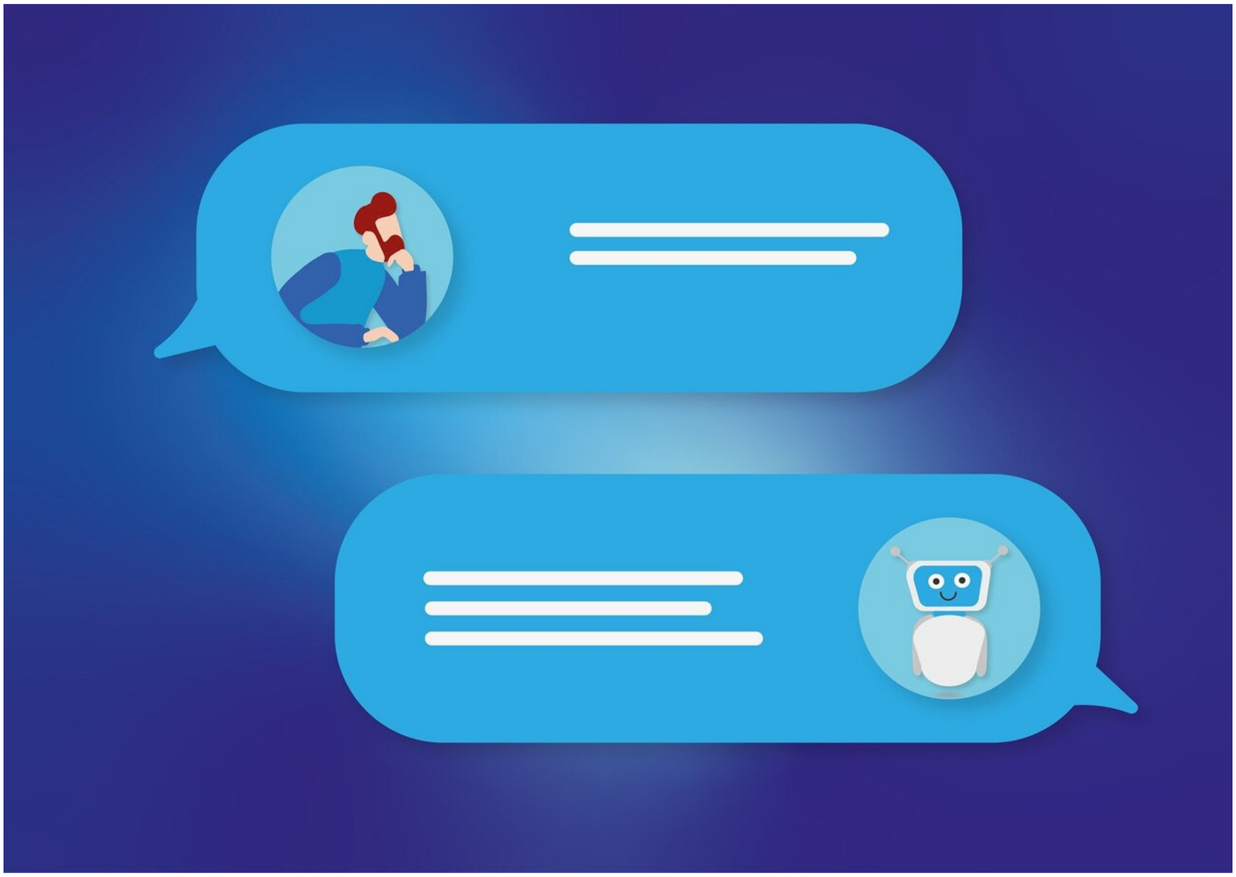
Medical Xpress

ChatGPT shows limited ability to recommend guidelines-based cancer treatments

August 24 2023



Credit: Pixabay/CC0 Public Domain

For many patients, the internet serves as a powerful tool for selfeducation on medical topics. With ChatGPT now at patients' fingertips, researchers from Brigham and Women's Hospital, a founding member of the Mass General Brigham health care system, assessed how consistently the artificial intelligence chatbot provides recommendations for cancer treatment that align with National Comprehensive Cancer Network (NCCN) guidelines.

Their findings, published in *JAMA Oncology*, show that in approximately one-third of cases, ChatGPT 3.5 provided an inappropriate ("non-concordant") recommendation, highlighting the need for awareness of the technology's limitations.

"Patients should feel empowered to educate themselves about their medical conditions, but they should always discuss with a clinician, and resources on the Internet should not be consulted in isolation," said corresponding author Danielle Bitterman, MD, of the Department of Radiation Oncology and the Artificial Intelligence in Medicine (AIM) Program of Mass General Brigham.

"ChatGPT responses can sound a lot like a human and can be quite convincing. But, when it comes to clinical decision-making, there are so many subtleties for every patient's unique situation. A right answer can be very nuanced, and not necessarily something ChatGPT or another large language model can provide."

The emergence of artificial intelligence tools in health has been groundbreaking and has the potential to positively reshape the continuum of care. Mass General Brigham, as one of the nation's top integrated academic health systems and largest innovation enterprises, is leading the way in conducting rigorous research on new and emerging technologies to inform the responsible incorporation of AI into care delivery, workforce support, and administrative processes.

Although medical decision-making can be influenced by many factors, Bitterman and colleagues chose to evaluate the extent to which ChatGPT's recommendations aligned with the NCCN guidelines, which are used by physicians at institutions across the country.

They focused on the three most common cancers (breast, prostate and lung cancer) and prompted ChatGPT to provide a treatment approach for each cancer based on the severity of the disease. In total, the researchers included 26 unique diagnosis descriptions and used four, slightly different prompts to ask ChatGPT to provide a treatment approach, generating a total of 104 prompts.

Nearly all responses (98%) included at least one treatment approach that agreed with NCCN guidelines. However, the researchers found that 34% of these responses also included one or more non-concordant recommendations, which were sometimes difficult to detect amidst otherwise sound guidance.

A non-concordant treatment recommendation was defined as one that was only partially correct; for example, for a locally advanced breast cancer, a recommendation of surgery alone, without mention of another therapy modality. Notably, complete agreement in scoring only occurred in 62% of cases, underscoring both the complexity of the NCCN guidelines themselves and the extent to which ChatGPT's output could be vague or difficult to interpret.

In 12.5% of cases, ChatGPT produced "hallucinations," or a treatment recommendation entirely absent from NCCN guidelines. These included recommendations of novel therapies, or curative therapies for non-curative cancers. The authors emphasized that this form of misinformation can incorrectly set patients' expectations about treatment and potentially impact the clinician-patient relationship.

Going forward, the researchers are exploring how well both patients and clinicians can distinguish between medical advice written by a clinician versus a large language model (LLM) like ChatGPT. They are also prompting ChatGPT with more detailed clinical cases to further evaluate its clinical knowledge.

The authors used GPT-3.5-turbo-0301, one of the largest models available at the time they conducted the study and the model class that is currently used in the open-access version of ChatGPT (a newer version, GPT-4, is only available with the paid subscription). They also used the 2021 NCCN guidelines, because GPT-3.5-turbo-0301 was developed using data up to September 2021. While results may vary if other LLMs and/or clinical guidelines are used, the researchers emphasize that many LLMs are similar in the way they are built and the limitations they possess.

"It is an open research question as to the extent LLMs provide consistent logical responses as oftentimes 'hallucinations' are observed," said first author Shan Chen, MS, of the AIM Program. "Users are likely to seek answers from the LLMs to educate themselves on health-related topics—-similarly to how Google searches have been used. At the same time, we need to raise awareness that LLMs are not the equivalent of trained medical professionals."

More information: Use of Artificial Intelligence Chatbots for Cancer Treatment Information, *JAMA Oncology* (2023). DOI: 10.1001/jamaoncol.2023.2954 , jamanetwork.com/journals/jamao … /jamaoncol.2023.2954

Provided by Brigham and Women's Hospital

Citation: ChatGPT shows limited ability to recommend guidelines-based cancer treatments (2023, August 24) retrieved 2 May 2024 from https://medicalxpress.com/news/2023-08-chatgpt-limited-ability-guidelines-based-cancer.html