Since the launch in November 2022 of the large language model (LLM) known as ChatGPT (OpenAI), health care organizations are rapidly embracing this new technology, and many now believe that artificial intelligence (AI), with LLMs and other forms of generative AI leading the way, will swiftly transform medical practice. Chatbots are already helping to redesign the medical workflow, schedule appointments, and even answer routine patient questions. Still very much an open question, however, is the role of these new technologies for improving medical diagnosis.

Optimists see value in an instantaneous review of a voluminous medical record, machine learning interpretation of diagnostic scans, and analysis of diverse data ranging from published literature to patient breathing patterns. Skeptics, on the other hand, point out that definitive diagnosis often involves a careful and nuanced history and bedside physical examination, neither of which come naturally to a machine, and neither of which are adequately quantified by human clinicians to be digitally digested. What is rarely acknowledged, however, is that the ability of AI tools (such as LLMs) to assist with medical diagnosis is not merely an answer to be discovered—it is an outcome to be influenced through policy.

Artificial intelligence (the branch of computer science concerned with endowing computers with the ability to simulate intelligent human behavior) incorporates multiple types of technologies (of which, LLMs are just 1 type). The most complex cognitive task in medicine is the act of diagnosing the cause of a patient’s symptoms. Errors in diagnosis have been estimated to account for nearly 800,000 deaths or permanent disabilities each year in the US, more than 80% of which are associated with cognitive errors or clinical reasoning failures. This burden creates a unique quality improvement opportunity for AI-based systems to save lives at public health scale.

Achieving this goal is not yet in reach. Large language models (such as ChatGPT) produce written text that appears to be written by a human (including using medical language that seems quite expert), but these AI models are not designed to distinguish fact from fiction. Sometimes they produce false statements unfaithful to the source text, nonsensical concepts, citations to fabricated sources, and the like—a phenomenon that has been anthropomorphically described as the system “hallucinating,” or, more properly, “confabulating.”

The lack of intrinsic truth in AI-generated text that seems expert (but may not be) makes LLMs a poor tool for guiding medical diagnosis. By contrast, AI-driven expert systems, which are modeled after human experts and trained in a supervised fashion on more narrowly defined, gold standard datasets (eg, a large databank of digital photographs of melanoma skin lesions with definitive diagnoses made by expert pathologists), are designed for accuracy and have been shown to rival or even exceed the diagnostic accuracy of experienced clinicians (eg, dermatologists viewing the same images). Pairing AI-driven expert systems with LLMs may ultimately offer the best of both worlds.

Even though AI, machine learning, and LLMs have matured to the point where the technologies themselves are no longer an impediment to success, a critical constraint persists—specifically, the lack of adequate gold standard data sources needed to train AI systems to achieve excellence in diagnosis. If AI systems are trained on faulty data, they will yield faulty results. Artificial intelligence systems that learn on faulty data will generally make the same mistakes that humans make or worse. For example, if electronic health record data are used to train AI systems, an AI system will (at a minimum) replicate current diagnostic errors or implicit human biases; in a worst-case scenario, AI systems will not only be frequently wrong in their diagnostic recommendations, but also will...
undermine clinical training. If AI-based diagnostic systems are deployed without adequate testing and monitoring after deployment, the quality of medical diagnosis can be expected to decline and patients will suffer.

Funding agencies have a critical role to play. The National Institutes of Health, the Agency for Healthcare Research and Quality, and others can invest in the development of databases specifically for the purpose of training AI systems in diagnosis. Such datasets for “visual diagnosis” based solely on interpretation of medical images (as in radiology, ophthalmology, pathology, and dermatology) are already being developed. However, comparable initiatives for the bulk of clinical diagnostic encounters across care settings are either nascent or do not exist. Training diagnostically accurate AI systems requires high-quality data at both the front end (patient demographics, symptoms, signs, and laboratory and radiographic findings) and at the back end (accurate final diagnoses, treatment effects, and morbidity or mortal outcomes).

These comprehensive and accurate datasets often cannot be extracted from routine practice. For example, a common clinical conundrum is the patient in the emergency department complaining of dizziness. The cause could be nonthreatening (such as benign positional vertigo) or dangerous (such as a brain stem stroke). The most useful diagnostic information comes from patient symptoms and bedside eye movement examinations, but these findings are rarely documented correctly. For AI systems to aid in differentiating inner ear disease from stroke, underlying datasets must include a reliable record of these diagnostically relevant patient histories and physical examinations, as well as key diagnostic test results and accurate final diagnoses. Such datasets must be built intentionally.

In addition to the role of funding agencies, regulators such as the US Food and Drug Administration should engage to set standards for the comprehensive and accurate databases needed to train AI diagnostic systems. Sponsors of such systems should also be required to have safety systems in place to check their outputs, with requirements for monitoring diagnostic errors and reporting serious harms. Payers (such as the Centers for Medicare & Medicaid Services) should create payment policies that require monitoring of diagnostic accuracy and pay more for better patient outcomes in diagnosis to incentivize adoption of diagnostic AI systems demonstrated to help patients.

The Turing test reveals whether an AI system exhibits intelligent behavior indistinguishable from that of a human. Researchers have reported that LLMs can pass the Turing test. However, for a medical diagnosis, it is obviously not enough for an AI system to approximate the cognitive skills of a layperson; it is also insufficient to be equivalent to an average clinician, given the current rate of diagnostic errors in clinical practice. An appropriate diagnostic excellence Turing test would determine whether an AI diagnostic system performs as well or better than top-performing clinicians in their areas of greatest diagnostic expertise.

A national effort to improve the quality of AI diagnostic systems could improve patient care over time. The goal should be to leverage the complementary strengths of humans and computers to yield better diagnoses than either one alone. Absent these steps, the risks of AI for diagnosis may overshadow the benefits, leading to worse health outcomes, biased care, and a less-skilled clinician workforce. The alternative future of diagnostic excellence through AI starts with recognizing the challenge. Dedicated resources and thoughtful public policies can drive substantial improvements in the care and health of patients.
Author Affiliations: Armstrong Institute Center for Diagnostic Excellence, School of Medicine, Johns Hopkins University, Baltimore, Maryland (Newman-Toker); Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Sharfstein).

Conflict of Interest Disclosures: Dr Newman-Toker reported receiving grants and nonfinancial support from Natus Inc/Otometrics, receiving nonfinancial support from Interacoustics, and receiving grants from the National Institutes of Health. No other disclosures were reported.

REFERENCES