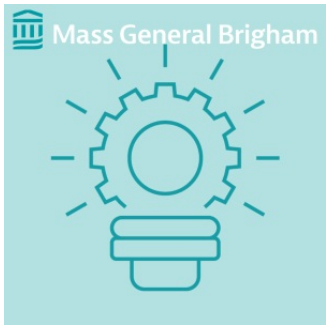
HealthManagement
Promoting Management and Leadership

ChatGPT Shows ‘Impressive’ Accuracy in Clinical Decision Making



- Mass General Brigham research highlights potential for generative AI to increase access and efficiency in healthcare
- Investigators found ChatGPT to be nearly 72 percent accurate across all medical specialties and phases of clinical care, and 77 percent accurate in making final diagnoses

A new study led by investigators from Mass General Brigham has found that ChatGPT was about 72 percent accurate in overall clinical decision making, from coming up with possible diagnoses to making final diagnoses and care management decisions. The large-language model (LLM) artificial intelligence chatbot performed equally well in both primary care and emergency settings across all medical specialties. The research team’s results are published in the *Journal of Medical Internet Research* .

“Our paper comprehensively assesses decision support via ChatGPT from the very beginning of working with a patient through the entire care scenario, from differential diagnosis all the way through testing, diagnosis, and management,” said corresponding author Marc Succi, MD, associate chair of innovation and commercialization and strategic innovation leader at Mass General Brigham and executive director of the MESH Incubator. “No real benchmarks exists, but we estimate this performance to be at the level of someone who has just graduated from medical school, such as an intern or resident. This tells us that LLMs in general have the potential to be an augmenting tool for the practice of medicine and support clinical decision making with impressive accuracy.”

Changes in artificial intelligence technology are occurring at a fast pace and transforming many industries, including health care. But the capacity of LLMs to assist in the full scope of clinical care has not yet been studied. In this comprehensive, cross-specialty study of how LLMs could be used in clinical advisement and decision making, Succi and his team tested the hypothesis that ChatGPT would be able to work through an entire clinical encounter with a patient and recommend a diagnostic workup, decide the clinical management course, and ultimately make the final diagnosis.

The study was done by pasting successive portions of 36 standardized, published clinical vignettes into ChatGPT. The tool first was asked to come up with a set of possible, or differential, diagnoses based on the patient’s initial information, which included age, gender, symptoms, and whether the case was an emergency. ChatGPT was then given additional pieces of information and asked to make management decisions as well as give a final diagnosis—simulating the entire process of seeing a real patient. The team compared ChatGPT’s accuracy on differential diagnosis, diagnostic testing, final diagnosis, and management in a structured blinded process, awarding points for correct answers and using linear regressions to assess the relationship between ChatGPT’s performance and the vignette’s demographic information.

The researchers found that overall, ChatGPT was about 72 percent accurate and that it was best in making a final diagnosis, where it was 77 percent accurate. It was lowest-performing in making differential diagnoses, where it was only 60 percent accurate. And it was only 68 percent accurate in clinical management decisions, such as figuring out what medications to treat the patient with after arriving at the correct diagnosis. Other notable findings from the study included that ChatGPT’s answers did not show gender bias and that its overall performance was steady across both primary and emergency care.

“ChatGPT struggled with differential diagnosis, which is the meat and potatoes of medicine when a physician has to figure out what to do,” said Succi. “That is important because it tells us where physicians are truly experts and adding the most value—in the early stages of patient care with little presenting information, when a list of possible diagnoses is needed.”

The authors note that before tools like ChatGPT can be considered for integration into clinical care, more benchmark research and regulatory guidance is needed. Next, Succi’s team is looking at whether AI tools can improve patient care and outcomes in hospitals’ resource-constrained areas.

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.

"Mass General Brigham sees great promise for LLMs to help improve care delivery and clinician experience," said co-author Adam Landman, MD, MS, MIS, MHS, chief information officer and senior vice president of digital at Mass General Brigham. "We are currently evaluating LLM solutions that assist with clinical documentation and draft responses to patient messages with focus on understanding their accuracy, reliability, safety, and equity. Rigorous studies like this one are needed before we integrate LLM tools into clinical care."

Source: Mass General Brigham

Published on : Tue, 22 Aug 2023