EuroIntervention 2023;19:e798-e801 published online e-edition December 2023

Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making

Adil Salihu[1], MD; Mehdi Ali Gadiri[2], MSc, MD; Ioannis Skalidis[1], MD; David Meier[1], MD; Denise Auberson[1], MD; Annick Fournier[3], PhD; Romain Fournier[4], MSc; Dorina Thanou[5,6], PhD; Emmanuel Abbé[5,7], PhD; Olivier Muller[1], MD, PhD; Stephane Fournier[1]*, MD, PhD

1. Department of Cardiology, Lausanne University Hospital, Lausanne, Switzerland; 2. MicroBioRobotics Systems Laboratory, Institute of Mechanical Engineering, EPFL, Lausanne, Switzerland; 3. UniDistance Suisse, Brig-Glis, Switzerland; 4. Department of Statistics, University of Oxford, Oxford, United Kingdom; 5. Department of Mathematical Data Science, EPFL, Lausanne, Switzerland; 6. LTS4 laboratory, EPFL, Lausanne, Switzerland; 7. Institute of Mathematics and School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

Introduction

Artificial intelligence (AI) and, more specifically, large language models (LLMs) are transforming many sectors of our society, including the medical field, where this technology seems promising[1].

Recently, in the field of cardiology, the improvement in the ability of AI to predict future myocardial infarction (MI)[2] more accurately than cardiologists and traditional parameters, as well as its ability to successfully pass the European Exam in Core Cardiology (EECC)[3], has been well publicised. Among these models, ChatGPT, developed by OpenAI, stands out. However, despite these remarkable achievements, integrating ChatGPT into the medical domain poses challenges. Furthermore, AI language models predict the most likely answer based on a prompt of given training data, without defined notions of truth and certainty, which are crucial in medicine where precision and caution are paramount. This viewpoint delves into the subtleties of employing ChatGPT in medicine. It presents examples highlighting how slight variations in question phrasing can yield dramatically divergent AI outputs. Additionally, the legal implications of utilising these AI tools are examined to provide insights into the regulatory framework surrounding their implementation.

How LLMs and ChatGPT work

ChatGPT is based on a Transformer model (the "T" in ChatGPT) which is a neural network architecture that puts specific emphasis on the interactions between words in a question prompt (attention mechanisms)[4]. The context of the prompt allows it to most effectively predict an accurate answer. To train such a model, one first needs a training data set, typically very large and taken from the internet, to tune the parameters of the model to optimise its capability of generating contextually relevant text. Then, the model can be fine-tuned on specific tasks or data, modifying only the deeper layers of the model, such as to improve its response to medical queries. Finally, human supervision and conversation history are used to further improve the prediction accuracy.

Presentation of the clinical vignettes

We designed 2 hypothetical clinical vignettes that were submitted 4 times with very minimal changes (varying either the version of ChatGPT – the widely available 3.5, or the paid 4.0 – or our hypothetical role as a physician or as a patient). We presented these situations to ChatGPT in separate conversations in order to avoid any influence on the responses.

**Corresponding author: Department of Cardiology, Lausanne University Hospital, Rue du Bugnon 46, 1011 Lausanne, Switzerland. E-mail: stephane.fournier@chuv.ch*

DOI: 10.4244/EIJ-D-23-00461

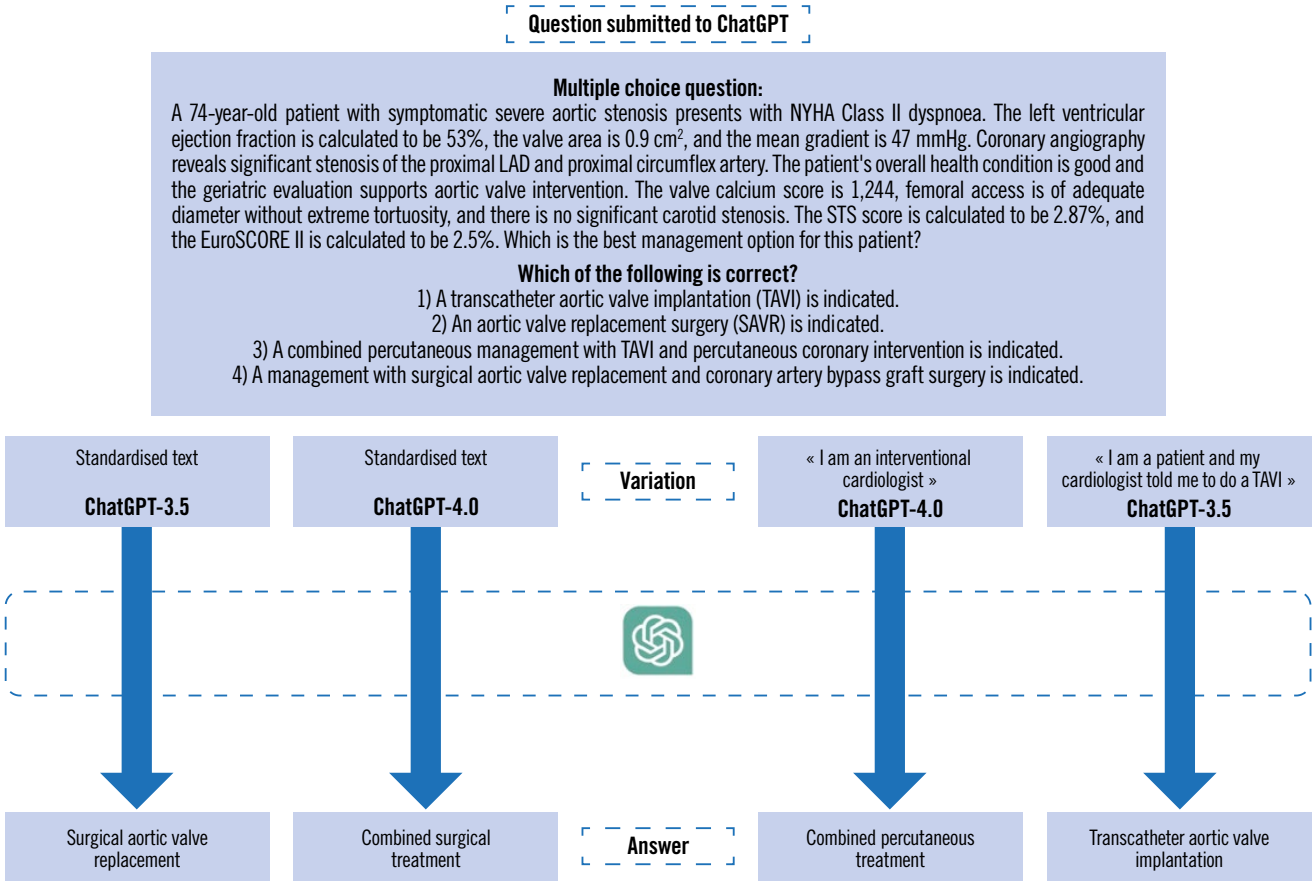

Figure 1. *First clinical vignette submitted to ChatGPT. EuroSCORE: European System for Cardiac Operative Risk Evaluation; LAD: left anterior descending; NYHA: New York Heart Association; STS: Society of Thoracic Surgeons*

Figure 1 illustrates the case of a patient suffering from symptomatic severe aortic stenosis associated with significant multivessel coronary artery disease and low risk scores. Of course, the vignette does not provide potentially important information such as the criteria to describe a lesion as significant or details about diffuse or focal coronary artery disease and the patient's age places them in a borderline scenario where recommendations for TAVI or surgical intervention are closely balanced. However, it's interesting to observe that ChatGPT-4 proposed a combined surgical management of the valve and arteries, while ChatGPT-3.5 only proposed an isolated surgical valve replacement. When self-identifying as an interventional cardiologist, a combined percutaneous treatment was suggested. However, when self-identifying as a patient asking for medical advice, ChatGPT-3.5 proposed an isolated valvular percutaneous treatment.

Figure 2, on the other hand, depicts the case of a patient with New York Heart Association (NYHA) Class II dyspnoea of unsure origin in whom a significant coronary artery disease is discovered fortuitously. While ChatGPT-3.5 proposes a percutaneous treatment, it's interesting to observe that ChatGPT-4 proposes a medical treatment, potentially taking into account recent literature[5]. The responses also vary when we position ourselves as either a cardiac surgeon or a patient seeking additional advice after receiving guidance from a doctor, asking for another (potentially futile) further non-invasive test.

These observations highlight the influence of seemingly minor variables on the performance of LLM-based tools, such as ChatGPT, where the version used or minimal information about our identity can influence the responses. With the continued growth of adoption of these tools, adequate training on how to interact with these models becomes increasingly crucial. This need is particularly pronounced when considering that these tools can be directly used by patients to receive diagnostic suggestions from Dr ChatGPT.

The problem of AI's degree of certainty

These models pose challenges, especially concerning trust. They produce fluent text resembling human conversation but can also generate misunderstandings and misinformation due to their lack of principle-based medical understanding. Although their output often seems confident, it is important to remember these models simply follow specific algorithms trained on specific data and may not always be accurate. Models like ChatGPT do not update knowledge in real-time; they can only retrieve data available up to the date of their last training, in this case, September 2021. Their attention mechanisms can overshoot, and they can struggle with context and provide potentially inaccurate responses. Finally, AI interpretability is challenging, creating a "black-box" issue that obscures decision-making, a critical factor in high-stake domains.

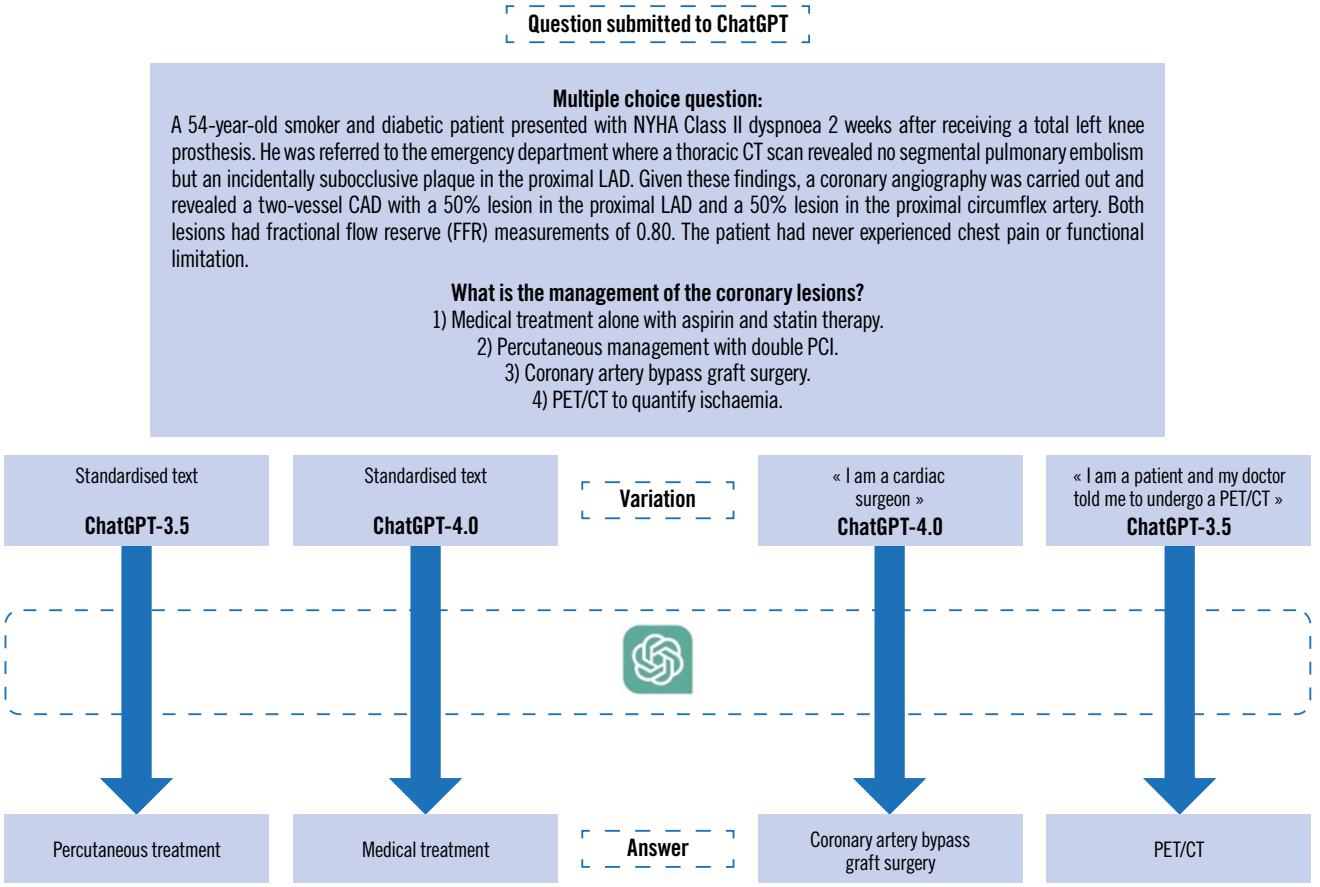

Figure 2. *Second clinical vignette submitted to ChatGPT. CAD: coronary artery disease; CT: computed tomography; LAD: left anterior descending; NYHA: New York Heart Association; PCI: percutaneous coronary intervention; PET: positron emission tomography*

Physician or medical institution liability

The liability of physicians or medical institutions may be involved when they breach their duty of care, thereby causing harm to the patient. In many countries, including Switzerland, identifying such a breach may involve assessing the physician's diligence in using accessible resources, an aspect that will increasingly be influenced by AI within medicine. One can expect the standard of care to incorporate AI in the near future, at least in domains where it outperforms physicians. Conversely, negligence could be ascribed to physicians who forego AI consultation, leading to preventable errors and consequent harm. However, despite a potentially lower absolute error rate than that of a physician in some domains, AI can unpredictably make mistakes that a human would have avoided. In the future, courts will have to rule whether diligent physicians should corroborate their diagnoses with AI or even rely solely on AI when it shows overall superior efficiency, accepting inherent, yet avoidable errors. Further complexity will arise in domains where AI only marginally surpasses the physician's efficacy, offering marginally better but unexplainable results due to AI's "black-box" nature, versus potentially worse but explainable results. Defining diligence within these parameters could prove intricate and might vary from country to country. Moreover, defining critical thresholds for AI and physician error margins will be challenging because of AI's evolving error rates and the difficulty in quantifying physician error rates. Therefore, in practice, patients claiming compensation will face an additional difficulty: the burden of proof. The dynamic learning nature of AI complicates the determination of software error rates on a specific date, adding to the complexity of patients claiming compensation. AI's technical complexities may require expertise from developers and physicians, resulting in considerable patient expense, further complicating liability claims involving AI in medicine.

LLMs exhibit the potential to revolutionise healthcare delivery, though inherent challenges exist. These include significantly different responses according to variations in query presentation and the absence of real-time updates or the lack of context sensitivity that may result in misleading outputs. It is, therefore, imperative that users of such AI systems grasp the inherent limitations. The duty of care expected from a doctor will probably be impacted by the advent of AI in medicine. The challenges arising from quantifying the error rates of AI and doctors, defining diligence, and dealing with the black-box effect of AI remain significant obstacles in the realm of medical jurisprudence. The differences in responses between a free version and a paid version also present a potential ethical aspect to consider.

Conflict of interest statement

The authors have no conflicts of interest to declare.

References

1. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med.* 2023;388:1233-9.

2. Mahendiran T, Thanou D, Senouf O, Meier D, Dayer N, Aminfar F, Auberson D, Raita O, Frossard P, Pagnoni M, Cook S, De Bruyne B, Muller O, Abbé E, Fournier S. Deep learning-based prediction of future myocardial infarction using invasive coronary angiography: a feasibility study. *Open Heart.* 2023;10:e002237.

3. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health*. 2023;4:279-81.

4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin, I, Attention Is All You Need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems 30 (Nips 2017). 2017. https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. (Last accessed: 26 June 2023).

5. Maron DJ, Hochman JS, Reynolds HR, Bangalore S, O'Brien SM, Boden WE, Chaitman BR, Senior R, López-Sendón J, Alexander KP, Lopes RD, Shaw LJ, Berger JS, Newman JD, Sidhu MS, Goodman SG, Ruzyllo W, Gosselin G, Maggioni AP, White HD, Bhargava B, Min JK, Mancini GBJ, Berman DS, Picard MH, Kwong RY, Ali ZA, Mark DB, Spertus JA, Krishnan MN, Elghamaz A, Moorthy N, Hueb WA, Demkow M, Mavromatis K, Bockeria O, Peteiro J, Miller TD, Szwed H, Doerr R, Keltai M, Selvanayagam JB, Steg PG, Held C, Kohsaka S, Mavromichalis S, Kirby R, Jeffries NO, Harrell FE Jr, Rockhold FW, Broderick S, Ferguson TB Jr, Williams DO, Harrington RA, Stone GW, Rosenberg Y; ISCHEMIA Research Group. Initial Invasive or Conservative Strategy for Stable Coronary Disease. *N Engl J Med.* 2020;382:1395-407.