Gynecology and Obstetrics Clinical Medicine 3 (2023) 203–206

Contents lists available at ScienceDirect

Gynecology and Obstetrics Clinical Medicine

journal homepage: www.keaipublishing.com/en/journals/gynecology-and-obstetrics-clinical-medicine

Short Communication

Comparative efficacy of ChatGPT 3.5, ChatGPT 4, and other large language models in gynecology and infertility research

Pallav Sengupta [a,*], Sulagna Dutta [b], Srikumar Chakravarthi [c], Ravindran Jegasothy [d], Ravichandran Jeganathan [e], Anuradha Pichumani [f]

[a] College of Medicine, Gulf Medical University, Ajman, United Arab Emirates
[b] School of Medical Sciences, Bharath Institute of Higher Education and Research, TN, India
[c] Faculty of Medicine, SEGi University, Kota Damansara, Malaysia
[d] Faculty of Medicine, MAHSA University, Jenjarom, Malaysia
[e] Department of Obstetrics & Gynaecology, Hospital Sultanah Aminah, Johor Bahru, Malaysia
[f] Sree Renga Hospital, Chengalpattu, Tamil Nadu, India

Dear Editor,

As experts in gynecology and infertility research, we have witnessed the rapid advancement of artificial intelligence (AI) technologies, particularly language models, which have greatly improved their capabilities. In this communication, we aim to compare the proficiency of advanced language models, such as Chat Generative Pre-Trained Transformer (ChatGPT) 3.5, ChatGPT 4, and others, in relation to our field.

AI-driven algorithms have potential to accelerate research by analyzing vast amounts of academic literature, identifying complex correlations, and synthesizing fragmented knowledge to form a cohesive understanding.[1] These advanced systems have the ability to revolutionize the way we conduct research, extract valuable insights, and translate them into effective clinical applications. The power of AI-based frameworks can speed up scientific progress and improve the challenging process of extracting valuable insights from extensive and complex existing knowledge repositories.[2]

ChatGPT 3.5, a cutting-edge generative pre-trained transformer by OpenAI, augments numerous disciplines, including content generation, linguistic metamorphosis, and complex data analysis. However, its knowledge repository, limited to pre-September 2021 data, requires supplementation for optimal utility.[3] ChatGPT 4, surpasses its antecedent in linguistic cognition, generalizability, and response quality. Enhanced contextual processing and discernment facilitate more refined interactions in gynecology and infertility research, while adeptly addressing intricate queries, thus elevating the investigative process to unprecedented levels.[4]

Compared to GPT-3, ChatGPT 3.5 showcased superior natural language generation, while ChatGPT 4 further refined this quality. The performance, however, varies among large language models(LLMs). Whenever gynecology or infertility related questions or instructions are provided, understanding of the context is vital for the LLMs in order to deliver coherent outputs. ChatGPT 3.5 enhanced comprehension of context and user input, a trait that ChatGPT 4 further sharpened, coupled with a better memory for user interactions. Yet, this understanding is inconsistent across LLMs. When synthesizing data, especially complex research in fields like gynecology or infertility, ChatGPT 3.5 effectively interprets findings, a capability ChatGPT 4 excels in by delivering coherent summaries. This ability, though, depends on the LLMs in use, with some requiring fine-tuning. In customization, ChatGPT 3.5 provides fine-tuning opportunities, especially in specialized fields. ChatGPT 4 enhances this feature, supporting extensive user-driven fine-tuning and domain adaptation, though the extent varies among LLMs. Ethically, ChatGPT 3.5 addressed biases and concerns in gynecologic or infertility data or information, better than GPT-3, and ChatGPT 4 strengthened these efforts with advanced bias control. However, ethical attentiveness differs among LLMs. Challenges include increased processing times due to model complexities, overfitting risks, limited interpretability, and higher deployment costs, with ChatGPT 4 demanding more resources and incurring greater costs than its predecessor. These concerns differ across LLMs. Focusing on gynecology, several studies validated the efficacy of ChatGPT. To assess the efficacy of ChatGPT, Kemp MW et al.[5] conducted a rigorous investigation within the context of a simulated clinical evaluation pertaining to the Royal College of Obstetricians and Gynecologists (RCOG) membership virtual Objective Structured Clinical Examination (OSCE). The advanced language model was subjected to a series of seven

* Corresponding author. Department of Biomedical Sciences, College of Medicine Gulf Medical University, Ajman, United Arab Emirates.
E-mail address: pallav_cu@yahoo.com (P. Sengupta).

Publishing services by Elsevier on behalf of KeAi

https://doi.org/10.1016/j.gocm.2023.09.002

Available online 17 September 2023

meticulously crafted, structured discourse queries, with its subsequent responses undergoing an unbiased assessment by a cohort of 14 certified examiners. These assessments were subsequently juxtaposed with the historical performance of human examinees. Remarkably, ChatGPT attained an average performance metric of 77.2%, surpassing human candidates across multiple knowledge domains relevant to obstetrics and gynecology. In another study by Santo DSE et al.,[6] the utility of ChatGPT as a resource for guidance during unanticipated labor events was scrutinized. The findings of this study demonstrated that ChatGPT possesses considerable potential as a valuable adjunctive instrument to assist individuals confronted with unforeseen labor situations. However, considering these reports and also limitations of an AI-driven LLMs in gynecology practice and infertility research, Grunebaum A et al.[7] have opined that ChatGPT possesses considerable efficacy in proffering foundational knowledge pertaining to the domain of obstetrics and gynecology, as corroborated by its elaborate, eloquent, erudite, and syntactically coherent responses to a multifarious assortment of queries (Table 1).[5–11]

Given the rapid expansion of AI, gynecology and infertility researchers and clinicians should stay abreast of state-of-the-art developments and incorporate these technological tools to improve the quality and rigor of their empirical work. Language models can aid in identifying therapeutic targets, analyzing clinical trial data, and exploring innovative treatments by processing vast scientific texts.[7] They enable rapid access to relevant data on gynecological pathologies, therapeutic interventions, and fertility alternatives through natural language inquiries and assimilation of evidence-based insights from extensive medical literature. They can also generate diagnostic hypotheses and offer informed therapeutic recommendations for specific conditions, augmenting the expertise of healthcare professionals.[12] This approach enhances the decision-making abilities of clinicians and empowers superior diagnostic accuracy and patient outcome prediction through data-driven algorithms. It expedites the mining of medical records via natural language processing, obviating laborious chart reviews. Furthermore, AI fosters educational innovation, enabling context-aware training modules, augmenting scholastic proficiency, and catalyzing groundbreaking advancements in healthcare.[13] (Fig. 1).

Elucidating the intricate mechanisms governing male and female infertility has long been a subject of paramount importance.[14] Leveraging the prodigious capabilities of LLMs presents a promising avenue to unravel the complex underpinnings of infertility and expedite advancements in this domain. By integrating LLMs into investigative frameworks, researchers and clinicians can synergistically consolidate disparate sources of information, facilitate hypothesis generation, and harness a robust knowledge base to propel future investigations.[1] Current knowledge about infertility is riddled with gaps, primarily due to the convoluted nature of reproductive processes and the myriad factors that modulate them. A comprehensive understanding of the molecular, cellular, and physiological mechanisms in both male and female infertility remains elusive. Consequently, the multifactorial etiology of infertility, encompassing genetic, epigenetic, and environmental factors, necessitates innovative approaches to bridge these lacunae in our understanding.[15–17] By consistently refreshing their information repository, LLMs can rapidly assimilate new findings and information from various domains associated with infertility, including genetics, endocrinology, embryology, and more. Their sophisticated analysis abilities can pinpoint previously missed patterns or links, resulting in innovative theories in infertility studies and therapeutic strategies, potentially tailoring treatments to individual needs. For example, utilizing LLMs to analyze high-dimensional, multifaceted data sets derived from genomic, transcriptomic, proteomic, and metabolomic studies can facilitate the identification of novel biomarkers and pathways implicated in infertility.[18] This knowledge can subsequently be employed to guide the development of targeted therapeutic interventions and bolster personalized medicine strategies.[18] LLMs can also analyze patient responses to the treatments based on intricate datasets. Moreover, LLMs can assist in uncovering novel gene-gene and gene-environment interactions, thereby illuminating the intricate interplay between genetic predispositions and environmental exposures in the context of infertility. By capitalizing on the capacity of LLMs to scrutinize vast troves of scientific literature, researchers and clinicians can derive contextually relevant, data-driven insights that can augment our understanding of the etiopathogenesis of infertility[12] (Fig. 1). Furthermore, use of these modalities can level the playing field in many nations where clinicians of varying experience and knowledge levels can be brought up to par in deciding on evidence-based

Table 1
Comparative view of ChatGPT 3.5, ChatGPT 4, and other LLMs, and studies in Gynecology Research.

Model/Feature	ChatGPT 3.5	ChatGPT 4	Other Large Language Models (LLMs)
Pros			
1. Text generation quality	Improved natural language generation compared to GPT-3[8,10,11]	Further advancements in linguistic performance, more refined answers[8,10]	Varying text generation quality across different LLMs[11]
2. Understanding of context	Enhanced comprehension of context and user input[8,10,11]	More accurate and precise context understanding, better long-term memory	Variable context understanding, dependent on the model[11]
3. Data synthesis	Capable of synthesizing complex gynecological research[11]	Increased efficacy in producing coherent summaries of research findings[11]	Capability dependent on specific LLMs, may require fine-tuning[11]
4. Customization	Allows limited fine-tuning for specific domains like gynecology[8,10,11]	Enhanced user-driven fine-tuning and domain adaptation[8,10,11]	Some models offer extensive fine-tuning possibilities[11]
5. Ethical considerations	Reduced biases and ethical concerns compared to GPT-3[11]	Stronger mechanisms to control biases and address ethical concerns[8,10,11]	Varying levels of ethical considerations across LLMs[11]
Cons			
1. Complexity and processing time	More complex model may result in increased latency[8,10,11]	Enhanced complexity, may require significant computational resources[11]	Model complexity may vary, leading to variable processing times[11]
2. Overfitting	Possible risk of overfitting to gynecological data[11]	Overfitting to specific gynecology and infertility data may occur[11]	Risk of overfitting can depend on the LLMs and domain[11]
3. Interpretability	Limited model interpretability[8,10,11]	Further limitations in model interpretability[8,10,11]	LLMs can vary in interpretability, often limited[11]
4. Deployment cost	Higher cost for deployment and maintenance compared to GPT-3[11]	Increased deployment cost due to larger model size and complexity[11]	Deployment costs can differ significantly among LLMs[11]
Studies on gynecology research			
Kemp MW et al., 2023[5]	The study evaluated the performance of ChatGPT in a mock clinical examination for membership of the Royal College of Obstetricians and Gynecologists (RCOG). The system was tested using seven structured discussion questions, and its responses were blind-scored by 14 qualified examiners, compared to historical human candidate performances. ChatGPT achieved an average score of 77.2%, outperforming human candidates in several subject domains.		
Santo DSE, 2023[6]	This study explored the use of ChatGPT for guidance during unexpected labour and found that ChatGPT can be a valuable tool to help people during unexpected labour		
Levin G et al., 2023[9]	ChatGPT proficiently generates text output with a natural, human-like style, akin to human-written text. It is an accessible resource for all users, contributing to its widespread utilization in obstetrics and gynecology.		

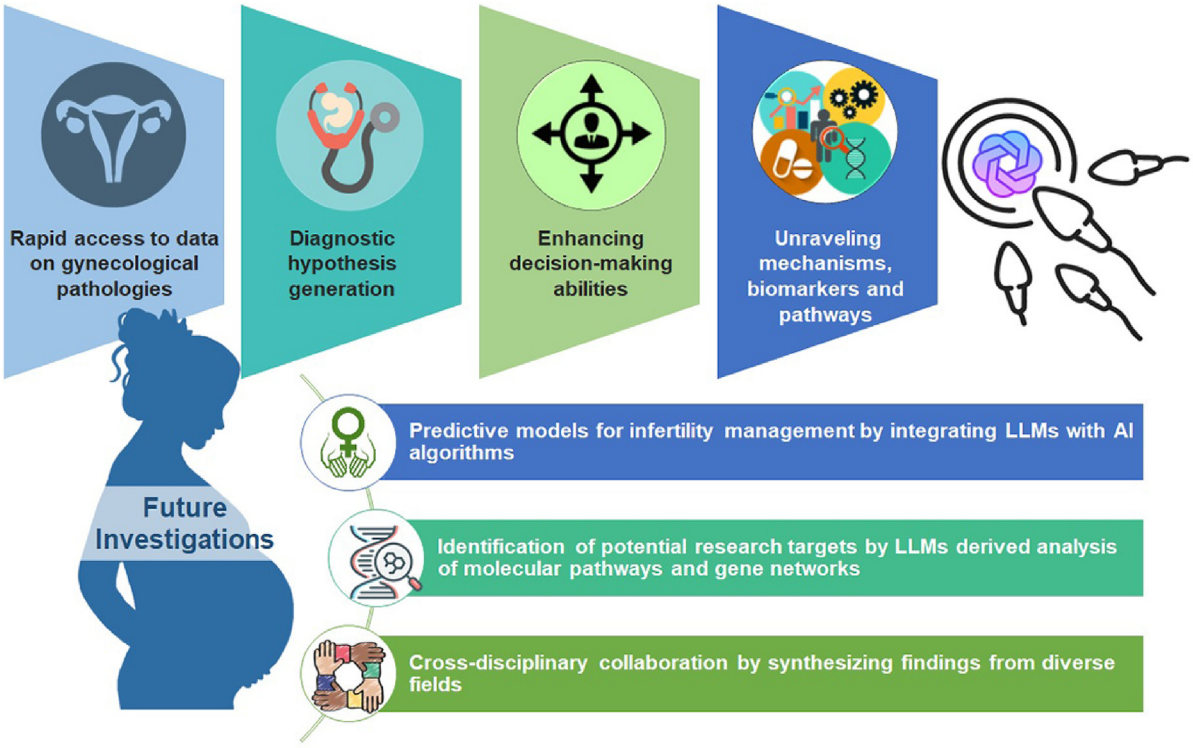

Fig. 1. Multi-faceted applications of language models in gynecology and infertility research.

treatment modalities and strategies in an exponentially evolving field. Patient safety can be strengthened by ensuring research-backed treatment.

LLMs can also be used to obtain directions for future investigations by integration of LLMs with machine learning and artificial intelligence algorithms to establish predictive models for infertility diagnosis, prognosis, and response to treatment; leveraging LLMs to identify potential targets for pharmacological intervention by analyzing molecular pathways and gene networks implicated in infertility[12]; employing LLMs to discern epigenetic modifications, such as DNA methylation and histone modifications, that may contribute to the pathogenesis of infertility and serve as targets for therapeutic modulation; exploiting LLMs to facilitate cross-disciplinary collaboration, by synthesizing findings from diverse fields, including endocrinology, immunology, and genetics, to elucidate the multifactorial nature of infertility; and also by utilizing LLMs to enhance patient counseling and education, by generating personalized, evidence-based recommendations grounded in the latest scientific advancements.[1]

Nevertheless, ChatGPT has certain limitations as it often miscodes human language and may occasionally provide inaccurate information due to constant updates and user adaptation. Researchers and clinicians should use ChatGPT judiciously to avoid plagiarism.[19] It is noteworthy that the training data of the model is static, necessitating cognizance of the users of its obsolescence. Moreover, adding citation features could improve its relevance in the research field. Besides ChatGPT, it also remains crucial to assess the efficacious potentialities of alternative extensive language models, such as BERT, T5, and GPT-Neo. Systems that update themselves regularly and release new versions will gain an advantage over their competitors. The review of 118 publications shows that ChatGPT can act as a healthcare "clinical assistant" for tasks like patient inquiries and research but struggles with issues like inconsistency, bias, and legality. Additionally, it has potential in academic writing but faces challenges like plagiarism and a lack of human-like qualities, questioning its authority as an author.[20] Thus, the effectiveness and reliability of the LLMs should be enhanced further, both in healthcare and academic settings, and the developers of ChatGPT may need to focus on addressing the identified issues such as bias, plagiarism, and improving human-like interactions.[21] The application of LLMs to gynecology and infertility research presents notable constraints, alluding to prolonged deployment times, paucity of data accessibility, dearth of expertise, and a demand for a further extensive investigation to thoroughly evaluate ethical, legal, and data security considerations. Temporal considerations factor significantly into LLMs deployment. The construction of these models requires an extended period for data aggregation, model training, and optimization, further delaying the implementation of such technologies in clinical practice.[12] In gynecology and infertility research, where time can often play a critical role, these protracted deployment durations can represent a substantial limitation. Accessibility and quality of data pose another significant barrier. Infertility and gynecological data, like other medical data, are subject to stringent privacy regulations, hindering the ease of data accessibility. In addition, given the sensitive nature of gynecological and fertility data, acquiring a sufficiently large and diverse dataset to train the LLMs can prove arduous.[12,13] The inherent complexity of gynecological conditions, their multifactorial etiologies, and heterogeneous presentations demand comprehensive, high-quality, and diverse datasets that can be challenging to amass. Regarding the human aspect, when it comes to LMMs, there is a lack of information driven by experts. This interdisciplinary skill set is requisite for effective LLMs implementation and tailoring the models to specific research questions. Without a critical mass of such expertise, the models risk misinterpretation or misuse. Further extensive studies are indispensable to holistically understand the applicability of LLMs in this domain. As LLMs grow more complex, there is an increasing need to evaluate not only their performance but also their reliability, transparency, and fairness. Protocols for fertility-related information usage must ensure informed consent, clarifying to patients that their data will be processed with this tool while maintaining confidentiality. Importantly, as the tool is data-driven, it avoids plagiarism concerns when multiple researchers investigate identical topics. Finally, and critically, the application of LLMs to gynecology and infertility research

carries with it substantial ethical, legal, and data security implications. The use of sensitive medical data demands stringent measures to ensure patient privacy and data security. Ethical considerations, such as ensuring fair and unbiased model outputs, maintaining transparency in model decisions, and obtaining informed consent from patients for data usage, also necessitate careful navigation. Legal frameworks regulating the use of AI in healthcare, often lagging behind technological advances, further complicate the deployment of LLMs in this field.[13]

In light of the rapidly expanding domain of artificial intelligence, it is incumbent upon investigators in the fields of gynecology and infertility research to maintain a comprehensive awareness of these state-of-the-art developments, judiciously incorporating such technological instruments to augment the quality and intellectual rigor of their empirical endeavors.[1] LLMs aid in the identification of prospective therapeutic targets, analysis of clinical trial data, and exploration of innovative treatment modalities through the processing of voluminous scientific texts.[1] LLMs enable expeditious access to germane data regarding gynecological pathologies, therapeutic interventions, and fertility alternatives, through the processing of natural language inquiries and assimilation of evidence-based insights from an extensive corpus of medical literature. They can also proffer specific diagnostic conjectures for gynecological- and infertility-associated conditions, supplementing the expertise of healthcare professionals and expediting the diagnostic trajectory.[7] A study revealed that Flan-T5, a publicly available LLM, efficiently phenotyped patients with postpartum hemorrhage (PPH), achieving a 0.95 positive predictive value and identifying 47% more patients than standard claims codes, even without manual annotation. Its ability to extract 24 detailed concepts allowed for the creation of intricate phenotypes and subtypes related to PPH, surpassing claims-based approaches and offering a flexible, easily updatable algorithm.[22] These can generate informed therapeutic propositions for specific gynecological and infertility challenges by appraising current medical research and guidelines, thus empowering clinicians and researchers in decision-making and enhancing comprehension of available options.[7]

This article serves both as a testament to the indispensable role AI is poised to play in catalyzing groundbreaking advancements within the realm of reproductive health research and as an exhortation to expeditiously capitalize on the plethora of possibilities furnished by this dynamic and ever-expanding technological milieu. Employing these sophisticated computational methods and algorithms can provide unprecedented insights and facilitate innovative solutions to overcome challenges in the gynecological and infertility research landscape.

Declaration of competing interest

None.

References

1. Sok S, Heng K. *ChatGPT for education and research: a review of benefits and risks*. 2023. https://doi.org/10.2139/ssrn.4378735. SSRN 4378735.
2. Baidoo-Anu D, Owusu Ansah L. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. 2023:4337484. https://doi.org/10.2139/ssrn.4337484.
3. Marron L. *Exploring the potential of ChatGPT 3.5 in higher education: benefits, limitations, and academic integrity. Handbook of research on redesigning teaching, learning, and assessment in the digital era*. IGI Global; 2023:326–349.
4. Dowling M, Lucey B. ChatGPT for (finance) research: the Bananarama conjecture. *Finance Res Lett*. 2023;53:103662.
5. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;229(2):172.e1–172.e12. https://doi.org/10.1016/j.ajog.2023.04.020.
6. Santo DSE, Joviano-Santos JV. Exploring the use of ChatGPT for guidance during unexpected labour. *Eur J Obstet Gynecol Reprod Biol*. 2023;285:208–209. https://doi.org/10.1016/j.ejogrb.2023.04.001.
7. Grünebaum A, Chervenak J, Pollet SL, et al. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023. https://doi.org/10.1016/j.ajog.2023.03.009.
8. Teebagy S, Colwell L, Wood E, et al. Improved performance of ChatGPT-4 on the OKAP Exam: a comparative study with ChatGPT-3.5. *medRxiv*. 2023, 2023.04.03.23287957.
9. Levin G, Meyer R, Yasmeen A, et al. ChatGPT-written OBGYN abstracts fool practitioners. *Am J Obstet Gynecol*. 2023. https://doi.org/10.1016/j.ajogmf.2023.100993.
10. Plevris V, Papazafeiropoulos G, Rios AJ. *Chatbots put to the test in math and logic problems: a preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. AI*. 2023;4(4):949–969. https://doi.org/10.3390/ai4040048.
11. Caramancion KM. *News Verifiers Showdown: a comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking*. 2023. arXiv:230617176.
12. Rao AS, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv*. 2023. https://doi.org/10.1101/2023.02.21.23285886.
13. Kleesiek J, Wu Y, Stiglic G, et al. An opinion on ChatGPT in health care—written by humans only. *J Nucl Med*. 2023;64(5):701–703. https://doi.org/10.2967/jnumed.123.265687.
14. Mustafa M, Sharifa A, Hadi J, et al. Male and female infertility: causes, and management. *IOSR J Dent Med Sci*. 2019;18:27–32.
15. Bhattacharya K, Sengupta P, Dutta S, et al. Obesity, systemic inflammation and male infertility. *Chem Biol Lett*. 2020;7(2):92–98.
16. Agarwal A, Sengupta P. *Male infertility: contemporary clinical approaches, andrology, ART and antioxidants. Oxidative stress and its association with male infertility*. 2020:57–68.
17. Darbandi M, Darbandi S, Agarwal A, et al. Reactive oxygen species-induced alterations in H19-Igf2 methylation patterns, seminal plasma metabolites, and semen quality. *J Assist Reprod Genet*. 2019;36:241–253. https://doi.org/10.1007/s10815-018-1350-y.
18. Patrinos GP, Sarhangi N, Sarrami B, et al. Using ChatGPT to predict the future of personalized medicine. *Pharmacogenomics J*. 2023;23(6):178–184. https://doi.org/10.1038/s41397-023-00316-9.
19. King MR, ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell Mol Bioeng*. 2023;16(1):1–2. https://doi.org/10.1007/s12195-022-00754-8.
20. Garg RK, Urs VL, Agrawal AA, et al. Exploring the role of Chat GPT in patient care (diagnosis and treatment) and medical research: a systematic review. *Health Promot Perspect*. 2023;13(3):183–191. Published 2023 Sep 11. doi:10.34172/hpp.2023.22.
21. Sengupta P, Dutta S. ChatGPT guidance for reproductive specialists: dr. Jekyll or Mr. Hyde? *EXCLI J*. 2023;22:911–914. https://doi.org/10.17179/excli2023-6120.
22. Alsentzer E, Rasmussen MJ, Fontoura R, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *medRxiv*. 2023;2023. https://doi.org/10.1101/2023.05.31.23290753.