How to Safely Integrate Large Language Models Into Health Care
Scott Gottlieb, MD; Lauren Silvis, JD

Long before ChatGPT popularized the use of artificial intelligence (AI) for everyday interactions, AI devices were being used to enhance medical care. The primary applications in health care have been machine learning tools for interpreting clinical images and diagnostic test results in fields such as radiology, pathology, and cardiology. The US Food and Drug Administration (FDA) has authorized more than 500 AI-based medical devices.¹ In addition, many AI tools have been incorporated into clinical decision support software to help inform the delivery of medical care.

Some clinical decision support tools are not regulated as medical devices even though they assist in interpreting data, often within electronic medical records, and provide valuable prompts to clinicians by using algorithms to recognize patterns in data and generate recommendations. Examples of tools that might not be actively regulated by the FDA could include software that analyzes a patient’s family history, genetic profile, and prior test results to recommend more frequent colonoscopies to screen for colon cancer. In these instances, the final decision is dependent on the clinician’s judgment, and the clinician is able to assess the independent variables that informed the treatment recommendation.

However, a growing number of clinical decision support tools are regulated as medical devices because they generate outputs that can drive a clinical decision, but do not allow the clinician to exercise complete control over the data. In such cases, the output of the tool may be based on its own pattern recognition and generated from analyzing very large data sets. Nonetheless, such tools are often trained on closed data sets like laboratory results, medical images, and confirmed diagnoses in which the findings have been rigorously confirmed. With the ability to verify the accuracy of the data on which these models were trained, regulators can apply the traditional regulatory framework with more certainty about the validity of results produced by these tools.² We should expect increasing efforts to develop more of these regulated tools.

A second category of AI devices that has garnered recent attention, but has had limited practical applications in medical care thus far, are large language models (LLMs) that encompass natural language processing. Large language models are a specific subset of machine learning designed to understand and generate human language by enabling computers to convert language and unstructured text into machine-readable, organized data. By incorporating data from countless individual decisions, these models can mimic a person’s responses by calculating the probability of each potential response and then selecting the most appropriate one—either by choosing the response with the highest likelihood of being correct or by sampling from a distribution of likely outcomes. A key difference between machine learning and LLMs is functionality in that machine learning devices are trained to perform specific tasks, whereas LLMs can understand and generate free-form text, potentially making them effective tools for expanding interactions with patients. For example, LLMs may generate prompts to obtain additional information about a patient’s symptoms or response to therapy.

Artificial intelligence tools based on LLMs are likely to permeate medicine more deeply than traditional machine learning–based technologies. Health care was an early sector to make use of predictive machine learning tools, and it is poised to be at the forefront of making practical use of LLMs. The seemingly “human” element of ChatGPT—its ability to generate comprehensive, intelligible text in response to complex inquiries—offers a promising opportunity for advancing the delivery of health care. We can envision an LLM-based AI tool that can decipher a vast array of health care data.
and provide in-depth insights into the diagnosis and management of patients. The best way to achieve effective adoption of these models is to proceed deliberately—the application of these LLMs should align with the steadily improving accuracy and rigor of these tools.

Large language models are poised to assist physicians in diagnosing, treating, and managing diseases for which we have abundant data to develop these AI tools and gauge their effectiveness. They could be especially valuable in managing high-volume and routine encounters and by providing more ongoing support to patients than what clinicians are able to achieve unassisted.

Large language models will be most effective in situations where they can enhance the ability to have high-value interactions with patients, especially in clinical settings when there is evidence that more frequent patient engagement can improve outcomes, and where clear thresholds indicate when a seemingly routine interaction becomes complex enough to necessitate the intervention of a physician. For these reasons, successful integration of LLMs will most likely be for conditions in which interventions are supported by comprehensive longitudinal data from clinical trials, observational studies, and various real-world data sources. In these scenarios, AI-guided interactions can be readily compared with decades of data and well-established clinical guidelines.

For example, consider the routine management of heart disease or diabetes. Large language models that provide daily recommendations for these conditions can offer patients continuous guidance to improve disease management and outcomes. This support would be especially helpful in situations when a similar volume of interactions with a medical professional may be impractical, such as adjusting prescribed treatments based on changes in health status or frequently monitoring a patient’s symptoms to identify changes in their condition more quickly. In these settings, the tools can expand the volume of encounters in which the frequency of interactions can improve outcomes and not merely replace the existing, high-value encounters with clinicians. Among some of the practical uses, an AI tool could help patients adjust medications for diabetes based on changes in diet or glucose levels, manage chief complaints commonly encountered in primary care settings, or help with the initial management of straightforward cases.

By initially developing and deploying these models in clinical situations when the disease trajectory and the associated treatment effects are well understood, developers can help instill confidence in the use of AI in health care. A carefully staged approach better prepares AI for use in more challenging settings, allowing physicians, patients, regulators, and product developers to enhance their understanding of how to generate the necessary data and assess a model’s safety and efficacy. The application of these tools will expand as larger, more reliable, and more representative data sets are used to train the LLMs, and more sophisticated models are developed to interpret a sufficiently vast number of clinical parameters.

To realize these opportunities, innovative methods will be needed to unlock and aggregate health care data, going beyond current legal frameworks to ensure interoperability and allow research on deidentified health information. Enhanced collaboration among health care systems, including data sharing and accessibility, will be crucial. By introducing AI in restricted health care environments initially, its applications can be carefully expanded across more medical settings over time. As data from diverse institutions are integrated, they can be used to refine the performance of AI models. Currently, only a handful of clinical LLMs are in use, and even the most advanced ones interpret a relatively limited number of parameters.

Focusing on routine medical encounters first, where ample data supports AI outputs, can also help address the issue of bias. With larger data sets, gaps in the underlying data can be identified more readily, including areas where the AI may be accurate or biased due to unrepresentative data.

Currently, developers of LLMs in health care primarily use them as digital assistants to streamline patient interactions. These tools collect data from patients and present clinicians with structured information. However, the functionality of current chatbots is deliberately constrained in that they do not offer diagnostic or treatment recommendations to patients because such actions would categorize them as regulated medical devices and subject them to FDA scrutiny. Many of these tools are not ready for such regulation, and few developers are inclined to undertake the process.
Although the current stage of development of LLMs does not enable clinicians to be removed from the decision-making process, these tools are poised to enhance medical care. In carefully chosen circumstances, they will oversee some routine aspects of health care, aiming to broaden the scope of patient engagement rather than replacing interactions with clinicians.