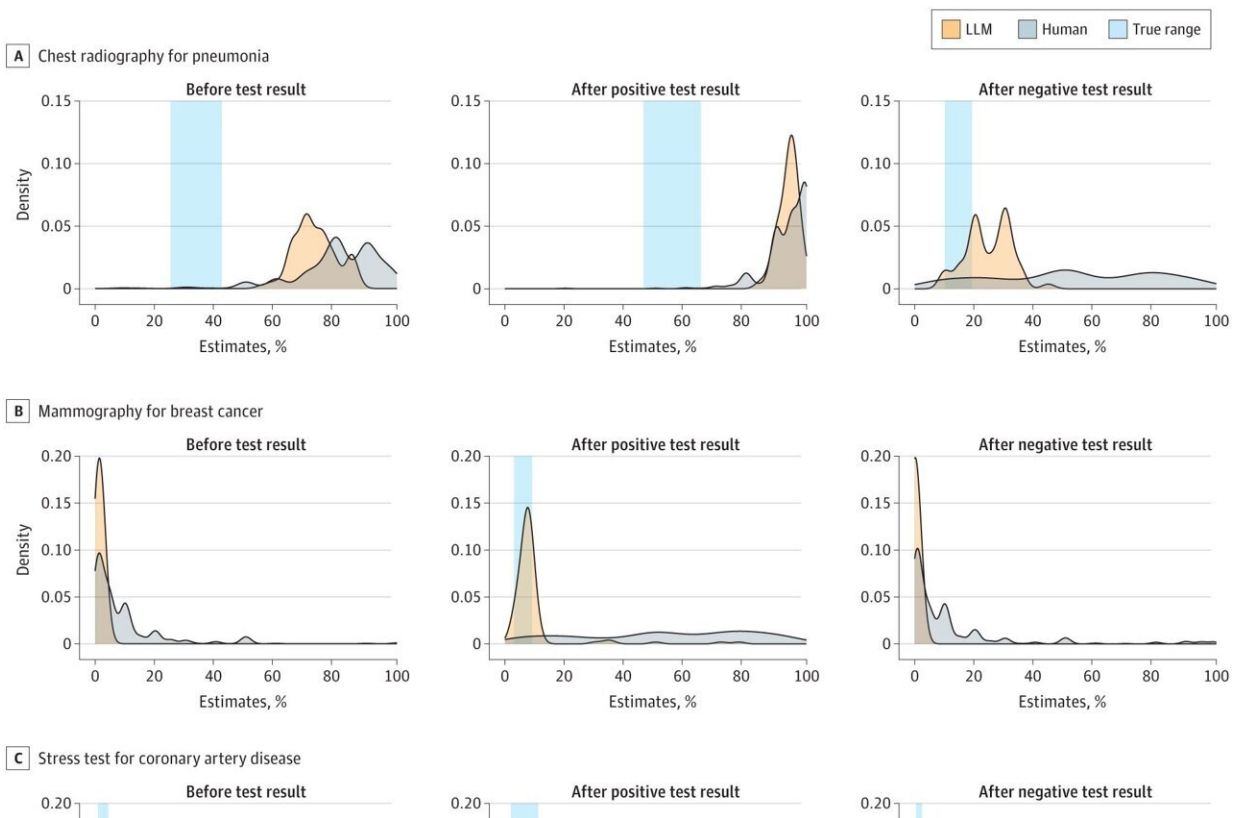# AI chatbot shows potential as diagnostic partner, researchers find

December 11 2023

Density Plots of the Distributions of Responses. LLM indicates large language model. Credit: *JAMA Network Open* (2023). DOI: 10.1001/jamanetworkopen.2023.47075

Physician-investigators at Beth Israel Deaconess Medical Center (BIDMC) compared a chatbot's probabilistic reasoning to that of human

clinicians. The findings, [published](#) in *JAMA Network Open*, suggest that artificial intelligence could serve as useful clinical decision support tools for physicians.

"Humans struggle with probabilistic reasoning, the practice of making decisions based on calculating odds," said the study's corresponding author, Adam Rodman, MD, an internal medicine physician and investigator in the Department of Medicine at BIDMC.

"Probabilistic reasoning is one of several components of making a diagnosis, which is an incredibly complex process that uses a variety of different cognitive strategies. We chose to evaluate probabilistic reasoning in isolation because it is a well-known area where humans could use support."

Basing their study on a previously published national survey of more than 550 practitioners performing probabilistic reasoning on five [medical cases](#), Rodman and colleagues fed the publicly available Large Language Model (LLM), Chat GPT-4, the same series of cases and ran an identical prompt 100 times to generate a range of responses.

The chatbot—just like the practitioners before them—was tasked with estimating the likelihood of a given diagnosis based on patients' presentation. Then, given test results such as chest radiography for pneumonia, mammography for [breast cancer](#), [stress test](#) for [coronary artery disease](#), and a urine culture for [urinary tract infection](#), the chatbot program updated its estimates.

When test results were positive, it was something of a draw; the chatbot was more accurate in making diagnoses than the humans in two cases, similarly accurate in two cases and less accurate in one case. But when tests came back negative, the chatbot shone, demonstrating more accuracy in making diagnoses than humans in all five cases.

"Humans sometimes feel the risk is higher than it is after a negative test result, which can lead to overtreatment, more tests, and too many medications," said Rodman.

But Rodman is less interested in how chatbots and humans perform toe-to-toe than in how highly skilled physicians' performance might change in response to having these new supportive technologies available to them in the clinic, added Rodman. He and his colleagues are looking into it.

"LLMs can't access the outside world—they aren't calculating probabilities the way that epidemiologists, or even poker players, do. What they're doing has a lot more in common with how humans make spot probabilistic decisions," he said.

"But that's what is exciting. Even if imperfect, their ease of use and ability to be integrated into clinical workflows could theoretically make humans make better decisions," he said. "Future research into collective human and artificial intelligence is sorely needed."

More information: Adam Rodman et al, Artificial Intelligence vs Clinician Performance in Estimating Probabilities of Diagnoses Before and After Testing, *JAMA Network Open* (2023). [DOI: 10.1001/jamanetworkopen.2023.47075](DOI:10.1001/jamanetworkopen.2023.47075)

Provided by Beth Israel Deaconess Medical Center