Study shows ChatGPT failed when challenging guideline for treating brain abscesses

April 29 2024

Credit: CC0 Public Domain

With artificial intelligence (AI) poised to become a fundamental part of clinical research and decision making, many still question the accuracy of ChatGPT, a sophisticated AI language model, to support complex diagnostic and treatment processes.

Now a new study, presented at this year's ESCMID Global Congress (formerly ECCMID) in Barcelona, Spain (27–30 April), which pitted ChatGPT against the ESCMID guideline for the management of brain abscesses, found that while ChatGPT seems able to give recommendations on key questions about diagnosis and treatment in most cases, some of the AI model's responses could put patients at risk.

The study was conducted by members of the ESCMID Study Group for Infectious Diseases of the Brain (ESGIB), and is published in *The Journal of Neurology.*

"Anything less than 100% is a failure when you're dealing with patient safety," says author Dr. Susanne Dyckhoff-Shen from LMU University Hospital Munich in Germany and a member of ESCMID. "While we are amazed by ChatGPT's knowledge on the management of brain abscesses, there are some key limitations when it comes to using the AI model as a medical device, including potential patient harm and the lack of transparency about which data are used to provide responses."

The ability of AI to rapidly assimilate, process, and interpret vast data sets offers tantalizing prospects. But are time-consuming processes to create medical guidelines still necessary, or could AI models trained on a wealth of scientific medical literature rival clinical experts in answering complex clinical questions?

Brain abscesses are a potentially life-threatening central nervous system (CNS) infection that require immediate identification and treatment to prevent severe neurological complications and even death.

Historically, the management of brain abscesses has been largely guided by clinical experience and limited studies, but in 2023 ESCMID fulfilled the need for a standardized approach by developing an international guideline.

To find out whether ChatGPT is able to professionally evaluate medical research and give scientifically valid recommendations, a European team of researchers tested the AI model to see whether it could accurately provide answers to 10 key questions on brain abscess diagnostics and treatment in comparison to the ESCMID guideline.

First, the researchers asked ChatGPT (version 4) to answer 10 questions that had been developed and appraised by the ESCMID committee for their brain abscess guideline without any additional information.

Then, ChatGPT was additionally primed with the text of the same scientific research articles that were used to develop the guideline before asking the same questions. This was done to see if ChatGPT could provide more aligned recommendations when given the same data used for guideline development.

The AI-generated responses were then compared to the recommendations of the ESCMID guideline by three independent infectious CNS disease experts for their clarity, alignment with the guideline, and patient risk.

Clear responses to most key questions

The researchers found that overall, for 17 out of 20 questions asked (with and without data input), ChatGPT's responses were clear on the management of patients with brain abscess, including grade of evidence and strength of recommendation, with clarity assessed at 80-90% (see link to poster in notes to editors).

However, the AI model did not provide clear enough answers to guide physicians on treatment decisions on withholding microbials until surgery and prophylactic antiepileptic treatment (questions 2 and 10).

More incorrect advice and risk to patients with data prompting

Without additional data input, ChatGPT's responses to 70% (7/10) of questions were very similar to the guideline recommendation. However, the AI model failed to come up with the correct advice on three questions relating to withholding microbials, consolidation therapy, and prophylactic antiepileptic treatment (questions 2, 8, and 10). Importantly, however, these incorrect responses would not have harmed patients (see link to poster in notes to editors).

Surprisingly, data input resulted in fewer correct answers (40%) including two recommendations that directly contradicted the guideline, that could have put patients at risk.

Question 6 about duration of antimicrobial therapy for bacterial brain abscess was answered by ChatGPT after data entry as "intravenous administration for about four weeks, followed by 12 weeks oral medication," but the ESCMID guideline recommends "a total duration of 6–8 weeks of intravenous antimicrobials…".

For question 7 about early transition to oral antimicrobials, after data input ChatGPT recommended that "an early switch to oral antibiotics during the first 14 days of treatment…seems to be associated with favorable outcomes in selected patients." However, the ESCMID guideline committee judged that there was insufficient evidence to provide a recommendation for this question.

In both cases, following ChatGPT's advice could have potentially led to patient harm.

"The fact that ChatGPT's recommendations were inferior after data entry might be due to an overvaluation of the few observational studies provided for key questions 6 and 7. For one of those, even the guideline committee was not able to give a recommendation as the evidence was insufficient to answer the question," says senior author Professor Mattias Klein from LMU University Hospital Munich in Germany and a member of the ESCMID committee which established the guideline.

"As the exact operating procedures of ChatGPT remain unclear, we speculate that while the AI model can process large amounts of data quickly, it may lack the ability to correctly classify and weigh the data based on their scientific quality. Moreover, it remains unclear which data are used for ChatGPT's responses as it does not disclose the sources of its answers, which risks dubious literature being used."

Dr. Dyckhoff-Shen adds, "It is alarming to think that patients could have come to harm if ChatGPT's advice on two key questions had been followed. The nuanced expertise of expert committees remains essential, especially to answer complex clinical queries. Blindly relying on AI could put patients at risk."

Nevertheless, the authors note that ChatGPT's knowledge was from before September 2021 and the questions in the study covered some extremely complex medical issues, some of which are controversial even among experts and for which hardly any robust data are available.

However, even when primed with the same research articles that were used to develop the ESCMID guideline, ChatGPT's advice aligned even less with the guideline. They recommend that the quality of ChatGPT should be reviewed on an ongoing basis following its evolution and further development.

The authors explain further that ChatGPT, like many AI models, has a cutoff date for the information it can access. For the current version as of today, it was last trained on data up until January 2022.

"This means that while it can provide responses based on a wide range of information, it doesn't have access to real-time data or events occurring after that date," says Dr. Dyckhoff-Shen.

"When we used ChatGPT for our study before August 2023, it was only trained on data up until September 2021. There was no possibility for us to get a more up-to-date version at that time because ChatGPT was not trained further yet. This is also the reason why we used a second approach by prompting ChatGPT with relevant scientific articles that were used by the ESGIB group to give recommendations in the ESCMID guideline so that we sort of 'manually' tried to get it more up-to-date.

"In the future, it would be interesting to re-assess ChatGPT's knowledge in the future after internal optimization processes. However, once the chatbot has access to the ESCMID guideline itself, it could just use the recommendations from the guideline thus rendering a comparison no longer useful."

More information: Susanne Dyckhoff-Shen et al, ChatGPT fails challenging the recent ESCMID brain abscess guideline, *Journal of Neurology* (2024). DOI: 10.1007/s00415-023-12168-1

Provided by European Society of Clinical Microbiology and Infectious

Diseases

Citation: Study shows ChatGPT failed when challenging guideline for treating brain abscesses (2024, April 29) retrieved 17 May 2024 from https://medicalxpress.com/news/2024-04-chatgpt-guideline-brain-abscesses.html