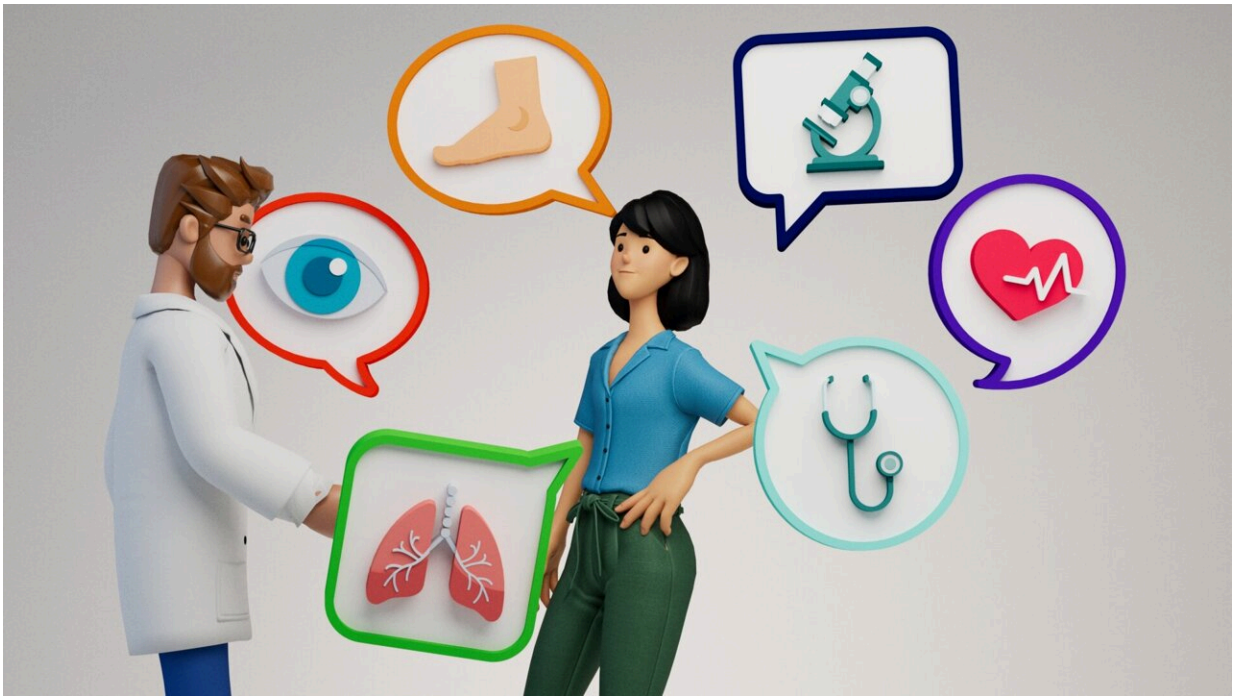
Leading AI models struggle to identify genetic conditions from patient-written descriptions, researchers find

August 14 2024



Clinical geneticists listen to patients describe their conditions as part of making diagnoses. While AI tools can interpret textbook-like medical questions, they struggle to analyze this same information when provided directly by patients. Credit: Ernesto del Aguila III, National Human Genome Research Institute

National Institutes of Health (NIH) researchers have discovered that

while artificial intelligence (AI) tools can make accurate diagnoses from textbook-like descriptions of genetic diseases, the tools are significantly less accurate when analyzing summaries written by patients about their own health.

These findings, reported in the *American Journal of Human Genetics*, demonstrate the need to improve these AI tools before they can be applied in health care settings to help make diagnoses and answer patient questions.

The researchers studied a type of AI known as a large language model, which is trained on massive amounts of text-based data. These models have the potential to be very helpful in medicine due to their ability to analyze and respond to questions and their often user-friendly interfaces.

"We may not always think of it this way, but so much of medicine is words-based," said Ben Solomon, M.D., senior author of the study and clinical director at the NIH's National Human Genome Research Institute (NHGRI).

"For example, electronic health records and the conversations between doctors and patients all consist of words. Large language models have been a huge leap forward for AI, and being able to analyze words in a clinically useful way could be incredibly transformational."

The researchers tested 10 different large language models, including two recent versions of ChatGPT. Drawing from medical textbooks and other reference materials, the researchers designed questions about 63 different genetic conditions. These included some well-known conditions, such as sickle cell anemia, cystic fibrosis and Marfan syndrome, as well as many rare genetic conditions.

These conditions can show up in a variety of ways among different

patients, and the researchers aimed to capture some of the most common possible symptoms.

They selected three to five symptoms for each condition and generated questions phrased in a standard format, "I have X, Y and Z symptoms. What's the most likely genetic condition?"

When presented with these questions, the large language models ranged widely in their ability to point to the correct genetic diagnosis, with initial accuracies between 21% and 90%. The best performing model was GPT-4, one of the latest versions of ChatGPT.

The success of the models generally corresponded with their size, meaning the amount of data the models were trained on. The smallest models have several billion parameters to draw from, while the largest have over a trillion.

For many of the lower-performing models, the researchers were able to improve the accuracy over subsequent experiments, and overall, the models still delivered more accurate responses than non-AI technologies, including a standard Google search.

The researchers optimized and tested the models in various ways, including replacing medical terms with more common language. For example, instead of saying a child has "macrocephaly," the question would say the child has "a big head," more closely reflecting how patients or caregivers might describe a symptom to a doctor.

Overall, the models' accuracy decreased when medical descriptions were removed. However, seven out of 10 of the models were still more accurate than Google searches when using common language.

"It's important that people without medical knowledge can use these

tools," said Kendall Flaharty, an NHGRI postbaccalaureate fellow who led the study.

"There are not very many clinical geneticists in the world, and in some states and countries, people have no access to these specialists. AI tools could help people get some of their questions answered without waiting years for an appointment."

To test the large language models' efficacy with information from real patients, the researchers asked patients from the NIH Clinical Center to provide short write-ups about their own genetic conditions and symptoms. These descriptions ranged from a sentence to a few paragraphs and were also more variable in style and content compared to the textbook-like questions.

When presented with these descriptions from real patients, the best-performing model made accurate diagnoses only 21% of the time. Many models performed much worse, even as low as 1% accurate.

The researchers expected the patient-written summaries to be more challenging because patients at the NIH Clinical Center often have extremely rare conditions. The models may therefore not have sufficient information about these conditions to make diagnoses.

However, the accuracies improved when the researchers wrote standardized questions about the same ultra-rare genetic conditions found among the NIH patients. This indicates that the variable phrasing and format of the patient write-ups was difficult for the models to interpret, perhaps because the models are trained on textbooks and other reference materials that tend to be more concise and standardized.

"For these models to be clinically useful in the future, we need more data, and those data need to reflect the diversity of patients," said Dr.

Solomon.

"Not only do we need to represent all known medical conditions, but also variation in age, race, gender, cultural background and so on, so that the data capture the diversity of patient experiences. Then these models can learn how different people may talk about their conditions."

Beyond demonstrating areas of improvement, this study highlights the current limitations of large language models and the continued need for human oversight when AI is applied in health care.

"These technologies are already rolling out in clinical settings," Dr. Solomon added. "The biggest questions are no longer about whether clinicians will use AI, but where and how clinicians should use AI, and where should we not use AI to take the best possible care of our patients."

More information: Evaluating Large Language Models on Medical, Lay Language, and Self-Reported Descriptions of Genetic Conditions, *The American Journal of Human Genetics* (2024). DOI: 10.1016/j.ajhg.2024.07.011.
www.cell.com/ajhg/fulltext/S0002-9297(24)00255-6

Provided by NIH/National Human Genome Research Institute