VIEWPOINT

AI IN MEDICINE

Artificial Intelligence in Clinical Diagnosis
Opportunities, Challenges, and Hype

Prathit A. Kulkarni, MD
Medical Care Line, Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas; and Infectious Diseases Section, Department of Medicine, Baylor College of Medicine, Houston, Texas.

Hardeep Singh, MD, MPH
Center for Innovations in Quality, Effectiveness, and Safety (IQuESt), Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, Texas; and Health Services Research Section, Department of Medicine, Baylor College of Medicine, Houston, Texas.

Viewpoint
pages 309, 311, 313, and 315

Corresponding Author: Hardeep Singh, MD, MPH, Michael E. DeBakey VA Medical Center, 2002 Holcombe Blvd, MS: 152, Houston, TX 77030 (hardeeps@bcm.edu).

ChatGPT, a generative artificial intelligence (AI) chatbot, has recently been hailed as a promising tool to improve health care quality. One study compared output from the AI chatbot for medical questions with answers from physicians[1]; other studies have evaluated the AI chatbot's responses to sample clinical vignettes.[2,3] A foundational aspect of high-quality health care—making a correct and timely diagnosis—remains a challenge in modern medicine despite decades of technological advances.[4,5] Therefore, any emerging technology with potential to reduce diagnostic errors warrants serious examination.

Recent literature provides some suggestions as to what role AI and the AI chatbot may have in assisting with diagnosis.[6-8] However, clinical diagnosis is both an art and a science, and is more challenging for AI to optimize than visual diagnostic interpretation, such as radiographic and pathologic diagnosis. Here, we provide a realistic overview of generative AI's role in clinical diagnosis to clarify hype, strengths, challenges, and future opportunities.

Diagnostic dilemmas are common in clinical medicine. Arriving at a patient's final diagnosis is a process that evolves over time and can include periods of uncertainty. One potential use of AI is to identify rare diagnoses or unusual presentations in particularly complex cases. Historically, clinicians relied on their memory to recall rare conditions or unusual clinical presentations if a patient had a certain constellation of signs, symptoms, and laboratory and radiographic findings. In modern times, clinicians often use internet search engines, online resources, or differential-diagnosis generators to assist with unexplained clinical presentations. Pending evaluation of effectiveness, the AI chatbot could become an additional tool to help with diagnostic dilemmas.

> [S]upport from AI when integrated effectively into clinician workflow can potentially improve quality of care…However, in its current form, the advantages of AI do not effectively account for certain real-world aspects of making a clinical diagnosis.

Conversely, AI applications must also address common diagnoses because most harm from diagnostic errors involves common medical conditions.[4,5] Effective use of AI chatbots could ensure clinicians have considered all common conditions that might explain a patient's presentation. For example, dyspnea on exertion, anemia, and hyponatremia are classic general medicine problems, but clinicians often rely on their memory when performing their diagnostic evaluation, a fallible approach. Additionally, laboratory or radiographic findings might not be interpreted correctly by clinicians. AI chat platforms can be consulted, potentially in real time, to ensure that obvious diagnostic possibilities have not been overlooked. Ideally, the platform would be embedded into the electronic health record (EHR) to make this consultation highly efficient.

AI also has the advantage of being able to scan a patient's medical record faster than a person can. Clinicians often spend long periods trying to decipher a patient's story and longitudinal journey by clicking through scores of notes, laboratory trends, radiology and pathology reports, and additional diagnostic data. With associated visualization platforms, AI could display these data in a more intuitive way and potentially assist with nuanced interpretation of such cumulative historical data.

Despite these potential benefits of AI, fundamental limitations and challenges require careful consideration as AI is further integrated into medical care. Of paramount significance is that the accuracy of data about the clinical case entered into the chatbot will determine the differential-diagnosis output. However, research has demonstrated that many diagnostic errors are related to core clinical skills, including history taking, physical examination, and other data-gathering activities.[4] Information gathered from these actions serves as the basis of what an AI chatbot would use to assist with diagnosis, and this information might be incomplete or incorrect.

Additionally, patient histories are, by nature, subjective. A patient who describes their pain as "stabbing" or as 10 out of 10 on a pain scale might give subtle cues that provide important context but which only a person can detect. Subjective and varying reports from patients are more difficult for AI to consistently use in an algorithmic way. Many patients also describe a myriad of symptoms. Sorting through the relative importance of each of these is often best done by a skillful clinician who can relate to the patient emotionally, understand subtle nuances, observe minor clues, and deduce the most salient features of what the patient is describing. Some patients might also be more reticent to describe certain symptoms or other medical history. It often requires a compassionate clinician to

make an interpersonal human bond with a patient to elicit a complete history. AI, at least in its current form, is likely unable to do this, and many scientific advances would be required for it to gain this ability in the future.

The accuracy of physical-examination findings can also pose limits to AI's effectiveness. A chatbot might arrive at a more accurate diagnosis if it knows the patient's jugular venous pressure is elevated or there is suprapubic tenderness. However, this assumes that the clinician has accurately assessed the patient's jugular venous pressure or has correctly detected suprapubic tenderness. This fundamental aspect of patient assessment—the accuracy of the practitioner's clinical examination—will remain paramount even with AI's increasing role in supporting clinical diagnosis.

Next, although AI might be faster than a person at scanning a medical record, this advantage also relies on the accuracy of the information scanned. EHRs have an abundance of inaccurate information, an important driver of which is copying and pasting.[9] Some of this inaccurate information includes key common diagnoses incorrectly added to the problem list, such as heart failure, chronic obstructive pulmonary disease, and chronic kidney disease. Other inaccurate information includes human errors in notes, such as incorrect dates, time frames, or other important patient information. Importantly, the quality of diagnostic data in the EHR will directly determine the accuracy of the AI-generated diagnosis.

Another significant challenge in advancing AI diagnosis is judging the accuracy of the diagnosis itself. Many medical diagnoses do not have a final confirmatory test. For example, common diagnoses, such as cellulitis, pneumonia, and decompensated heart failure, rely on clinicians agreeing on a final diagnosis, which does not occur in all clinical cases. AI is likely most helpful in diagnosing conditions that are objectively or numerically true, such as a diagnosis of acute kidney injury when a patient's creatinine level has risen from 1.0 mg/dL to 3.5 mg/dL over 3 days. However, determining whether a patient truly had pneumonia is less straightforward. The general constellation of findings for this diagnosis includes systemic symptoms such as fever, cough with sputum production, and neutrophilic leukocytosis. However, the presence or absence of any one of these findings does not definitively confirm or refute the diagnosis. Assessing the accuracy of AI's ability to diagnose differs from examining its ability to beat a person at chess or solving mathematical problems. In these classic examples of AI success, everyone agrees on the accuracy of the final product. Conversely, the final diagnostic "answer" might remain uncertain in many instances, even if the patient improves. Clinical diagnosis often involves gathering and interpreting imprecise and nonnumerical diagnostic data and includes the possibility that no objective pathognomonic test exists for the diagnosis in question. Thus, judging AI accuracy in clinical diagnosis is particularly challenging outside of well-structured case vignettes.

Finally, consider a hypothetical scenario in which a chatbot has finally achieved the ability to diagnose better than people. Even if this were achieved, at the present time, AI cannot replicate in-person, human-to-human empathy. A recent study suggested that the AI chatbot was more empathetic than physicians, but an important caveat was this study considered only written interactions.[1] Most patients want their treating clinician to be accurate in their diagnosis but also empathetic during verbal and in-person communication. This core aspect of the patient-clinician interaction at its most fundamental and human level is expected even in modern times, and it is unlikely AI will replace in-person human connection soon.

In sum, support from AI when integrated effectively into clinician workflow can potentially improve quality of care, including diagnosis. However, in its current form, the advantages of AI do not effectively account for certain real-world aspects of making a clinical diagnosis. These factors include fundamental clinical skills required for clinical diagnosis, such as history taking and physical examination, accuracy of EHR data, uncertainty in accuracy of a final diagnosis, and in-person empathy and communication. These considerations must be addressed as AI is used to improve diagnosis in the 21st century.

ARTICLE INFORMATION

Published Online: July 6, 2023. doi:10.1001/jama.2023.11440

Conflict of Interest Disclosures: Dr Kulkarni reported receiving grants from the Gordon and Betty Moore Foundation/Society of Bedside Medicine and Vessel Health Inc. Dr Singh reported receiving grants from the Houston Veterans Administration (VA) Health Services Research and Development (HSR&D) Center for Innovations in Quality, Effectiveness, and Safety (CIN13-413), the VA HSR&D Service (IIR17-127), the VA National Center for Patient Safety, the Agency for Healthcare Research and Quality (R18 HS029347 and RO1HS27363), and the Gordon and Betty Moore Foundation and serving as co-chair of The Leapfrog Group's National Advisory Group for Recognizing Excellence in Diagnosis.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

REFERENCES

1. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

2. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;20(4):3378. doi:10.3390/ijerph20043378

3. Hailu R, Beam A, Mehrotra A. ChatGPT-assisted diagnosis: is the future suddenly here? STAT. February 13, 2023. Accessed May 5, 2023. https://www.statnews.com/2023/02/13/chatgpt-assisted-diagnosis/

4. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. *JAMA Intern Med*. 2013;173(6):418-425. doi:10.1001/jamainternmed.2013.2777

5. Schiff GD, Hasan O, Kim S, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Arch Intern Med*. 2009;169(20):1881-1887. doi:10.1001/archinternmed.2009.333

6. Haug CJ, Drazen JM. artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388(13):1201-1208. doi:10.1056/NEJMra2302038

7. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

8. Haupt CE, Marks M. AI-generated medical advice: GPT and beyond. *JAMA*. 2023;329(16):1349-1350. doi:10.1001/jama.2023.5321

9. Vaghani V, Wei L, Mushtaq U, Sittig DF, Bradford A, Singh H. Validation of an electronic trigger to measure missed diagnosis of stroke in emergency departments. *J Am Med Inform Assoc*. 2021;28(10):2202-2211. doi:10.1093/jamia/ocab121