ESC European Society of Cardiology
European Heart Journal - Digital Health (2023) **4**, 366–367
https://doi.org/10.1093/ehjdh/ztad043

LETTER TO THE EDITOR

ChatGPT fails the test of evidence-based medicine

Wilhelm Haverkamp [1,*], Jonathan Tennenbaum[2], and Nils Strodthoff[3]

[1]Department of Cardiology Campus Virchow Clinic of German Heart Center Charité, Charité—University Medicine Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; [2]Center for the Philosophy of Science, University of Lisbon, Lisbon, Portugal; and [3]Department of Health Sciences, School of Medicine and Health Services, Oldenburg University, Oldenburg, Germany

Online publish-ahead-of-print 13 July 2023

Commentary article to: 'Use of large language models for evidencebased cardiovascular medicine', by I. Skalidis *et al*. https://doi.org/10.1093/ehjdh/ztad041.

We have read with great interest the article 'ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?'.[1] The results of the reported investigation are remarkable, although not entirely surprising given earlier results with medical examinations.[2] To us, however, the euphoric tone of the article should be balanced by more reflection on the problematic aspects, including the risks, posed by the application of Chat Generative Pre-trained Transformer (ChatGPT) and similar systems in medicine, especially in sensitive areas such as cardiology. This is missing in the article.

We are concerned that enthusiasm about the capabilities of ChatGPT might convey the impression that ChatGPT is a competent and reliable source of information for clinical practice, which it certainly is not. In addition, as we shall indicate below, ChatGPT is far from meeting the standards of evidence-based medicine. At first glance, the authors' finding, that ChatGPT provided correct answers to about 60% of randomly sampled questions from the European Exam in Core Cardiology, is impressive. But what about the remaining 40% of questions, for which ChatGPT gave wrong or indeterminate answers? The wrong answers are just as important as the correct ones when judging the suitability of ChatGPT applications in medicine. Here an investigation is called for. Even 10% wrong answers could represent a significant medical risk, if physicians and others were to place too much trust in the system. Moreover, no one would regard a medical student, who answered 100% of the questions correctly, but has never seen a patient, as an authority to be consulted on cardiology!

In some respects, the success of ChatGPT resembles that of a clever student who has no real grasp of the subject but managed to smuggle a laptop into the examination room. The most obvious difference is that ChatGPT is orders of magnitude faster and integrates a vastly larger data base. There is growing recognition, in other domains, of the biases and risks that can arise from excessive trust or even dependence on ChatGPT, both regarding the accuracy of provided information, ethical and moral aspects, legal considerations, and so on. These issues are aggravated by the notorious lack of transparency of such systems. We have found that ChatGPT, when confronted with facts that contradict one of its statements, will often acknowledge 'Yes, you are right', but give no reason for its error, leaving open the question, whether a similar error might occur again. We have observed a tendency for ChatGPT to omit or 'smooth over' essential facts, required in order to make competent judgements concerning specific issues of practical importance.

The limitations and risks involved in utilizing ChatGPT become especially clear when its performance is measured against the standard of 'evidenced-based medicine', generally regarded as one of the pillars of present-day medicine and a major source of progress in recent times. Evidence-based medicine calls for integrating the best available evidence from scientific research, clinical studies and expertise, and the individual patient's desires and values.[3] Today, diagnostic and therapeutic options have to undergo approval processes that require them to meet the

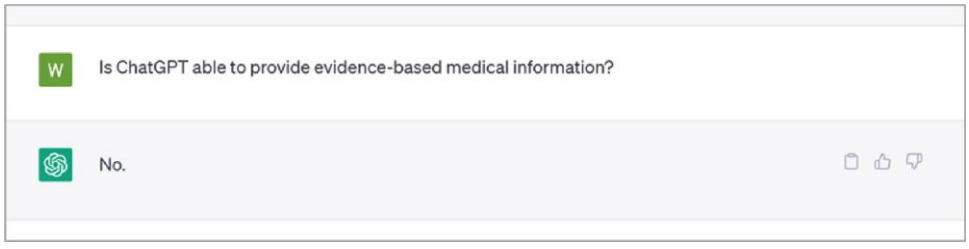

Figure 1 A question addressed to ChatGPT, and the answer given by the chatbot.

* Corresponding author. Email: wilhelm.haverkamp@dhzc-charite.de

criteria of evidence-based medicine.[4] This also applies to computer-based systems designed to assist physicians in making medical decisions. Interestingly, when ChatGPT is asked whether it is able to provide medical information according to the criteria of evidence-based medicine, the first answer, like many of ChatGPT's answers, is rather evasive. When pressed to provide a direct yes or no answer, the answer is 'No' (*Figure 1*). This aspect, which significantly restricts the utility of ChatGPT, is not given much emphasis in current discussions, but should be taken seriously.

It is important to recognize that the training data and methods used by the programme are not extensively documented. Indeed, most of these have been kept secret up to now. ChatGPT itself, when directly asked, is not able to specify the extent to which medical guidelines have been incorporated into its training. Standard of evidence-based medicine calls for transparency in the sources used and a careful selection of those sources. It should be clear which guidelines are utilized, allowing for comprehensibility. Additionally, the information needs to be kept up to date. The last cut-off date for the knowledge and training data for ChatGPT was September 2021.

Overall, ChatGPT has numerous gaps and limitations that require a thorough discussion among medical experts. This discussion should lead to a more objective evaluation of the clinical relevance und applicability of ChatGPT. Unfortunately, discussions concerning the capabilities of chatbots tend to be ambiguous and problematic. We think it is unlikely that ChatGPT, even in improved versions, will ever be able to provide competent, reliable, and trustworthy information on the level demanded by evidence-based medicine. Perhaps, more specialized systems will emerge in the future, which address the concerns voiced in this letter. Time will tell.

Funding

None declared.

Conflict of interest: None declared.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

References

1. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, *et al.* ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* 2023;**4**:279–281.
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al.* Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;**2**:e0000198.
3. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. *BMJ* 1996;**312**:71–72.
4. Zanca F, Brusasco C, Pesapane F, Kwade Z, Beckers R, Avanzo M. Regulatory aspects of the use of artificial intelligence medical software. *Semin Radiat Oncol* 2022;**32**: 432–441.