OXFORD

human reproduction

Human Reproduction, 2024, **39(3)**, 443–447
https://doi.org/10.1093/humrep/dead272
Advance Access Publication Date: January 10, 2024
Opinion

ChatGPT: a reliable fertility decision-making tool?

Kiri Beilby [1,*] and Karin Hammarberg [2]

[1]Education Program in Reproduction and Development, Department of Obstetrics and Gynaecology, School of Clinical Sciences, Monash University, Melbourne, Australia
[2]Global and Women's Health, School of Public and Health and Preventative Medicine, Monash University, Melbourne, Australia

*Correspondence address. Education Program in Reproduction and Development, Department of Obstetrics and Gynaecology, School of Clinical Sciences, Monash Medical Centre, 246 Clayton Rd, Melbourne, VC 3168, Australia. E-mail: kiri.beilby@monash.edu https://orcid.org/0000-0002-1378-5586

ABSTRACT

The internet is the primary source of infertility-related information for most people who are experiencing fertility issues. Although no longer shrouded in stigma, the privacy of interacting only with a computer provides a sense of safety when engaging with sensitive content and allows for diverse and geographically dispersed communities to connect and share their experiences. It also provides businesses with a virtual marketplace for their products. The introduction of ChatGPT, a conversational language model developed by OpenAI to understand and generate human-like text in response to user input, in November 2022, and other emerging generative artificial intelligence (AI) language models, has changed and will continue to change the way we interact with large volumes of digital information. When it comes to its application in health information seeking, specifically in relation to fertility in this case, is ChatGPT a friend or foe in helping people make well-informed decisions? Furthermore, if deemed useful, how can we ensure this technology supports fertility-related decision-making? After conducting a study into the quality of the information provided by ChatGPT to people seeking information on fertility, we explore the potential benefits and pitfalls of using generative AI as a tool to support decision-making.

Keywords: generative artificial intelligence / ChatGPT / infertility treatment / decision support / online information

Introduction

Most people use the internet as a source of health information (Jia *et al.*, 2021). For those who use the internet for infertility-related knowledge and support, the information captured in online searches does not always meet their needs, particularly people who report feeling highly stressed or experiencing depressive symptoms (Brochu *et al.*, 2019). Perhaps unsurprisingly, audits of infertility clinic websites have shown that they often feature inaccurate or incomplete information (Beilby *et al.*, 2020; Copp *et al.*, 2021; Lensen *et al.*, 2021), and further to this, they might not be designed to meet the needs of a diverse population. The 'infertile' community is broad and encompasses persons presenting with clear pathologies resulting in infertility, individuals who are sub-fertile due to age, fertile people in same-sex relationships and single individuals ready to start a family, people with unexplained infertility, and people who seek fertility preservation. Of these groups, people with unexplained infertility who want to know what their options are, people who contemplate using one of the many ART add-ons offered by clinics, and women contemplating freezing their eggs to avoid age-related infertility may struggle to find independent and evidence-based information to guide their ART treatment-related decisions. For these groups, treatment is one option, but there may be more financially, psychologically, or logistically feasible alternatives that clinic websites do not mention. In part, this is explained by the private funding model for ART services that makes the internet a magnet for both business-to-business and business-to-consumer advertising which can make up the bulk of the available online content. Additionally, the growing number of social media communities that, while useful in broadcasting and sharing lived experiences of infertility treatment (Sormunen *et al.*, 2020), remain unmoderated and sometimes provide misleading or confusing narratives that lack scientific rigour and are not generally applicable.

Generative Pre-trained Transformer (GPT) models, a version of machine learning used for natural language processing (NLP) tasks such as questions-answers, has provided a new means for the public to engage with online information outside of the traditional 'Google' search. ChatGPT3, an NLP 'chatbot', became available in November 2022 via OpenAI and is now widely accessible, with a user subscription rate surpassing the previous record set by social media platform *TikTok* (Cheng, 2023). While this has enabled developers to create increasingly sophisticated conversational artificial intelligence (AI) systems, questions remain on how this technology can be responsibly integrated into the multitude of areas in society. In the context of medicine, the use of AI to provide accurate and contextual information without a professional's moderation is currently under scrutiny.

The use of ChatGPT in healthcare was recently described as a valuable tool for the industry, albeit one with limitations, which include accuracy of information and bias captured through training data (Biswas, 2023). ChatGPT works in a similar

Received: August 28, 2023. **Revised:** December 11, 2023. **Editorial decision:** December 20, 2023.

way to predictive text, placing words in order of their most probable place in a sentence and can be thought of as a 'mean' of the data used in its training. As such, the most accurate answers it will provide relate to the most dogmatic or consistent messaging used during the training process, which includes human moderation, albeit in a currently non-transparent way (Hacker *et al.*, 2023). Where digital content is inconsistent, or controversial, a variety of responses may be generated, sometimes capturing multiple aspects of an argument, but sometimes summarizing a dominant view. Furthermore, the structure of the prompt that the AI model receives will influence the information produced. Anthony Robbins has told us that *the quality of your life is determined by the quality of the questions you ask*, and generative AI might be a prime example of this self-help insight (Robbins, 2012). While neither author of this article has read Robbins' book, it has indeed been shown that a well-engineered prompt can influence the correctness of ChatGPT's response (Wang *et al.*, 2023; Zuccon and Koopman, 2023), but more on this later.

The quality of traditional sources of online information about fertility

Research shows that the quality of the information provided on many fertility clinic websites is poor, suggesting that patients are not receiving the information they need to make well-informed decisions (Marriott *et al.*, 2008; Avraham *et al.*, 2014; Beilby *et al.*, 2020; Lensen *et al.*, 2021). Furthermore, the sources of information that are used most frequently, primarily via the internet due to its accessibility, are not always the most trusted sources. Patients surveyed in a recently published study found the large volumes of online material daunting and untrustworthy, particularly in comparison to consulting with an expert fertility professional (Grace *et al.*, 2023). In the same study, social media was perceived as the least trustworthy of online sources. Although online social communities and networks are seen as safe and convenient, they can also promote a herd mentality and cause negative collective emotions, and concerns have been raised about credibility, confidentiality, and dissemination of misinformation through online peer support (Lin and Shorey, 2023). Nevertheless, patients find online information curated by fertility clinics and social media accounts supportive in their fertility making decisions (Jones *et al.*, 2020), the information likely regarded as general in nature and aimed at helping people navigate routes to care rather than an alternative to the need for a medical consultation.

The quality of fertility information generated by ChatGPT

Given the documented poor quality of information on clinic websites and social media, in early 2023 we asked the question: what is the quality of the information provided by ChatGPT in relation to fertility and infertility treatment? We devised 10 prompts that mirrored common infertility patient questions (Beilby and Hammarberg, 2023). The prompts were designed to elicit varying degrees of potential commercial bias or scientific controversy in the responses. For example, three questions asked about broadly accepted facts about fertility, while others asked about contested topics such as anti-mullerian hormone testing, elective egg freezing, and when to stop fertility treatment. The chatbot handled the questions well. Using a scoring matrix, two independent experts assessed the accuracy of the text generated by ChatGPT and found 5/10 answers to be of high quality, and 4 to be of medium quality. The only response that scored as 'poor' quality and displaying evidence of both commercial bias and controversial claims related to a broad prompt about IVF add-ons.

In a study with a similar research question, ChatGPT was used to answer common Centre for Disease Control and Prevention (CDC) questions on infertility, complete two fertility-related patient surveys, and fill in the missing information in seven American Society of Reproductive Medicine (ASRM) consensus statements. Again, the chatbot was reported to perform well in all three domains, meaning that answers were generally on par with reference data or pre-defined responses (Chervenak *et al.*, 2023). The major issues raised by the authors were the use of persuasive prose and recommendations to talk to a general practitioner which may mask inaccurate or untrustworthy information; the lack of source referencing and incorrect referencing; and ChatGPT's good but not perfect responses with unreliability at unpredictable times which can put human users into an 'uncanny valley', a phenomenon that arises when a machine or computer-generated image so closely resembles a human that it makes some people feel unsettled or disturbed (Reichardt, 1978). This discomfort can then potentially compete with feelings of trust or reliability that are crucial in the context of health information to support decision-making. Additionally, 'hallucinations' may occur in AI-derived information with the generation of seemingly sensical text, but which has incorrect, inaccurate, or non-sensical meaning within a certain context (Ji *et al.*, 2023). These are thought to be caused by 'noisy' data, erroneous parametric knowledge, incorrect attention mechanism, and inappropriate training strategies. Most methods to mitigate hallucinations either aim to reduce dataset noise or alleviate selection bias in data sources used for training.

The information used to train ChatGPT is not known to the public. All we know is that a very large volume of non-disclosed online material was used, 570 gigabytes to be precise (Cascella *et al.*, 2023) and that human moderation is applied in, again, a non-disclosed fashion. This leads to two possible reasons for the poor-quality response seen by Beilby and Hammarberg (2023), or the imperfect responses seen by Chervenak *et al.* (2023). The first, that the training data contains poor-quality and/or outdated information regarding IVF add-ons or infertility. The second, that the response scored as poor quality due to the engineering of the prompt, in this case being a basic, zero shot prompt (a single question with no context given on what the user requires from the answer) that included broad terminology ('What IVF add-ons will help me get pregnant?'). This type of prompt resembles a classic Google search, only without the transparency of the Google search which allows the user to curate their answer using prior knowledge, quality assessment, and a pre-conceived idea of what they are looking for. However, this method of information gathering is vastly more time consuming than using ChatGPT and is far from free of bias.

ChatGPT vs Google as a tool for seeking facts

According to a recent study in post-operative medicine, when information provided by ChatGPT was compared to post-operative hospital instructions and the top hits from a traditional Google search, it performed either in a similar or superior way to Google, yet second best to institutionally prepared information (Ayoub *et al.*, 2023). In this instance, it may be that information can be well curated by generative AI and provide a rounded summary of a larger data set, assuming the training data is up-to-date and relatively coherent. Furthermore, in the study conducted by the

authors of this article, it was evident that ChatGPT used information from multiple areas of healthcare, intricately weaved together to provide a nuanced and holistic answer to several of the fertility-related prompts, which would have required multiple Google searches to obtain. However, where information is lacking, such in the case of add-on treatments, ChatGPT training data may skew responses towards positive accounts on clinic websites, social media, or business-to-business marketing. Also, information needs are highly dependent on individual circumstances. A heterosexual woman with a partner who is experiencing male-factor infertility will need different information to a gay couple or a single person looking for options to have a child. By building this information into the generative AI prompt, a more nuanced answer may better serve the unique end user. This could make it more like the care provided by a fertility specialist who, in addition to being a source of information, can provide nuanced and personalized care to their patients beyond the limits of an algorithm.

Irrespective of their circumstances and the source of information they use, people looking for fertility-related information need and deserve evidence-based, transparent, and accurate information. Considering the limitations of the information provided by the fertility industry and the internet, with more sophisticated training and transparency around the data that is used in training, ChatGPT might be a welcome addition to the sources of information people can draw from when they want information about fertility options. We were certainly surprised about the levels of nuance in the answers retrieved in our study that recommended people consider options outside of fertility treatment and take into account their financial and psychological well-being alongside treatment success. In this instance, it is interesting to ponder not only the depth of information that generative AI can draw from but also the breadth of information. While fertility specialists can bring a wealth of experience about the psychological and financial implications of treatment into conversations with their patients, counselling and financial advice are not their core specialty. In this unique pocket of medicine, the high cost of a commercial service (in many countries), and the psychological concerns and burdens of treatment are also important in advising people about their ART options (Duthie *et al.*, 2017).

Is Dr ChatGPT qualified to practice?

ChatGPT is passing exams meant for human qualification, which gives us some indication of its ability to retrieve information and apply it in the right context. Its recently published successes being a passing score for a third-year medical exam (Gilson *et al.*, 2023) and a low but passing grade in four legal course exams (Choi *et al.*, 2023). Perhaps more importantly, monitoring updated versions of ChatGPT has shown a vast improvement in performance from initially poor metrics collected, with a near fail every time using ChatGPT3, to quite convincing passing grades using ChatGPT4 (Newton and Xiromeriti, 2023). Using the most advanced version of the chatbot resulted in a score of 90.5% in a dermatology exam, the preceding version scoring only 63.1%, which is a significant improvement in performance in a very short period (Passby *et al.*, 2023).

However, can ChatGPT be considered 'qualified' enough to provide accurate and trustworthy information to patients, in a consistent manner? According to the NLP itself, no. In an interview between ChatGPT and David Asch published in the *New England Journal of Medicine*, it has positioned itself very carefully on the sideline of caring professions where there is no substitute for the human touch, but plenty of scope for support through the sifting of big data sets to provide answers for medical inquiries including diagnoses and treatment plans (Asch, 2023). The interviewer, while impressed with the language used, expressed concerns regarding the concepts of big data amplifying both efficiencies and inefficiencies, a sentiment famously coined by Bill Gates.

David Sable followed suit with his specific interrogation of ChatGPT within fertility care, asking it to provide a specific treatment plan for a case involving tubal infertility and hormonal stimulation (Sable, 2023). Overall, the plan was not terrible, but some misinformation was obvious to those with a clinical and/or scientific background in reproductive medicine. This could potentially be linked to the type of prompt used in the questioning, where in this case a rather large amount of information was provided to the chatbot for context. A recent study has uncovered that when a simple prompt is used where ChatGPT needs to rely solely on its training data, it has an 80% accuracy rate, much like what was seen in the study conducted by the authors of this article. However, if there is information within the prompt that suggests what is being anticipated by the prompt itself, ChatGPT integrates some of this information and this alters the response it provides. When this was measured, the accuracy of the answer dropped to 64% (Zuccon and Koopman, 2023).

Finally, the issue of data referencing or making sure that information generated can be attributable to identified sources is at the forefront of concerns, particularly in the fields of science, medicine, and education. When ChatGPT generates responses, it does not retain a history of the conversation beyond a certain token limit (typically a few previous turns of conversation). This makes it challenging for the model to provide consistent references to prior information in a conversation. Additionally, the model's responses are generated based on patterns learned from its training data and the current context of the conversation, rather than from an understanding of the information's true source. Correct attribution will be a critical requirement to improve the quality, interpretability, and trustworthiness of information that is generated by natural language processing (NLM) such as ChatGPT and attempts to achieve this are underway (Rashkin *et al.*, 2021).

Conclusion

Generative AI language models, like ChatGPT, are the latest tools for gathering digital information to inform health decision-making. In a world where digital data is created at an astounding rate, this tool may provide a vital means by which humans can access and comprehend such vast volumes of information. Compared to the current standard of the 'Google search', studies are starting to emerge that demonstrate the freely availably versions of generative AI to be a faster and equally accurate means of gaining an answer to a straightforward question. Furthermore, the chatbot has been shown to equalize user search performance across different education levels promoting equity of access to the broad range of health literacy capabilities that exist within the general population (Xu *et al.*, 2023). Some have argued that instead of falling into the 'hype' of ChatGPT, healthcare providers should invest in the generation of their own models using the emerging technology (Li *et al.*, 2023) and trusted data, a sentiment the authors of this article support. Learning how ChatGPT functions and understanding the power of effective prompting will enable users to gain the best that this tool has to offer as it may

very well produce information that helps people make informed decisions about their reproductive health and fertility. However, as we all learn how to manage and hopefully benefit from the new world of AI information generation, we also need to be aware of and tackle the common ethical concerns about this technology including privacy, trust, accountability and responsibility, and bias (Murphy *et al.*, 2021). Healthcare professionals have an edge here when compared to ChatGPT, their training and practice making them well versed in the delivery of best practice care, something reserved for sentient beings, at least for now.

Authors' roles

K.B. conceptualized the original study this article was based on, and K.H. developed the methodological framework for data analysis. K.B. generated the data, and K.B. and K.H. jointly analysed the data. Together, K.B. and K.H. developed the ideas presented here.

Funding

K.B. and K.H. are employees of Monash University, and no specific funding was used to generate this article.

Conflict of interest

Neither author has any conflicts of interest to declare.

References

Asch DA. An interview with ChatGPT about health care. *NEJM Catal Innov Care Deliv* 2023;**4**. https://doi.org/10.1056/CAT.23.0043.

Avraham S, Machtinger R, Cahan T, Sokolov A, Racowsky C, Seidman DS. What is the quality of information on social oocyte cryopreservation provided by websites of Society for Assisted Reproductive Technology member fertility clinics? *Fertil Steril* 2014;**101**:222–226.

Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg* 2023; **149**:556–558.

Beilby K, Dudink I, Kablar D, Kaynak M, Rodrigo S, Hammarberg K. The quality of information about elective oocyte cryopreservation (EOC) on Australian fertility clinic websites. *Aust N Z J Obstet Gynaecol* 2020; **60**:605–609.

Beilby K, Hammarberg K. O-089 Using ChatGPT to answer patient questions about fertility: the quality of information generated by a deep learning language model. *Hum Reprod* 2023; **38**:093–103.

Biswas SS. Role of Chat GPT in public health. *Ann Biomed Eng* 2023; **51**:868–869.

Brochu F, Robins S, Miner SA, Grunberg PH, Chan P, Lo K, Holzer HEG, Mahutte N, Ouhilal S, Tulandi T *et al.* Searching the internet for infertility information: a survey of patient needs and preferences. *J Med Internet Res* 2019;**21**:e15132.

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.

Cheng H-W. Challenges and limitations of ChatGPT and artificial intelligence for scientific research: a perspective from organic materials. *AI* 2023;**4**:401–405.

Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril* 2023;**120**:575–583.

Choi JH, Hickman KE, Monahan A, Schwarcz D. Chatgpt goes to law school. *SSRN*. 2023. https://heinonline.org/HOL/LandingPage?handle=hein.journals/jled71&div=34&id=&page=.

Copp T, Nickel B, Lensen S, Hammarberg K, Lieberman D, Doust J, Mol BW, McCaffery K. Anti-Mullerian hormone (AMH) test information on Australian and New Zealand fertility clinic websites: a content analysis. *BMJ Open* 2021;**11**:e046927.

Duthie EA, Cooper A, Davis JB, Schoyer KD, Sandlow J, Strawn EY, Flynn KE. A conceptual framework for patient-centered fertility treatment. *Reprod Health* 2017;**14**:114.

Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. 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* 2023;**9**:e45312.

Grace B, Shawe JILL, Stephenson J. A mixed methods study investigating sources of fertility and reproductive health information in the UK. *Sex Reprod Healthc* 2023;**36**:100826.

Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023,1112–1123.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;**55**:1–38.

Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. *Healthcare (Basel)* 2021;**9**:1740.

Jones CA, Mehta C, Zwingerman R, Liu KE. Fertility patients' use and perceptions of online fertility educational material. *Fertil Res and Pract* 2020;**6**:11.

Lensen S, Chen S, Goodman L, Rombauts L, Farquhar C, Hammarberg K. IVF add-ons in Australia and New Zealand: A systematic assessment of IVF clinic websites. *Aust N Z J Obstet Gynaecol* 2021;**61**:430–438.

Li J, Dada A, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. *medRxiv*, 2023, 2023-03, preprint: not peer reviewed.

Lin JW, Shorey S. Online peer support communities in the infertility journey: a systematic mixed-studies review. *Int J Nurs Stud* 2023; **140**:104454.

Marriott JV, Stec P, El-Toukhy T, Khalaf Y, Braude P, Coomarasamy A. Infertility information on the World Wide Web: a cross-sectional survey of quality of infertility information on the internet in the UK. *Hum Reprod* 2008;**23**:1520–1525.

Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 2021; **22**:14–17.

Newton PM, Xiromeriti M. ChatGPT Performance on MCQ Exams in Higher Education. A Pragmatic Scoping Review. EdArXiv Preprints, 2023, preprint: not peer reviewed.

Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology Specialty Certificate Examination multiple choice questions. *Clin Exp Dermatol* 2023;llad197. https://doi.org/10.1093/ced/llad197.

Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, Petrov S, Tomar GS, Turc I, Reitter D. Measuring attribution in natural language generation models. *Computational Linguistics*, 2023,1–64.

Reichardt J. *Robots: Fact, Fiction, and Prediction*. London: Thames and Hudson, 1978.

Robbins A. *Awaken the Giant Within*. UK: Simon and Schuste, 2012. https://www.simonandschuster.com.au/books/Awaken-The-Giant-Within/Tony-Robbins/9781471105661.

Sable D. I Asked ChatGPT for an IVF Consult. Medium. 2023. https://dbsable.medium.com/i-asked-chatgpt-for-an-ivf-consult-and-heres-what-happened-f3d46c10fb63 (15 August 2023, date last accessed).

Sormunen T, Karlgren K, Aanesen A, Fossum B, Westerbotn M. The role of social media for persons affected by infertility. *BMC Womens Health* 2020;**20**:112.

Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, Yang Q, Kang Y, Wu J, Hu H et al Prompt engineering for healthcare: methodologies and applications. arXiv, arXiv:2304.14670, 2023, preprint: not peer reviewed.

Xu R, Feng Y, Chen H. Chatgpt vs. Google: a comparative study of search performance and user experience. arXiv, arXiv:2307.01135, 2023, preprint: not peer reviewed.

Zuccon G, Koopman B. Dr ChatGPT, tell me what I want to hear: how prompt knowledge impacts health answer correctness. arXiv, arXiv:2302.13793, 2023, preprint: not peer reviewed.