ratings. The rating for the evidence of the treatment proposal is based on the expert's critical appraisal

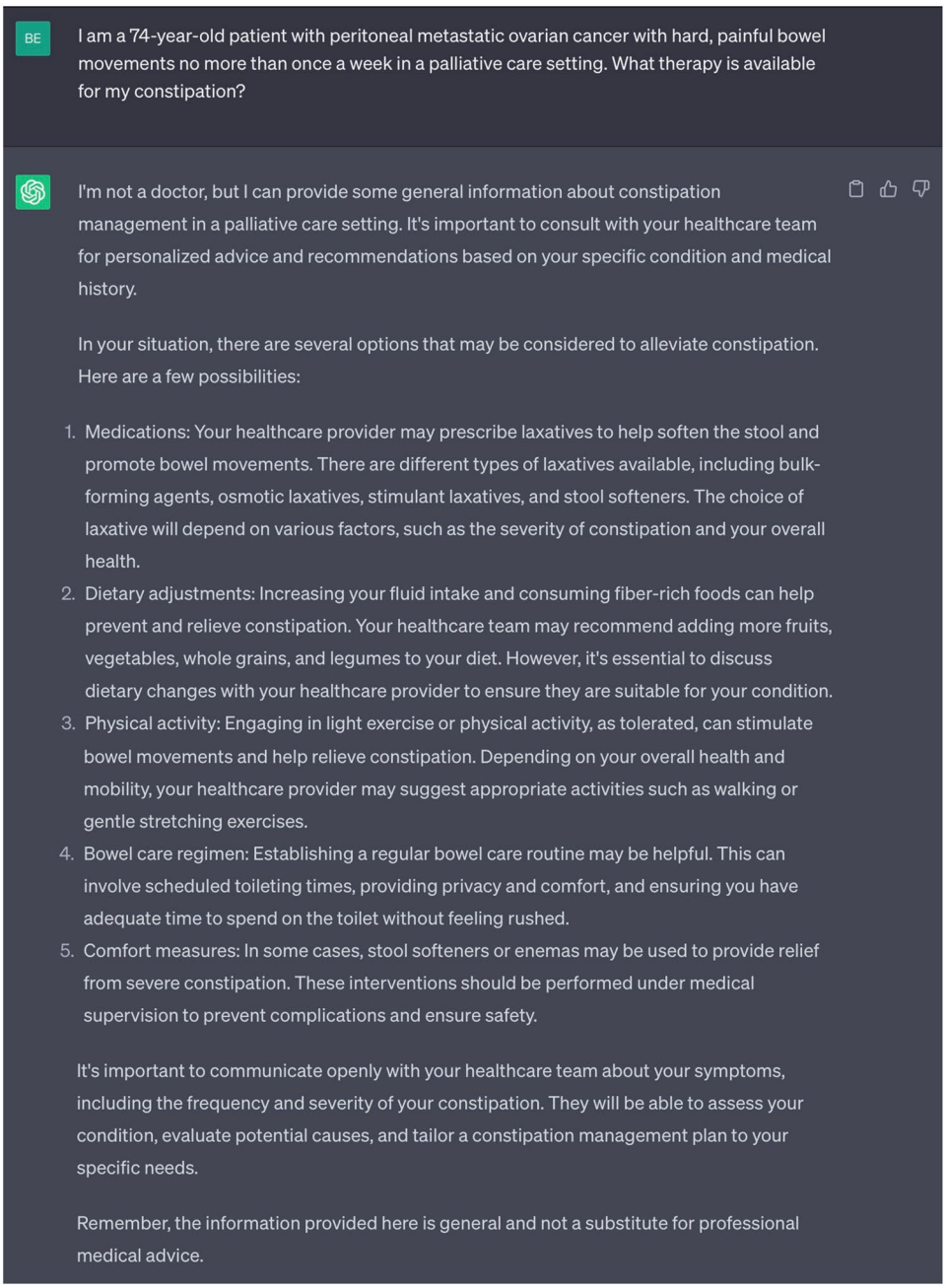

Fig. 1 Exemplary, English language answer of ChatGPT is shown

One reason for this lack of context consideration may be that language models, such as the one tested here, are so-called autoregressive models, which have only a limited amount of text as input length that can be meaningfully put into the model [11]. The models, therefore, largely lack the ability for the differentiated, medical, human comprehension of complex conditions, which must be evaluated in larger context, taking into account detailed information and based on informed reasoning [12, 13]. This issue is well-known and numerous research, as well as commercial, stakeholders are working to enable greater lengths of input to these models to improve contextual consideration [14].

The omission or only limited mention of more invasive therapeutic methods such as specific oncological therapy including chemotherapeutic agents, or radiation therapy in patients with osseous metastases is also known from preliminary work in other fields, in which, for example, invasive surgical measures are always placed in second place to conservative, less burdensome forms of therapy and reference is made to a medical consultation with regard to their

evaluation [15]. This seems to correspond to a deliberately cautious interpretation of the ChatGPT language model, in order not to prejudge medical treatment by a prematurely given evaluation. Another indication of this is that, at the time the survey was created, the language model integrated into the Microsoft Bing search engine merely acts as a classic online search engine for medical questions and does not provide a text response. Ultimately, the warning before each answer, in which the language model explicitly points out that it is merely the output of a language model with a suggestive character and not a medical answer, is also to be understood under this safety aspect in the case of medical questions.

The complete omission of recommendations, on the other hand, is potentially dangerous and sometimes withholds important information from patients and practitioners in their decision-making process regarding the choice of therapy. For example, in the case of constipation symptoms, the sometimes important therapy with peripheral opioid antagonists was not listed for these patients who often receive opioid therapy [16]. In the case of the patient with vomiting, potentially highly acute ileus symptoms are also not sufficiently taken into account, and thus the potential need for urgent care is disregarded by the language model. Especially in situations relevant to emergency medicine, these language models are, therefore, not yet fully usable and are not useful as sole therapy decision makers [17]. Rather, these systems are conceivable as support systems for medical decision-making processes, so-called decision support systems, but also as basic advice for patients prior to a planned medical consultation [18]. In this case, the overview-like presentation character of the answers merely represents a supportive entry into further clinical decision-making processes and enables patients to have an informed, pre-structured discussion without having to filter them from the multitude of (false) information available on the Internet, as is the case with classic search engine-based information [19, 20].

Furthermore, the generally polite way in which ChatGPT deals with patients' inquiries is striking. In addition to the warning that it is not medical advice, the language model usually expresses regret about the patient's situation. This polite, quasi human way of the language model has already been noticed in other studies, in which, among other things, the quality of the transmission of serious findings to patients in discharge letters was examined [21]. Here, the factor "humanity" of the answers was explicitly evaluated and in this study, it showed itself to be on a similarly high level as humanly created discharge letters. In addition, the phrasing 'I'm sorry to hear' suggests empathy in the reply, almost as though the LLM wants the answer recipient to feel understood. This field of medical ethics and AI and the effects of the answers of conversational chatbots on the patient is still fairly young, but of high interest to both the clinical, educational, as well as research oriented medical community and ethical frameworks are currently being developed [22, 23].

It should be noted that the present study deals with fictitious case vignettes and not with concrete clinical cases. This was necessary to avoid an ethically questionable forwarding of sensitive patient data to an AI system and is, against this background, common practice in the preparation these research works [17]. Furthermore, the case vignettes were evaluated not only by palliative experts, but also by gynecologists working in oncological surgery. This may explain subtle differences in the evaluation of the ChatGPT statements, depending on whether they were made against the background of a general palliative symptom control or in dependence of an entity-specific guideline taking into account also metastasized tumors in a palliative situation. Ultimately, however, this interdisciplinary assessment corresponds to the everyday clinical treatment of these diseases and thus allows, in our view, a clinically realistic assessment while accepting possible inhomogeneity of the numerical assessment.

Conclusion

Language models have in principle a high potential in the general counseling of our patients. The responses provide an overview of most, basically available treatment options of important core symptoms of palliative care, but are thereby rather to be understood as general advice, without the claim to absolute completeness, or detailed contextual consideration of an individual treatment situation. Against this background, ChatGPT also issues a corresponding warning to the person asking that additional medical advice must be obtained. This is particularly important for invasive therapies, and therapies, where the LLM is missing awareness for a potential emergency situation. As an outlook on the further use of language models, it can be stated that further technical development of AI will certainly make more precise and, above all, more context-appropriate answers possible in the future [24]. The use of these models by our patients can, therefore, be assumed to increase in the future. For our field, we should accompany the currently rapidly progressing evolution of these language models to be able to adequately react to inquiries of patients and their relatives without medical knowledge. From an ethical and legal perspective alone, we are still obliged to advise our patients on their treatment, irrespective of whether differential therapeutic planning with an AI system has been carried out by the patient or another practitioner [25]. However, knowing these systems, they can certainly support the counseling process of more informed patients in the future, as our answers show

the at times superficial but adequate quality of the answers. Against this background, it is not unexpected but reassuring that the direct applicability of the answers was rated lowest. The benefit of these systems currently and in the near future lies in the supportive consultation. The ultimate evaluation and selection of appropriate therapies lies with the physicians and patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00404-023-07272-6.

Author contributions All authors contributed to the study conception and design. EMB, EFS, IJB, CK and VH rated and commented the LLM therapy recommendations. Material preparation, data collection and analysis were performed by EMB and BJB. The first draft of the manuscript was written by EMB, BJB and DT and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received concerning the preparation of this manuscript.

Data availability All data are available within the manuscript and the appendix.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Only fictitious case vignettes were used. No actual patient data were analyzed, or used in any way. All experts consented to participate in the study and to the publication of their answers. No ethical committee vote was obtained.

Consent to participate All experts consented to participate in the study.

Consent to publish All experts consented to the publication of their answers.

References

1. Ruby D (2023) 57+ ChatGPT Statistics 2023 (Updated Data With Infographics). https://www.demandsage.com/chatgpt-statistics/. Accessed 21 May 2023
2. Kung TH, Cheatham M, Medenilla A et al (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS digital health 2:e0000198
3. Kemp MW, Logan SJ, Dimri PS et al (2023) ChatGPT outscored human candidates in a virtual objective structured clinical examination (OSCE) in obstetrics and gynecology. Am J Obstet Gynecol. https://doi.org/10.1016/j.ajog.2023.04.020
4. Sanchez-Ramos L, Lin L, Romero R (2023) Beware of references when using ChatGPT as a source of information to write scientific articles. Am J Obstet Gynecol. https://doi.org/10.1016/j.ajog.2023.04.004
5. Levin G, Meyer R, Kadoch E, Brezinov Y (2023) Identifying ChatGPT-written OBGYN abstracts using a simple tool. Am J Obstet Gynecol 5(6):100936. https://doi.org/10.1016/j.ajogmf.2023.100936
6. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med 388:1233–1239
7. Ray PP (2023) Bridging the gap: integrating ChatGPT into obstetrics and gynecology research—a call to action. Arch Gynecol Obstet. https://doi.org/10.1007/s00404-023-07129-y
8. Grünebaum A, Chervenak J, Pollet SL et al (2023) The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. https://doi.org/10.1016/j.ajog.2023.03.009
9. Lukac S, Dayan D, Fink V et al (2023) Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet. https://doi.org/10.1007/s00404-023-07130-5
10. Li J, Dada A, Kleesiek J, Egger J (2023) ChatGPT in healthcare: a taxonomy and systematic review. medRxiv 9(1):e001568. https://doi.org/10.1101/2023.03.30.23287899
11. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. OpenAI, San Francisco
12. Duong D, Solomon BD (2023) Analysis of large-language model versus human performance for genetics questions. medRxiv. https://doi.org/10.1038/s41431-023-01396-8
13. Potapenko I, Boberg-Ans LC, Hansen S et al (2023) Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol. https://doi.org/10.1111/aos.15661
14. Bulatov A, Kuratov Y, Burtsev MS (2023) Scaling transformer to 1M tokens and beyond with RMT. arXiv Prepr arXiv:230411062
15. Xie Y, Seth I, Hunter-Smith DJ et al (2023) Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 47(5):1985–1993. https://doi.org/10.1007/s00266-023-03338-7
16. Fernández-Montes A, de Velasco G, Aguín S et al (2021) Insights into the use of peripherally acting μ-opioid receptor antagonists (PAMORAs) in oncologic patients: from scientific evidence to real clinical practice. Curr Treat Options Oncol 22:1–19
17. Bushuven S, Bentele M, Bentele S et al (2023) ChatGPT, can you help me save my child's life?-Diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases–an in-silico analysis. Res Square. https://doi.org/10.2120/rs.3.rs-2910261/v1
18. Liu S, Wright AP, Patterson BL et al (2023) Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocad072
19. Corrales DM, Wells AE, Radecki Breitkopf C et al (2018) Internet use by gynecologic oncology patients and its relationship with anxiety. J Health Commun 23:299–305
20. Lawrentschuk N, Abouassaly R, Hewitt E et al (2016) Health information quality on the internet in gynecological oncology: a multilingual evaluation. Eur J Gynaecol Oncol 37:478–483
21. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS (2023) Using ChatGPT to write patient clinic letters. Lancet Digit Health 5:e179–e181
22. Maboloc CR (2023) Chat GPT: the need for an ethical framework to regulate its use in education. J Public Health. https://doi.org/10.1093/pubmed/fdad125
23. Fournier-Tombs E, McHardy J (2023) A medical ethics framework for conversational artificial intelligence. J Med Internet Res 25:e43068

24. King MR (2023) The future of AI in medicine: a perspective from a Chatbot. Ann Biomed Eng 51:291–295
25. Zhang J, Zhang Z (2023) Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 23:1–15