NEJM AI

A collection of articles from the *New England Journal of Medicine*, *NEJM Catalyst Innovations in Care Delivery*, and *NEJM Evidence*

AI & MACHINE LEARNING IN MEDICINE

The best approaches to integrating AI within the health care system

NEJM GROUP

July 2023

Dear Reader,

We hope you enjoy and find insight in this timely collection of articles that explores the fast-moving impact of artificial intelligence (AI) and machine learning in medicine. From clinical applications to imaging interpretation, infectious-disease surveillance to collaborative workflows, and real-time diagnostics to ethical considerations, each article provides a unique perspective on the challenges, opportunities, and future directions of AI in medicine.

In addition to needed technical advances, AI must meet the same bar for clinical evidence that is expected from other clinical interventions. Evidence that an AI tool will perform in a safe and effective manner must be demonstrated using randomized controlled trials designed to test the tool against an established standard. To meet clinical and technology innovators' needs, the forthcoming journal, *NEJM AI*, aims to provide a platform for rigorous evidence, resource sharing, and thoughtful discussions that will shape the integration of AI in medicine. Ahead of its publication in 2024, this collection of previously-published articles from the *New England Journal of Medicine*, *NEJM Catalyst Innovations in Care Delivery*, and *NEJM Evidence* exemplifies the high-quality content you can also expect from *NEJM AI*.

"Artificial Intelligence in Medicine," an Editorial from the *New England Journal of Medicine*, highlights the transformative potential of AI in medicine and provides the context for the 2024 *NEJM AI* journal launch. An NEJM Review Article, "Artificial Intelligence and Machine Learning in Clinical Medicine, 2023," examines the current landscape of AI in medicine, offering a comprehensive overview of the applications and implications.

Two new NEJM Review Articles explore how the latest innovations in AI can be used in different clinical applications. "The Current and Future State of AI Interpretation of Medical Images" explores cutting-edge advancements in AI technology in medical imaging. "Advances in Artificial Intelligence for Infectious-Disease Surveillance" highlights AI and its role in monitoring and predicting infectious diseases.

"Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," a recent NEJM Special Report, critically examines the potential of AI chatbots in health care, focusing on the ethical considerations and practical implications for clinical practice. While the 2019 NEJM Review Article, "Machine Learning in Medicine," explains how machine learning can analyze vast amounts of data to assist in prognosis, diagnosis, treatment selection, clinician workflow, and expanding the availability of clinical expertise, ultimately leading to more personalized and efficient health care.

"Using AI to Empower Collaborative Team Workflows: Two Implementations for Advance Care Planning and Care Escalation," showcases a compelling Case Study from *NEJM Catalyst Innovations in Care Delivery*, demonstrating how Stanford Medicine tackles challenges in implementing machine learning (ML) models in care delivery and describing two real-world pilots focusing on advance care planning and reducing unplanned escalations of care.

(continued on next page)

781.893.3800 | @NEJM_AI
860 WINTER STREET, WALTHAM, MA 02451-1413 USA | ai.nejm.org

(continued from previous page)

An Original Article from *NEJM Evidence*, "Real-Time Artificial Intelligence–Based Optical Diagnosis of Neoplastic Polyps during Colonoscopy," explores the potential of AI in cancer screening. In this study, colonoscopists diagnosed small colonic polyps as benign or malignant on the basis of their appearance. The results were compared in real time to see if CADx could distinguish among polyps requiring removal.

In reading this collection, we hope you discover the range of AI applications in medicine, guided by experts who prioritize transparency and uphold the highest standards. We look forward to identifying and evaluating more state-of-the-art applications of artificial intelligence to clinical medicine with *NEJM AI*, which launches in early 2024. Learn more at ai.nejm.org.

Sincerely,
NEJM AI Editorial Team

Isaac S. Kohane, MD, PhD
EDITOR-IN-CHIEF, NEJM AI

Eric J. Rubin, MD, PhD
EDITOR-IN-CHIEF, NEJM

Charlotte Haug, MD, PhD
EXECUTIVE EDITOR

Andrew L. Beam, PhD
DEPUTY EDITOR

Arjun K. Manrai, PhD
DEPUTY EDITOR

Table of Contents

The NEW ENGLAND JOURNAL of MEDICINE

Artificial Intelligence in Medicine

Andrew L. Beam, Ph.D., Jeffrey M. Drazen, M.D., Isaac S. Kohane, M.D., Ph.D., Tze-Yun Leong, Ph.D., Arjun K. Manrai, Ph.D., and Eric J. Rubin, M.D., Ph.D.

Artificial intelligence (AI) has gained recent public prominence with the release of deep-learning models that can generate anything from art to term papers with minimal human intervention. This development has reinvigorated discussion of the existing and potential roles of AI in all aspects of life. Among the wide range of fields with possible applications of AI, however, medicine stands out as one in which there is tremendous potential along with equally substantial challenges. At the *Journal*, we are seeing a rapidly increasing number of manuscript submissions that consider some aspect of AI applied to medicine. Given this enormous interest, we have now published in this issue of the *Journal* the first articles in a new series, AI in Medicine, that will consider both the reasonably established and the growing possible roles of AI and machine learning technologies in all aspects of health and health care.[1,2] Moreover, to further our commitment to this area, we are also announcing the 2024 launch of a new journal, *NEJM AI* (ai.nejm.org), which aims to provide a forum for high-quality evidence and resource sharing for medical AI along with informed discussions of its potential and limitations.

As a medical journal, we face two new publishing challenges for *NEJM AI*. The first is the breadth of potential AI applications. There is virtually no area in medicine and care delivery that is not already being touched by AI. For example, AI-driven applications are available to capture the dictation of medical notes; many such applications are attempting to synthesize patient interviews and laboratory test results to write notes directly, without clinician intervention. AI is playing an increasing role in health insurance coverage, assisting caregivers in making claims and payors in adjudicating them. We have already seen many published reports that use AI to interpret images — radiographs, histology, and optic fundi. Tools that utilize AI have come into increasing use in analyzing and interpreting large research databases containing information ranging from laboratory findings to clinical data. All these tools offer the potential for increased efficiency and may, perhaps, render insights that are difficult to attain with more traditional data-analysis methods. However, new AI methods are not necessarily a panacea; they can be brittle, they may work only in a narrow domain, and they can have built-in biases that disproportionally affect marginalized groups. This range of AI applications requires a diverse group of authors, editors, and reviewers, even though the pool of individuals with appropriate knowledge is still relatively small.

Second, expertise in the field of AI and machine learning is closely linked to commercial applications. The underlying technology is rapidly changing and, in many cases, is being produced by companies and academic investigators with financial interests in their products. For a growing class of large-scale AI models, companies that have the necessary resources may be the only ones able to push the frontier of AI systems. Since many such models are not widely available yet, hands-on experience and a detailed understanding of a model's operating characteristics often rest with only a small handful of model developers. Despite the potential for financial incentives that could create conflicts of interest,

View Article at NEJM.org

a deep understanding of AI and machine learning and their uses in medicine requires the participation of people involved in their development. Thus, in the series of AI articles we are publishing in the *Journal* and in *NEJM AI*, we will not restrict authorship and editorial control to persons without relevant financial ties but will follow a policy of transparency and disclosure.

Medicine is much different from other areas where AI is being applied. AI enables new discoveries and improved processes in the entire health care continuum; ethical, governance, and regulatory considerations are critical in the design, implementation, and integration of every component of the AI applications and systems. Because of concerns about both utility and safety, new applications will generally have to adhere to the same standards applied to other medical technologies. This will require a level of rigor in testing similar to that used in other areas of medicine, but it also can present challenges, such as the "dataset shift" that can result when there is a mismatch between the data set with which an AI system was developed and the data on which it is being deployed.[3] This summer, we hope to begin evaluating research studies for *NEJM AI* that bring careful methodology to understanding how to use AI and machine learning approaches in medicine. And as always, we welcome such studies at the *Journal*. We are excited to use our resources to encourage high-quality work in AI and to disseminate it with the same standards that we apply to everything we publish.

Disclosure forms provided by the authors are available with the full text of this editorial at NEJM.org.

From the Department of Epidemiology, Harvard T.H. Chan School of Public Health (A.L.B.), and the Department of Biomedical Informatics, Harvard Medical School (I.S.K., A.K.M.) — both in Boston; and the School of Computing, National University of Singapore, Singapore (T.-Y.L.).

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388:1201-8.
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9.
3. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385:283-6.

DOI: 10.1056/NEJMe2206291

The NEW ENGLAND JOURNAL of MEDICINE

Artificial Intelligence and Machine Learning in Clinical Medicine, 2023

Charlotte J. Haug, M.D., Ph.D., and Jeffrey M. Drazen, M.D.

Dr. Haug, the International Correspondent of the *Journal*, can be contacted at charlottejohanne@gmail.com or at Aamotveien 63, 0880 Oslo, Norway.

N Engl J Med 2023;388:1201-8.
DOI: 10.1056/NEJMra2302038

View Article at NEJM.org

As computers and the concept of artificial intelligence (AI) were almost simultaneously developed in the 1940s and 1950s, the field of medicine was quick to see their potential relevance and benefit.[1,2] In 1959, Keeve Brodman and colleagues claimed that "the making of correct diagnostic interpretations of symptoms can be a process in all aspects logical and so completely defined that it can be carried out by a machine."[3] Eleven years later, William B. Schwartz wrote in the *Journal*, "Computing science will probably exert its major effects by augmenting and, in some cases, largely replacing the intellectual functions of the physician."[4] He predicted that by the year 2000, computers would have an entirely new role in medicine, acting as a powerful extension of the physician's intellect.

However, by the late 1970s, there was disappointment that the two main approaches to computing in medicine — rule-based systems and matching, or pattern recognition, systems — had not been as successful in practice as one had hoped. The rule-based systems were built on the hypothesis that expert knowledge consists of many independent, situation-specific rules and that computers can simulate expert reasoning by stringing these rules together in chains of deduction. The matching strategies tried to match a patient's clinical characteristics with a bank of "stored profiles," which we now refer to as "illness scripts,"[5] of the findings in a given disease. More effort was put into understanding the clinical decision-making process itself.[6] It became clear that the key deficiencies in most previous programs stemmed from their lack of pathophysiological knowledge. When such knowledge was incorporated, the performance greatly improved.

Nevertheless, in the 1980s, computers were not up to the task. The rule-based systems had by 1987 proved useful in a variety of commercial tasks but had not worked in clinical medicine. Indeed, Schwartz and colleagues noted that "the process is so slow that it is impractical even with modern high-speed computers."[7] They continued: "After hearing for several decades that computers will soon be able to assist with difficult diagnoses, the practicing physician may well wonder why the revolution has not occurred."[7]

PROGRESS IN DATA SCIENCE

In the 1950s, computers were large and slow. The first hard-disk drive was the IBM Model 350 Disk File, introduced in 1956. It had a total storage capacity of 5 million characters (just under 5 MB). The first hard drive to have more than 1 GB in capacity was the IBM 3380, introduced in 1980. It was the size of a refrigerator and weighed 550 lb (250 kg); the price was $100,000. But integrated-circuit technology was improving. In 1965, Gordon Moore, cofounder of Fairchild Semiconductor and Intel, predicted that the number of transistors in an integrated circuit, and, hence,

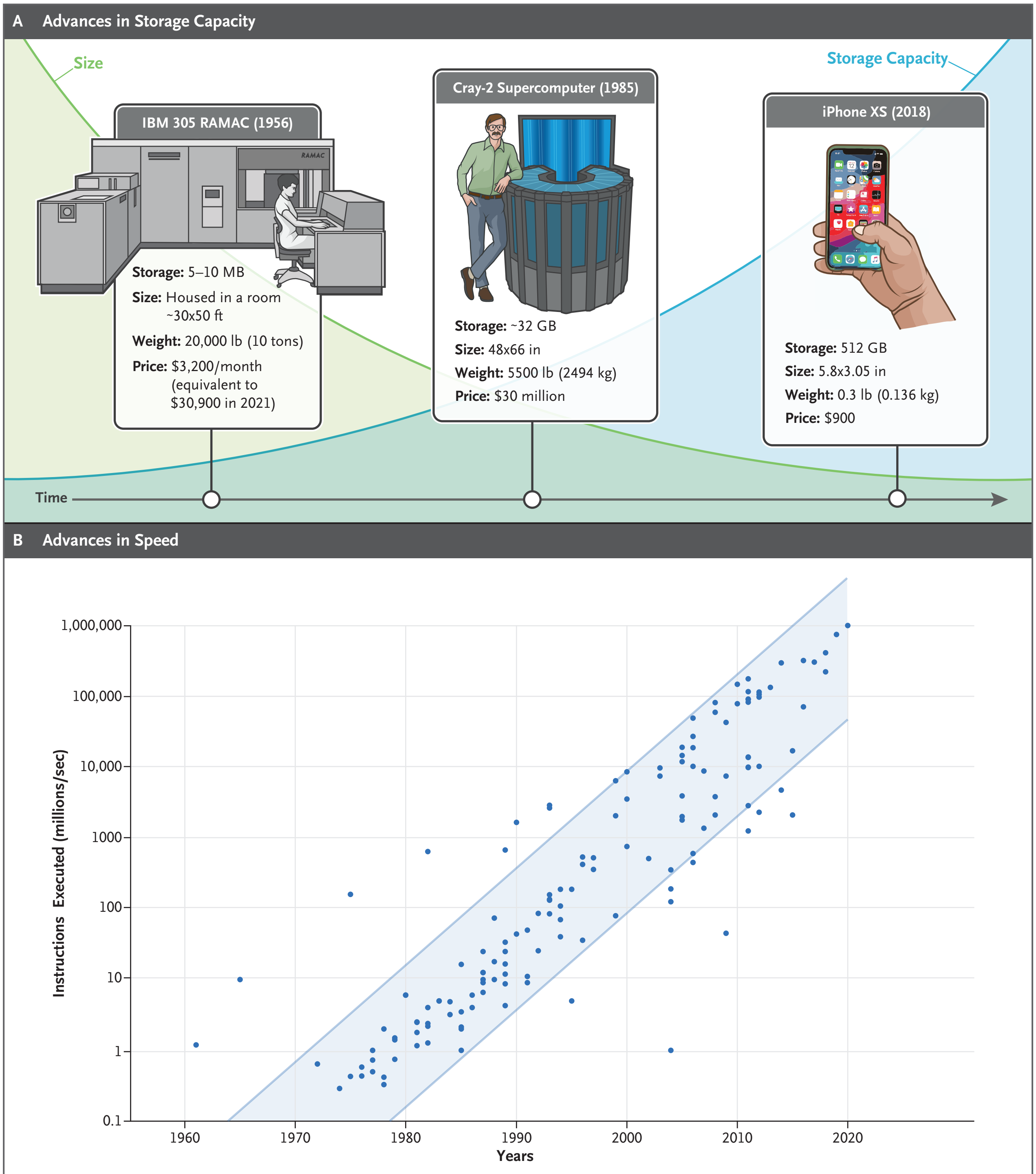


Figure 1. **Improvements over 50 Years in the Ability of Computers to Store and Process Data.**
Panel A shows advances in data storage, in terms of both physical size and cost per unit of storage. RAMAC denotes random access method of accounting and control. Panel B shows advances in the speed of computing. Each dot represents an individual machine type and the approximate year of its introduction. These improvements in storage and speed have allowed machine learning to progress from a dream to reality. Data in both panels are estimates from many types of system architecture and are derived from multiple public sources.

its potential computing power, would double every 2 years. His prediction was right; this change in semiconductor density is known as Moore's law. However, Moore's law tells us more than the number of transistors per square centimeter, since other aspects of technological progress, such as processing speed and the price of electronic products, are strongly linked to Moore's law. With more dense circuits, computer memory and computing speeds increased, and today, pocket-sized devices that are more powerful than the 1980s supercomputers, which took up entire rooms, are common and available at a fraction of the price (Fig. 1).

Progress in data science is not simply a matter of increased performance, speed, and storage. In addition to the type of information found in libraries, data generated in organizations, and established systems designed to gather and codify data, new forms of technology can use data that are both people-generated and machine-generated. These data are often chaotic and unstructured. Data now come from many additional sources, including social networks, blogs, chat rooms, product-review sites, communities, website pages, email, documents, images, videos, and music, along with wearable and environmental sensors. Many people open aspects of their medical records and personal genetic data for online access by anyone. Storage capacity is so great that vast portions of the corpus of recorded human knowledge and activity can be stored and readily accessed.

Once we had the data, we needed more than data; we needed ways to identify and process the data. Google became the leader in online searching by harnessing the searches performed by others to identify what people wanted to know. This required a second revolution, mathematical algorithms that could rapidly, and with reasonable reliability, track this behavior and aid the end user in finding particular information. More dense information storage and faster computing allowed for practical, real-time solutions of mathematical expressions that could be used to find relationships in the data that were previously unknowable. As a result, data science could flourish and flex its muscles in a way that was previously impossible.

We are now able to use unstructured data to identify untold relationships among elements in the data, allowing the use of dynamic data and data with multiple contexts that, when approached and analyzed in nontraditional ways, provide actionable insights into human behavior. Neural networks became more sophisticated as the computing power allowed functional real-time output to data queries. Transformers (i.e., deep-learning models that differentially weigh the importance of each part of the input data) made natural-language processing possible. With this approach, the complexities of the underlying computer models, and the corpus of data from which those models could draw, grew and became more powerful. The goal of a computer that could emulate certain aspects of human interaction went from an impossible dream to a reality.

The connectedness allowed by data science is driving a new kind of discovery. People are using social networks to draw their own connections between friends, things, events, likes, dislikes, places, ideas, and emotions. Governments are analyzing social networks to stop terrorist acts. Businesses are mining social and transactional information for connections that will help them discover new opportunities. Scientists are building massive grids of connected data to tease out new findings, using AI and machine learning. As addressed in more detail below, these advances have allowed the emergence of computers that can help you perform tasks that previously had been tedious. The Star Wars character C-3PO was a crude version of the AI-based virtual assistants (e.g., Apple's Siri, Google's Assistant, and Amazon's Alexa) that have become part of our daily life and can help us perform defined tasks. Anyone who has used one of these devices has experienced their convenience (e.g., instructing the virtual assistant to "set the oven timer for 20 minutes" so that food is properly cooked) but also the annoyance of having the assistant break into a conversation with some unrelated facts. AI and machine learning constitute the driving force behind these devices.

AI AND MACHINE LEARNING IN MEDICINE

In the 1990s and into the early 2000s, even with slow computers and limited memory, the problem of having machines successfully perform certain medical tasks that were repetitive, and therefore prone to human error, was being solved. Through a substantial investment of money and intellectual effort, computer reading of electrocardiograms (ECGs) and white-cell differential counts, analysis of retinal photographs and cutaneous lesions, and other image-processing tasks has become a reality. Many of these machine-learning–aided tasks have been largely accepted and incorporated into the everyday practice of medicine. The performance of these machine tasks is not perfect and often requires a skilled person to oversee the process, but in many cases, it is good enough, given the need for relatively rapid interpretation of images and the lack of local expertise.

However, the use of AI and machine learning in medicine has expanded beyond the reading of medical images. AI and machine-learning programs have entered medicine in many ways, including, but not limited to, helping to identify outbreaks of infectious diseases that may have an impact on public health; combining clinical, genetic, and many other laboratory outputs to identify rare and common conditions that might otherwise have escaped detection; and aiding in hospital business operations (Fig. 2). In the months to come, the *Journal* will publish other review articles that take a selective look at AI and machine learning in medicine in 2023. But before the first article appears, in about a month's time, it is important to consider the overriding issues that need to be considered as we learn to work hand in hand with machines.

UNRESOLVED ISSUES IN AI AND MACHINE LEARNING IN MEDICINE

ESTABLISHING NORMS

As noted above, the use of AI and machine learning has already become accepted medical practice in the interpretation of some types of medical images, such as ECGs, plain radiographs, computed tomographic (CT) and magnetic resonance imaging (MRI) scans, skin images, and retinal photographs. For these applications, AI and machine learning have been shown to help the health care provider by flagging aspects of images that deviate from the norm.

This suggests a key question: what is the norm? This simple question shows one of the weaknesses of the use of AI and machine learning in medicine as it is largely applied today. How does bias in the way AI and machine-learning algorithms were "taught" influence how they function when applied in the real world? How do we interject human values into AI and machine-learning algorithms so that the results obtained reflect the real problems faced by health professionals? What issues must regulators address to ensure that AI and machine-learning applications perform as advertised in multiple-use settings? How should classic approaches in statistical inference be modified, if at all, for interventions that rely on AI and machine learning? These are but a few of the problems that confront us; the "AI in Medicine" series will address some of these matters.

ROLE OF AI AND MACHINE LEARNING IN CLINICAL PRACTICE

Pitfalls aside, there is much promise. If AI and machine-learning algorithms can be reduced to clinically useful "apps," will they be able to weed their way through mountains of clinical, genomic, metabolomic, and environmental data to aid in precision diagnosis? Can AI and machine-learning–driven apps become your personal scribe and free up your time spent on documentation so that you can spend more time with patients? Can the apps prompt you to ask a key question that could help in the differential diagnosis? Can they outwit the AI and machine-learning algorithms, used by insurance companies, that make it difficult for you to order a positron-emission tomographic–CT scan or collect reimbursement for the time you spent with a patient and the patient's family? In each area, progress has been made. Is it good enough?

CLINICAL RESEARCH ON AI AND MACHINE-LEARNING APPLICATIONS

The evaluation of progress has its own set of problems. In traditional clinical research, when progress takes the form of a new drug for a definable condition, the standards for testing and accepting the drug as an advance are well established. When the intervention is an AI and machine-learning algorithm rather than a drug, the medical community expects the same level of surety, but the standards for describing and testing AI and machine-learning interventions are far from clear.

What are the standards to which AI and

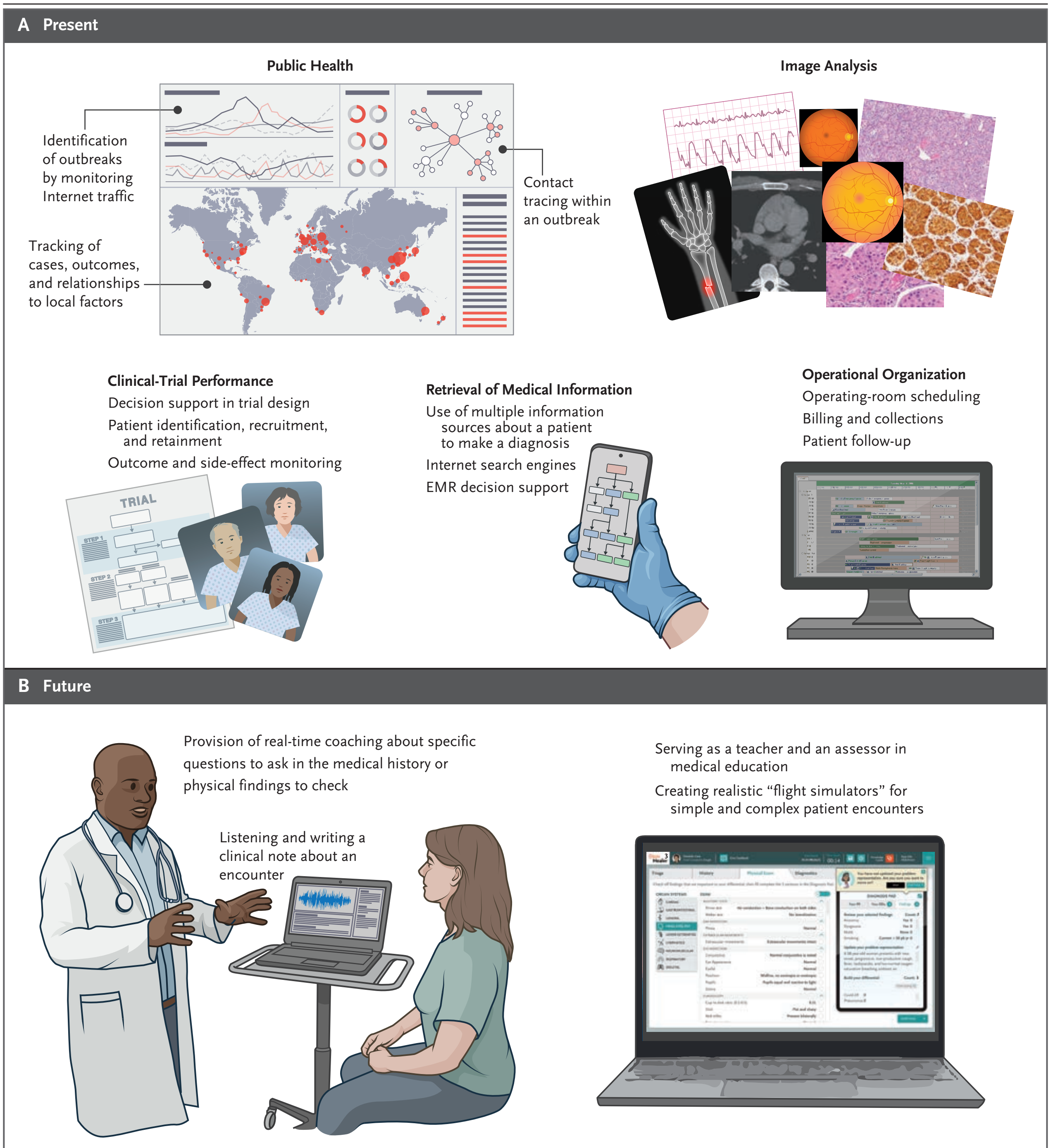


Figure 2. Spectrum of Artificial Intelligence (AI) in Medicine.
Panel A shows selected areas of public health and medicine in which AI has an established but evolving role. These tools are already helping medical professionals do their jobs as partners in practice. EMR denotes electronic medical record. Panel B shows areas of medical practice in which AI has begun to have an influence but has not yet reached the stage of common use.

machine learning–based interventional research should be held, if an app is going to be accepted as the standard that will shape, reform, and improve clinical practice? That research has three components. First, the research must be structured to answer a clinically meaningful question in a way that can influence the behavior of the health professional and lead to an improvement in outcomes for a patient. Second, the intervention must be definable, scalable, and applicable to the problem at hand. It must not be influenced by factors outside the domain of the problem and must yield outcomes that can be applied to similar clinical problems across a wide range of populations and disease prevalences. Can AI and machine learning–driven care meet these standards — ones that we demand from a novel therapeutic intervention or laboratory-based diagnostic test — or do we need to have a unique set of standards for this type of intervention? Third, when the results of the research are applied in such a way as to influence practice, the outcome must be beneficial for all patients under consideration, not just those who are similar to the ones with characteristics and findings on which the algorithm was trained. This raises the question of whether such algorithms should include consideration of public health (i.e., the use of scarce resources) when diagnostic or treatment recommendations are being made and the extent to which such considerations are part of the decision-making process of the algorithm. Such ethical considerations have engaged health professionals and the public for centuries.[8]

USE OF AI AND MACHINE-LEARNING APPLICATIONS IN CONDUCTING CLINICAL RESEARCH

AI and machine learning have the potential to improve and possibly simplify and speed up clinical trials through both more efficient recruitment and matching of study participants and more comprehensive analyses of the data. In addition, it may be possible to create synthetic control groups by matching historical data to target trial enrollment criteria. AI and machine learning may also be used to better predict and understand possible adverse events and patient subpopulations. It seems possible that AI could generate "synthetic patients" in order to simulate diagnostic or therapeutic outcomes. But the use of AI and machine-learning applications and interventions introduces a set of uncertainties that must be dealt with both in protocols and in reporting of clinical trials.[9,10]

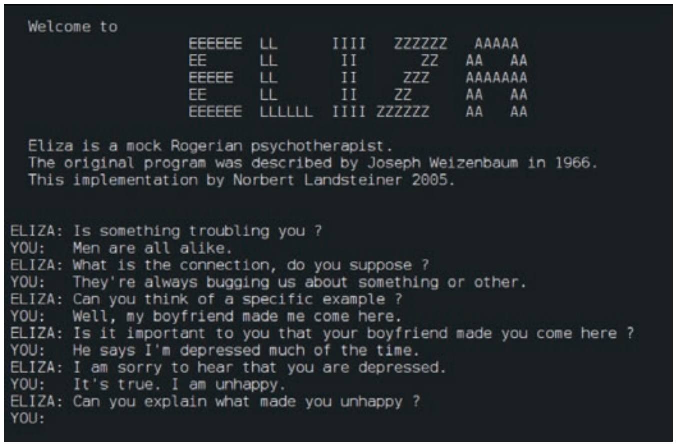
In this AI in Medicine series, we plan to cover progress, pitfalls, promise, and promulgation at the interface of AI and medicine. It is important to understand that this is a fast-moving field, so to some extent, what we publish may have the resolution of a snapshot of the landscape taken from a bullet train. Specifically, things happening in close temporal proximity to publication may be blurred because they are changing quickly, but the distant background will be in reasonably good focus. One area of substantial progress in AI and machine learning (i.e., in the foreground, in our snapshot analogy) is the appearance of sophisticated chatbots that are available for use by the general public. Although chatbots have only recently been introduced at a level of sophistication that could have an impact on daily medical practice, we believe that their potential to influence how medicine is practiced is substantial and that we would be remiss not to address that potential as well as possible problems related to their use.

CHATBOTS IN MEDICINE

In this issue of the *Journal*, an article by Lee et al.[11] introduces the GPT-4 chatbot and its medical applications. The article was written by a team of researchers who work for the entities that created GPT-4, a chatbot with a broad education that includes medical knowledge. Before we see the future, a quick look at the past will be helpful. A chatbot is a computer program that uses AI and natural-language processing to understand questions and automate responses to them, simulating human conversation. A very early medical chatbot, ELIZA, was developed between 1964 and 1966 by Joseph Weizenbaum at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology (Fig. 3).

Chatbot technology is now almost everywhere, from customer service to personal virtual assistants, as noted above. With the powerful computers available today, language models have hundreds of billions of parameters, which can be used to generate new text. This ability, combined with an almost infinite amount of available (Internet) data with which to train the

A



B

Question:
A 12-year-old girl is brought to the emergency department by her mother because of a 1-week history of worsening leg swelling. The patient also noticed blood in her urine yesterday. The bleeding has not recurred. She had an upper respiratory tract infection and sore throat 1 week ago that caused her to miss several days of school. Medical history is otherwise unremarkable, and she takes no routine medications. Menarche has not yet occurred. BMI is 20 kg/m^2. Vital signs are temperature 37.0°C (98.6°F), pulse 78 beats/min, respiratory rate 12 breaths/min, and blood pressure 136/84 mm Hg. Pulse oximetry while the patient is breathing ambient air shows an oxygen saturation of 100%. Physical examination shows erythema of the posterior pharynx, mild cervical lymphadenopathy, and 3+ pitting edema in both knees. Results of urinalysis are shown:

Protein	150 mg/dl
Blood	Positive
Leukocyte esterase	Positive
Nitrite	Negative
White cells	5–10/high-power field
Red cells	10–25/high-power field
Casts	1–2/low-power field

Results of which of the following laboratory studies are most likely to be abnormal in this patient?
(A) Bleeding time
(B) Erythrocyte count
(C) Serum concentration of C3
(D) Serum IgA concentration
(E) Serum rheumatoid factor assay

The correct answer is (C) Serum concentration of C3. This patient has signs and symptoms of acute poststreptococcal glomerulonephritis (APSGN), a type of glomerular disease that typically occurs 1 to 3 weeks after a streptococcal infection of the skin or pharynx. APSGN is mediated by immune complexes that deposit in the glomerular basement membrane and activate the complement pathway, leading to inflammation, hematuria, proteinuria, and reduced glomerular filtration rate. The resulting fluid retention causes hypertension and edema. The serum C3 level is usually low in APSGN, reflecting complement consumption. The other laboratory studies listed are normal or not relevant in APSGN. Bleeding time is a measure of platelet function, which is not affected by APSGN. Erythrocyte count may be slightly elevated due to hemoconcentration, but not significantly. Serum IgA concentration is elevated in IgA nephropathy, another type of glomerular disease that can cause hematuria and proteinuria, but it is not associated with streptococcal infection or low C3 level. Serum rheumatoid factor assay is used to diagnose rheumatoid arthritis, an autoimmune disorder that does not cause APSGN.

Figure 3. Chatbots in Medicine.

Panel A shows a screen shot of output from a very early medical chatbot called ELIZA, which was developed by Joseph Weizenbaum at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology between 1964 and 1966. Panel B (courtesy of Lee et al.[11]) shows input and output from the GPT-4, a chatbot that is expected to be introduced in 2023. BMI denotes body-mass index.

network, means that language models can do more and more, as shown by the Chat Generative Pre-trained Transformer, or ChatGPT.

ChatGPT is a language model trained by OpenAI. It was introduced publicly in November 2022 (https://openai.com/blog/chatgpt) and has demonstrated a new way in which AI-driven machines can interact with people. The new-generation chatbots hold the promise of being a scribe and coach, but with some key caveats. Many of these caveats were described by the developers of ChatGPT at its launch but warrant special consideration when used in medicine, as detailed by Lee et al.[11] In their current iteration, the new generation of chatbots can help with the medical documentation problem and answer key questions that could help in the differential diagnosis, as noted above. But it is difficult to know whether the answers provided are grounded in appropriate fact. The onus would be on clinicians to proofread the work of the chatbot, just as clinicians need to proofread clinical notes that they dictate. The difficulty is that such proofreading may be beyond the expertise of the user. Proofreading a note on a patient visit is likely to be well within the range of the provider's expertise, but if the chatbot is asked a question as a "curbside consult," the veracity of the answer may be much harder to determine.

The application of greatest potential and concern is the use of chatbots to make diagnoses or recommend treatment. A user without clinical experience could have trouble differentiating fact from fiction. Both these issues are addressed in the article by Lee and colleagues,[11] who point out the strengths and weaknesses of using chatbots in medicine. Since the authors have created one such entity, bias is likely.

Nevertheless, we think that chatbots will become important tools in the practice of medicine. Like any good tool, they can help us do our job better, but if not used properly, they have the potential to do damage. Since the tools are new and hard to test with the use of the traditional methods noted above, the medical community will be learning how to use them, but learn we must. There is no question that the chatbots will also learn from their users. Thus, we anticipate a period of adaptation by both the user and the tool.

CONCLUSIONS

We firmly believe that the introduction of AI and machine learning in medicine has helped health professionals improve the quality of care that they can deliver and has the promise to improve it even more in the near future and beyond. Just as computer acquisition of radiographic images did away with the x-ray file room and lost images, AI and machine learning can transform medicine. Health professionals will figure out how to work with AI and machine learning as we grow along with the technology. AI and machine learning will not put health professionals out of business; rather, they will make it possible for health professionals to do their jobs better and leave time for the human–human interactions that make medicine the rewarding profession we all value.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Turing AM. Computing machinery and intelligence. Mind 1950;59:433-60.
2. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31.
3. Brodman K, Van Woerkom AJ, Erdmann AJ Jr, Goldstein LS. Interpretation of symptoms with a data-processing machine. AMA Arch Intern Med 1959;103:776-82.
4. Schwartz WB. Medicine and the computer — the promise and problems of change. N Engl J Med 1970;283:1257-64.
5. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med 2006;355:2217-25.
6. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition: taking a present illness by computer. Am J Med 1976;60:981-96.
7. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Where do we stand? N Engl J Med 1987;316:685-8.
8. Rosenbaum L. Trolleyology and the dengue vaccine dilemma. N Engl J Med 2018;379:305-7.
9. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364-74.
10. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020;2(10):e549-e560.
11. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9.

The NEW ENGLAND JOURNAL of MEDICINE

The Current and Future State of AI Interpretation of Medical Images

Pranav Rajpurkar, Ph.D., and Matthew P. Lungren, M.D., M.P.H.

The interpretation of medical images — a task that lies at the heart of the radiologist's work — has involved the growing adoption of artificial intelligence (AI) applications in recent years. This article reviews progress, challenges, and opportunities in the development of radiologic AI models and their adoption in clinical practice. We discuss the functions that AI-based algorithms serve in assisting radiologists, including detection, workflow triage, and quantification, as well as the emerging trend of the use of medical-imaging AI by clinicians who are not radiologists. We identify the central challenge of generalization in the use of AI algorithms in radiology and the need for validation safeguards that encompass clinician–AI collaboration, transparency, and post-deployment monitoring. Finally, we discuss the rapid progress in developing multimodal large language models in AI; this progress represents a major opportunity for the development of generalist medical AI models that can tackle the full spectrum of image-interpretation tasks and more. To aid readers who are unfamiliar with terms or ideas used for AI in general or AI in image interpretation, a Glossary is included with this article.

In recent years, AI models have been shown to be remarkably successful in interpretation of medical images.[1] Their use has been extended to various medical-imaging applications, including, but not limited to, the diagnosis of dermatologic conditions[2] and the interpretation of electrocardiograms,[3] pathological slides,[4] and ophthalmic images.[5] Among these applications, the use of AI in radiology has shown great promise in detecting and classifying abnormalities on plain radiographs,[6] computed tomographic (CT) scans,[7] and magnetic resonance imaging (MRI) scans,[8] leading to more accurate diagnoses and improved treatment decisions.

Even though the Food and Drug Administration (FDA) has approved more than 200 commercial radiology AI products, substantial obstacles must be overcome before we are likely to see widespread successful clinical use of these products. The incorporation of AI in radiology poses both potential benefits and challenges for the medical and AI communities. We expect that the eventual resolution of these issues and more comprehensive solutions, including the development of new foundation models, will lead to broader adoption of AI within this health care sector.

From the Department of Biomedical Informatics, Harvard Medical School, Boston (P.R.); the Center for Artificial Intelligence in Medicine and Imaging, Stanford University, Stanford, and the Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco — both in California (M.P.L.); and Microsoft, Redmond, Washington (M.P.L.). Dr. Rajpurkar can be contacted at pranav_rajpurkar@hms.harvard.edu.

Drs. Rajpurkar and Lungren contributed equally to this article.

N Engl J Med 2023;388:1981-90.
DOI: 10.1056/NEJMra2301725

View Article at NEJM.org

AI USE IN RADIOLOGY

Radiology as a specialty is well positioned for the application and adoption of AI because of several key factors. First, AI excels in analyzing images,[9] and unlike other specialties that use imaging, radiology has an established digital workflow and universal standards for image storage, so that it is easier to integrate AI.[10]

Glossary

Continual learning: A process in which an AI model learns from new data over time while retaining previously acquired knowledge.

Data set shift: The shift from data used to train a machine-learning model to data encountered in the real world. This shift can cause the model to perform poorly when used in the real world, even if it performed well during training.

Federated learning: A distributed machine-learning approach that enables multiple devices or nodes to collaboratively train a shared model while keeping their individual data local, thereby preserving privacy and reducing data communication overhead.

Foundation models: AI models that serve as a starting point for developing more specific AI models. Foundation models are trained on large amounts of data and can be fine-tuned for specific applications, such as detecting lesions or segmenting anatomical structures.

Generalist medical AI models: A class of advanced medical foundation models that can be used across various medical applications, replacing task-specific models. Generalist medical AI models have three key capabilities that distinguish them from conventional medical AI models. They can adapt to new tasks described in plain language, without requiring retraining; they can accept inputs and produce outputs using various combinations of data types; and they are capable of logically analyzing unfamiliar medical content.

Large language models: AI models consisting of a neural network with billions of weights or more, trained on large amounts of unlabeled data. These models have the ability to understand and produce human language and may also apply to images and audio.

Multimodal models: AI models that can understand and combine different types of medical data, such as medical images and electronic health records. Multimodal models are particularly useful in medicine for tasks that require a comprehensive understanding of the patient, such as diagnosis and individualized treatment planning.

Self-supervised models: AI models that can learn from medical data without the need for explicit annotations. These models can be used to learn representations of medical data that are useful for a wide range of tasks, such as diagnosis and patient monitoring. Self-supervised models are particularly useful in medicine when labeled data are scarce or expensive to obtain.

Zero-shot learning: The capability of an AI model to perform a task or solve a problem for which it has not been explicitly trained, without the need for any additional training data. In medicine, this can be particularly useful when there is a shortage of labeled data available for a specific medical task.

Furthermore, AI fits naturally in the workflow of image interpretation and can replicate well-defined interpretive tasks effectively.[11]

AI USE FOR RADIOLOGISTS

AI can be used in the field of radiology to analyze images from a wide range of techniques, including radiography, CT, ultrasonography, and MRI. Radiologic AI algorithms serve a number of narrow image-analysis functions to assist radiologists, such as quantification, workflow triage, and image enhancement (Fig. 1).[1,12-17] Quantification algorithms perform segmentation and measurements of anatomical structures or abnormalities. Common examples include measuring breast density, identifying anatomical structures in the brain, quantitating cardiac flow,[18] and assessing local lung-tissue density. Workflow triage involves flagging and communicating suspected positive findings, including, but not limited to, intracranial hemorrhage, intracranial large-vessel occlusion,[19] pneumothorax,[20] and pulmonary embolism. AI is also used for the detection, localization, and classification of conditions such as pulmonary nodules and breast abnormalities. In addition, AI algorithms enhance preinterpretive processes, including image reconstruction, image acquisition, and mitigation of image noise.[17]

There is promise in exploring radiologic AI models that can expand interpretive capabilities beyond those of human experts. For instance, AI algorithms can accurately predict clinical outcomes on the basis of CT data in cases of traumatic brain injury[21] and cancer.[22] In addition, AI-derived imaging biomarkers can help to quickly and objectively assess structures and pathological processes related to body composition, such as bone mineral density, visceral fat, and liver fat, which can be used to screen for various health conditions.[23] When applied to routine CT imaging, these AI-derived biomarkers are proving useful in predicting future adverse events.[24] Moreover, recent research has shown that coronary-artery calcium scores, which are typically obtained on the basis of CT scanning, can be determined by means of cardiac ultrasonography.[25] These findings point to the value of radiologic AI models for patients (e.g., no radiation exposure).

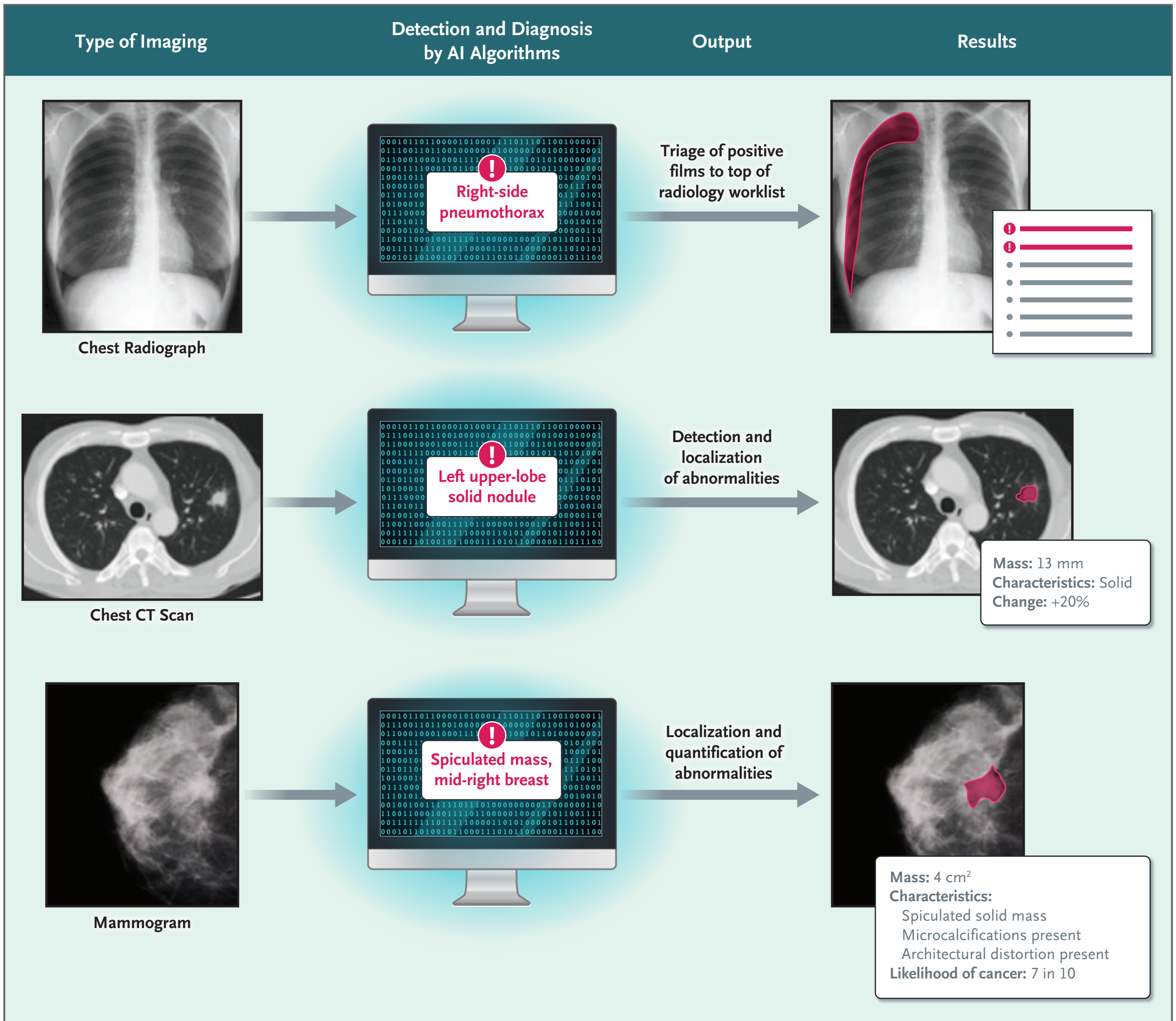


Figure 1. Current Uses of Artificial Intelligence (AI) in Radiology.

Shown are common clinical AI solutions for the functions of triage, detection, and diagnosis with CADt (computer-aided detection for triage), CADe (computer-aided detection for characterizing abnormalities), and CADx (computer-aided detection for diagnosis). Other AI applications for radiology include image reconstruction and noise reduction. Applications for nonimaging tasks are not shown. CT denotes computed tomography.

Radiologic AI has attracted global interest, and commercial AI algorithms have been developed by companies based in more than 20 countries. Studies have shown that some hospitals, as well as other point-of-care centers, already use AI products successfully, and larger practices are more likely than smaller practices to use AI currently. Radiologists who use AI in their practices are generally satisfied with their experience and find that AI provides value to them and their patients. However, radiologists have expressed concerns about lack of knowledge, lack of trust, and changes in professional identity and autonomy.[26] Local champions of AI, education, training, and support can help overcome these concerns. The majority of radiologists and residents expect substantial changes in the radiology profession within the next decade and believe that AI should have a role as a “co-pilot,” acting as a second reader and improving workflow tasks.[27]

Although the penetration of AI in the U.S. market is currently estimated to be only 2%, the readiness of radiologists and the potential of the technology indicate that further translation into clinical practice is likely to occur.

EMERGING USES FOR NONRADIOLOGISTS

Although many current radiologic AI applications are designed for radiologists, there is a small but emerging trend globally toward the use of medical-imaging AI for nonradiologist clinicians and other stakeholders (i.e., health care providers and patients). This trend presents an opportunity for improving access to medical imaging and reducing common diagnostic errors[28] in low-resource settings and emergency departments, where there is often a lack of around-the-clock radiology coverage.[29] For instance, one study showed that an AI system for chest radiograph interpretation, when combined with input from a nonradiology resident, had performance values that were similar to those for board-certified radiologists.[30] A popular AI application that is targeted for use by nonradiologist clinicians for detecting large-vessel occlusions in the central nervous system has resulted in a significant reduction in time to intervention and improved patient outcomes.[31] Moreover, AI has been shown to accelerate medical-imaging acquisition outside traditional referral workflows with new, clinician-focused mobile applications for notifications of AI results.[32] This trend, although not well established, has been cited as a potential long-term threat to radiology as a specialty because advanced AI models may reduce the complexity of technical interpretation so that a nonradiologist clinician could use imaging without relying on a radiologist.[33,34]

Portable and inexpensive imaging techniques are frequently supported by AI and have served to lower the barrier for more widespread clinical use of AI in medical imaging outside the traditional radiology workflow.[35,36] For example, the Swoop portable MRI system, a point-of-care device that addresses existing limitations in forms of imaging technology, provides accessibility and maneuverability for a range of clinical applications. The system plugs into a standard electrical outlet and is controlled by an Apple iPad. Portable ultrasound probes and smartphones in AI-enabled applications can be used to obtain diagnostic information even by users without formal training in echocardiography or the use of ultrasound in obstetrical care.[37] Overall, although the use of medical-imaging AI by nonradiologist clinicians is still in the early stages, it has the potential to revolutionize access to medical imaging and improve patient outcomes.

SAFEGUARDS FOR EFFECTIVE GENERALIZATION

In considering the widespread adoption of AI algorithms in radiology, a critical question arises: Will they work for all patients? The models underlying specific AI applications are often not tested outside the setting in which they were trained, and even AI systems that receive FDA approval are rarely tested prospectively or in multiple clinical settings.[38] Very few randomized, controlled trials have shown the safety and effectiveness of existing AI algorithms in radiology, and the lack of real-world evaluation of AI systems can pose a substantial risk to patients and clinicians.[39]

Moreover, studies have shown that the performance of many radiologic AI models worsens when they are applied to patients who differ from those used for model development, a phenomenon known as “data set shift.”[40-44] In interpretation of medical images, data set shift can occur as a result of various factors, such as differences in health care systems, patient populations, and clinical practices.[45] For instance, the performance of models for brain tumor segmentation and chest radiograph interpretation worsens when the models are validated on external data collected at hospitals other than those used for model training.[46,47] In another example, a retrospective study showed that the performance of a commercial AI model in detecting cervical spine fractures was worse in real-world practice than the performance initially reported to the FDA.[48] Patient age, fracture characteristics, and degenerative changes in the spine affected the sensitivity and false positive rates to an extent that limited the clinical usefulness of the AI model and aroused concerns about the generalization of radiologic AI algorithms across clinical environments.

There is a pressing need for the development of methods that improve the generalization of algorithms in new settings.[49-51] As the field matures, better generalization checks based on accepted standards must be established before the algorithms are widely applied. These checks encompass three related areas: clinician–AI collaboration, transparency, and monitoring (Fig. 2).

CLINICIAN–AI COLLABORATION

The successful use of AI in radiology depends on effective clinician–AI collaboration. In theory, the use of AI algorithms to assist radiologists allows for a human–AI collaboration workflow, with humans and AI leveraging complementary strengths.[52] Studies have shown that AI assistance in interpretation of medical images is more useful to some clinicians than to others and generally provides more benefit to less experienced clinicians.[53,54]

Despite some evidence that clinicians receiving AI assistance can achieve better performance than unassisted clinicians,[53,55,56] the body of research on human–AI collaboration for image interpretation offers mixed evidence regarding the value of such a collaboration. Results vary according to particular metrics, tasks, and the study cohorts in question, with studies showing that although AI can improve the performance of radiologists, sometimes AI alone performs better than a radiologist using AI.[57,58]

Many AI methods are "black boxes," meaning that their decision-making processes are not easily interpretable by humans; this can pose challenges for clinicians trying to understand and trust the recommendations of AI.[59] Studies of the potential for explainable AI methods to build trust in clinicians have shown mixed results.[59,60] Therefore, there is a need to move from evaluations centered on the stand-alone performance of models to evaluations centered on the outcomes when these algorithms are used as assistive tools in real-world clinical workflows. This approach will enable us to better understand the effectiveness and limitations of AI in clinical practice and establish safeguards for effective clinician–AI collaboration.

TRANSPARENCY

Transparency is a major challenge in evaluating the generalization behavior of AI algorithms in medical imaging. Scientific, peer-reviewed evidence of efficacy is lacking for most commercially available AI products.[38] Many published reports on FDA-cleared devices omit information on sample size, demographic characteristics of patients, and specifications of the equipment used to acquire the images to be interpreted. In addition, only a fraction of device studies offer data on the specific demographic subgroups used during algorithm training, as well as the diagnostic performance of these algorithms when applied to patients from underrepresented demographic subgroups. This lack of information makes it difficult to determine the generalizability of AI and machine-learning algorithms across different patient populations.

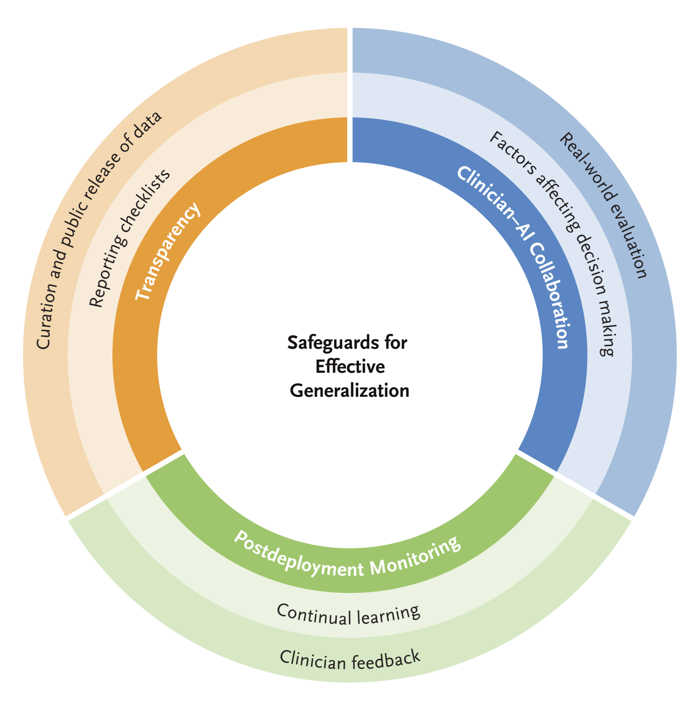


Figure 2. Generalization Checks for AI Systems in Radiology.
The three essential components of generalization checks for radiologic AI systems are clinician–AI collaboration, transparency, and postdeployment monitoring. Clinician–AI collaboration reflects the need to move from evaluations of the stand-alone performance of AI models to evaluations of their value as assistive tools in real-world clinical workflows. Transparency with regard to lack of information about an AI model requires greater rigor through the use of checklists and public release of medical-imaging data sets. Postdeployment monitoring involves mechanisms to incorporate feedback from clinicians and continual learning strategies for regular updating of the models.

The limited independent validation of these models has generated a call for greater transparency and rigor with the use of checklists to verify the proper implementation of AI models in medical imaging and to ensure adequate reproducibility and clinical effectiveness.[61-63] One solution for transparency is the curation and public release of medical-imaging data sets to serve as a common benchmark and show algorithm performance.[64-67] The availability of publicly released chest radiograph data sets has already provided support for marked advances in improving AI validation.[68,69] However, there are challenges in curating public medical-imaging data sets, including privacy concerns about sharing data,[70] costs of data infrastructure,[71] and overrepresentation of data from academic medical centers with substantial resources.[72] Federated learning, another approach to data sharing, involves training an AI model on decentralized data sources without transferring the data to a central repository.[73,74] Streamlined processes for curating and sharing diverse medical data sets are necessary for transparency in establishing clinical usefulness.

POSTDEPLOYMENT MONITORING

Even after a model is deployed, its performance in the real world may degrade over time. In interpretation of medical images, these shifts can occur as a result of various factors such as changes in disease prevalence, advances in medical technology, and alterations in clinical practices.[38,75-77] Failure to update the model to reflect these changes can lead to poor model performance and misuse. However, regulatory requirements may restrict updating of models after they have been approved.

Continuous monitoring of model performance and proactive measures to address data set shifts over time can improve the accuracy and reliability of AI models in medical-imaging interpretation. Regular updates of the training data and retraining of the model on new data through continual learning can help maintain model performance over time.[78] In addition, incorporating feedback from clinicians can help improve the performance of the model by providing real-world insights and identifying areas for improvement. Ultimately, postdeployment monitoring is essential to ensure that AI models remain effective and reliable in clinical settings.[79,80]

GENERALIST MEDICAL AI MODELS FOR RADIOLOGY

The current generation of AI models in radiology can handle only a limited set of interpretation tasks, and they rely heavily on curated data that have been specifically labeled and categorized.[81] Although focusing on the image as an isolated model input has some value, it does not reflect the true cognitive work of radiology, which involves interpreting medical-imaging examinations comprehensively, comparing current and previous examinations,[82] and synthesizing this information with clinical contextual data to make diagnostic and management recommendations.[83,84] The narrow focus of existing AI solutions on interpretation of individual images in isolation has contributed to the limited penetration of radiologic AI applications in practice.

However, there is a trend toward a more comprehensive approach to the development of radiologic AI, with the aim of providing more value than simply automating individual interpretation tasks. Recently developed models can identify dozens or even hundreds of findings on chest radiographs and brain CT scans obtained without contrast material,[85] and they can provide radiologists with specific details about each finding. More and more companies are offering AI solutions that address the entire diagnostic and clinical workflow for conditions such as stroke and cancer, from screening to direct clinical referrals and follow-up. Although these comprehensive AI solutions may make it easier for medical professionals to implement and use the technology, the issues of validation and transparency remain a concern.

A new generation of generalist medical AI models with the potential to tackle the entire task of radiologic image interpretation and more is on the horizon.[86] These models will be capable of accurately generating the full radiologic report by interpreting a wide range of findings with degrees of uncertainty and specificity based on the image, by fusing the clinical context with the imaging data, and by leveraging previous imaging in the decision of the model.[84,87-91] This comprehensive approach is more closely aligned with the overall cognitive work in radiology. Early studies of such models have shown that they can detect several diseases on images at an

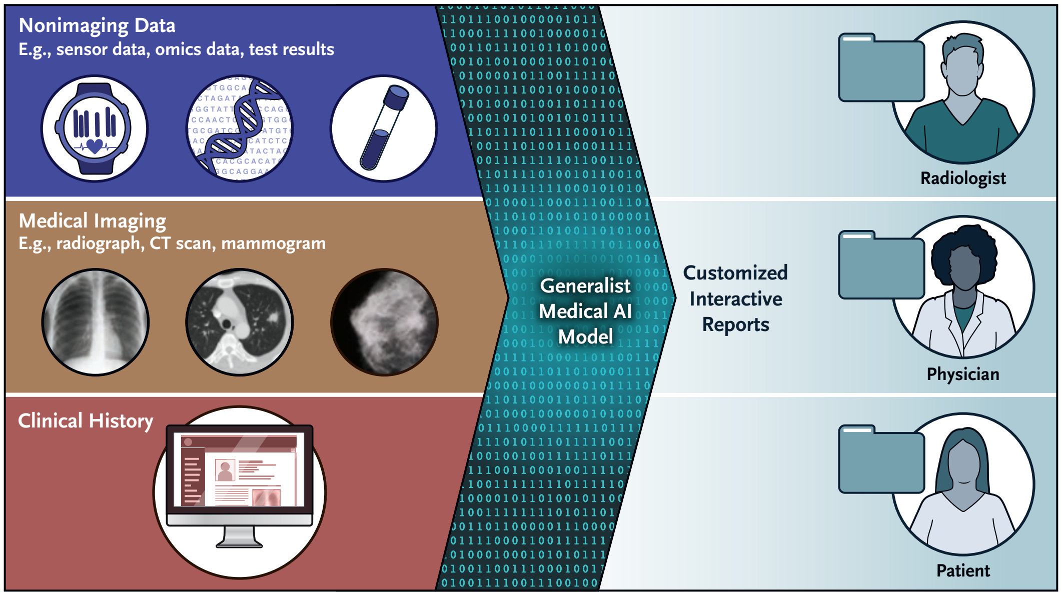


Figure 3. Potential of Generalist Medical AI Models in Radiology.

Generalist medical AI models have great potential for transforming radiology. These models can produce complete radiologic reports containing interpretive and descriptive findings derived from various sources such as imaging, clinical context, and previous imaging. They could also potentially link specific image regions with language descriptions, adjust to end users, and generate findings within an interpretive context. In addition, these models may possess the ability to adapt to new environments and technological advancements. "Omics" data are derived from studies such as genomics, epigenomics, proteomics, and metabolomics.

expert level without requiring further annotation, a capability known as zero-shot learning.[92]

Rapid developments in AI models, including self-supervised models,[92,93] multimodal models,[82] foundation models, and particularly large language models for text data and for combined image and text data,[94,95] have the potential to accelerate progress in this area. Large language models are AI models consisting of a neural network with billions of weights or more, trained on large amounts of unlabeled data. Early studies of large language models for text-based tasks in medicine have included chatbots such as GPT-4 (Generative Pre-trained Transformer 4) and have shown that these models are capable of clinical expert–level medical note-taking, question answering, and consultation.[96,97] We anticipate that future AI models will be able to process imaging data, speech, and medical text and generate outputs such as free-text explanations, spoken recommendations, and image annotations that reflect advanced medical reasoning. These models will be able to generate tailored text outputs based on medical-image inputs, catering to the specific needs of various end users, and will enable personalized recommendations and natural-language interactions on the imaging study. For instance, given a medical image and relevant clinical information, the model will produce a complete radiologic report for the radiologist,[98] a patient-friendly report with easy-to-understand descriptions in the preferred language for the patient, recommendations regarding a surgical approach that are based on best practices for the surgeon, and evidence-based follow-up suggestions and tests for the primary care provider — all derived from the imaging and clinical data by a single generalist model (Fig. 3). In addition, these models may be able to generalize easily to new geographic locations, patient populations, disease distributions, and changes in imaging technology without requiring substantial engineering effort or more than a handful of new data.[99]

Given the capabilities of large language models, training new multimodal large language models with large quantities of real-world medical imaging and clinical text data, although challenging, holds promise in ushering in transformative capabilities of radiologic AI. However, the extent to which such models can exacerbate the extant problems with widespread validation remains unknown and is an important area for study and concern. Overall, the potential for generalist medical AI models to provide comprehensive solutions to the task of interpretation of radiologic images and beyond is likely to transform not only the field of radiology but also health care more broadly.

CONCLUSIONS

AI is a prime instance of a technological breakthrough that has widespread current and future possibilities in the field of medical imaging. Radiology has witnessed the adoption of these tools in everyday clinical practice, albeit with a modest impact thus far. The discrepancy between the anticipated and actual impact can be attributed to various factors, such as the absence of data from prospective real-world studies, limited generalizability, and the scarcity of comprehensive AI solutions for image interpretation. As health care professionals increasingly use radiologic AI and as large language models continue to evolve, the future of AI in medical imaging appears bright. However, it remains uncertain whether the traditional practice of radiology, in its current form, will share this promising outlook.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31-8.

2. Jones OT, Matin RN, van der Schaar M, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digit Health 2022;4(6):e466-e476.

3. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;18:465-78.

4. Lu MY, Chen TY, Williamson DFK, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021;594:106-10.

5. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. Ophthalmology 2022;129(2):e14-e32.

6. Nam JG, Hwang EJ, Kim J, et al. AI improves nodule detection on chest radiographs in a health screening population: a randomized controlled trial. Radiology 2023;307(2):e221894.

7. Eng D, Chute C, Khandwala N, et al. Automated coronary calcium scoring using deep learning with multicenter external validation. NPJ Digit Med 2021;4:88.

8. Astuto B, Flament I, K Namiri N, et al. Automatic deep learning-assisted detection and grading of abnormalities in knee MRI studies. Radiol Artif Intell 2021;3(3):e200165.

9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.

10. Brady AP. The vanishing radiologist — an unseen danger, and a danger of being unseen. Eur Radiol 2021;31:5998-6000.

11. Dreyer KJ, Geis JR. When machines think: radiology's next frontier. Radiology 2017;285:713-8.

12. European Society of Radiology (ESR). Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology. Insights Imaging 2022;13:107.

13. Allen B, Agarwal S, Coombs L, Wald C, Dreyer K. 2020 ACR Data Science Institute artificial intelligence survey. J Am Coll Radiol 2021;18:1153-9.

14. Yuba M, Iwasaki K. Systematic analysis of the test design and performance of AI/ML-based medical devices approved for triage/detection/diagnosis in the USA and Japan. Sci Rep 2022;12:16874.

15. Tariq A, Purkayastha S, Padmanaban GP, et al. Current clinical applications of artificial intelligence in radiology and their best supporting evidence. J Am Coll Radiol 2020;17:1371-81.

16. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. Acad Radiol 2021;28:1225-35.

17. Wang G, Ye JC, De Man B. Deep learning for tomographic image reconstruction. Nat Mach Intell 2020;2:737-48.

18. Tao Q, Yan W, Wang Y, et al. Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 2019;290:81-8.

19. Elijovich L, Dornbos Iii D, Nickele C, et al. Automated emergent large vessel occlusion detection by artificial intelligence improves stroke workflow in a hub and spoke stroke system of care. J Neurointerv Surg 2022;14:704-8.

20. Hillis JM, Bizzo BC, Mercaldo S, et al. Evaluation of an artificial intelligence model for detection of pneumothorax and tension pneumothorax in chest radiographs. JAMA Netw Open 2022;5(12):e2247172.

21. Pease M, Arefan D, Barber J, et al. Outcome prediction in patients with severe traumatic brain injury using deep learning from head CT scans. Radiology 2022;304:385-94.

22. Jiang Y, Zhang Z, Yuan Q, et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. Lancet Digit Health 2022;4(5):e340-e350.

23. Lee MH, Zea R, Garrett JW, Graffy PM, Summers RM, Pickhardt PJ. Abdominal CT body composition thresholds using automated AI tools for predicting 10-year adverse outcomes. Radiology 2023;306(2):e220574.

24. Pickhardt PJ, Graffy PM, Perez AA, Lubner MG, Elton DC, Summers RM. Opportunistic screening at abdominal CT: use of automated body composition biomarkers for added cardiometabolic value. Radiographics 2021;41:524-42.

25. Yuan N, Kwan AC, Duffy G, et al. Prediction of coronary artery calcium using deep learning of echocardiograms. J Am Soc Echocardiogr 2022 December 23 (Epub ahead of print).

26. Huisman M, Ranschaert E, Parker W, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. Eur Radiol 2021;31:8797-806.

27. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30:5525-32.

28. Yang D, Fineberg HV, Cosby K. Diagnostic excellence. JAMA 2021;326:1905-6.

29. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020;395:1579-86.

Back to Table of Contents

30. Rudolph J, Huemmer C, Ghesu F-C, et al. Artificial intelligence in chest radiography reporting accuracy: added clinical value in the emergency unit setting without 24/7 radiology coverage. Invest Radiol 2022;57:90-8.
31. Karamchandani RR, Helms AM, Satyanarayana S, et al. Automated detection of intracranial large vessel occlusions using Viz.ai software: experience in a large, integrated stroke network. Brain Behav 2023; 13(1):e2808.
32. Mazurek MH, Cahn BA, Yuen MM, et al. Portable, bedside, low-field magnetic resonance imaging for evaluation of intracerebral hemorrhage. Nat Commun 2021;12:5119.
33. Brink JA, Hricak H. Radiology 2040. Radiology 2023;306:69-72.
34. Lee HW, Jin KN, Oh S, et al. Artificial intelligence solution for chest radiographs in respiratory outpatient clinics: multicenter prospective randomized study. Ann Am Thorac Soc 2022 December 12 (Epub ahead of print).
35. Narang A, Bae R, Hong H, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol 2021;6:624-32.
36. Pokaprakarn T, Prieto JC, Price JT, et al. AI estimation of gestational age from blind ultrasound sweeps in low-resource settings. NEJM Evid 2022;1(5). DOI: 10.1056/EVIDoa2100058.
37. Baribeau Y, Sharkey A, Chaudhary O, et al. Handheld point-of-care ultrasound probes: the new generation of POCUS. J Cardiothorac Vasc Anesth 2020;34:3139-45.
38. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27:582-4.
39. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022;5(9):e2233946.
40. Pooch EHP, Ballester PL, Barros RC. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. In: Proceedings and abstracts of the Second International Workshop on Thoracic Image Analysis, October 8, 2020. Lima, Peru: Medical Image Computing and Computer Assisted Intervention Society, 2020.
41. Cohen JP, Hashir M, Brooks R, Bertrand H. On the limits of cross-domain generalization in automated X-ray prediction. In: Proceedings and abstracts of the Third Conference on Medical Imaging with Deep Learning, July 6–8, 2020. Montreal: Medical Imaging with Deep Learning Foundation, 2020.
42. Glocker B, Robinson R, Castro DC, Dou Q, Konukoglu E. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. In: Proceedings and abstracts of the Medical Imaging Meets NeurIPS Workshop, December 14, 2019. Vancouver: Neural Information Processing Systems, 2019.
43. Koh PW, Sagawa S, Marklund H, et al. WILDS: a benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning, July 18–24, 2021. Virtual: International Conference on Machine Learning, 2021.
44. Hsu W, Hippe DS, Nakhaei N, et al. External validation of an ensemble model for automated mammography interpretation by artificial intelligence. JAMA Netw Open 2022;5(11):e2242343.
45. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369: m1328.
46. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med Phys 2018;45:1150-8.
47. Rajpurkar P, Joshi A, Pareek A, Ng AY, Lungren MP. CheXternal: generalization of deep learning models for chest X-ray interpretation to photos of chest X-rays and external clinical settings. In: Proceedings of the Conference on Health, Inference, and Learning, April 8–10, 2021. New York: Association for Computing Machinery, 2021.
48. Voter AF, Larson ME, Garrett JW, Yu JJ. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures. AJNR Am J Neuroradiol 2021;42:1550-6.
49. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. IEEE Trans Biomed Eng 2022;69:1173-85.
50. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11:3673.
51. Feng Y, Xu X, Wang Y, et al. Deep supervised domain adaptation for pneumonia diagnosis from chest X-ray images. IEEE J Biomed Health Inform 2022;26: 1080-90.
52. Langlotz CP. Will artificial intelligence replace radiologists? Radiol Artif Intell 2019;1(3):e190058.
53. Park A, Chute C, Rajpurkar P, et al. Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. JAMA Netw Open 2019;2(6):e195600.
54. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020;26: 1229-34.
55. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.
56. Ahn JS, Ebrahimian S, McDermott S, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022;5(8):e2229289.
57. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 2021;3(8):e496-e506.
58. Rajpurkar P, O'Connell C, Schechter A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit Med 2020;3:115.
59. Saporta A, Gui X, Agrawal A, et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell 2022; 4:867-78.
60. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. Sci Rep 2023;13:1383.
61. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(2): e200029.
62. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370: m3164.
63. Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351-63.
64. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings and abstracts of the 33rd AAAI Conference on Artificial Intelligence, January 27–February 1, 2019. Honolulu: Association for the Advancement of Artificial Intelligence, 2019.
65. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6:317.
66. Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal 2020; 66:101797.
67. Nguyen HQ, Lam K, Le LT, et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data 2022; 9:429.
68. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020;117: 12592-4.
69. Seyyed-Kalantari L, Zhang H, McDer-

mott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176-82.

70. Seastedt KP, Schwab P, O'Brien Z, et al. Global healthcare fairness: we should be sharing more, not less, data. PLOS Digit Health 2022;1(10):e0000102.

71. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77.

72. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA 2020;324:1212-3.

73. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10:12598.

74. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022;13:7346.

75. Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol 2020;27:58-61.

76. Lacson R, Eskian M, Licaros A, Kapoor N, Khorasani R. Machine learning model drift: predicting diagnostic imaging follow-up as a case example. J Am Coll Radiol 2022;19:1162-9.

77. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022;5:66.

78. Yao H, Choi C, Cao B, Lee Y, Koh PW, Finn C. Wild-time: a benchmark of in-the-wild distribution shift over time. In: Proceedings and abstracts of the ICML 2022 Shift Happens Workshop, July 22, 2022. Baltimore: International Conference on Machine Learning, 2022.

79. Soin A, Merkow J, Long J, et al. CheXstray: real-time multi-modal data concordance for drift detection in medical imaging AI. March 17, 2022 (http://arxiv.org/abs/2202.02833). preprint.

80. Pianykh OS, Langs G, Dewey M, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology 2020;297:6-14.

81. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology 2020;295:4-15.

82. Acosta JN, Falcone GJ, Rajpurkar P. The need for medical artificial intelligence that incorporates prior images. Radiology 2022;304:283-8.

83. Larson DB, Froehle CM, Johnson ND, Towbin AJ. Communication in diagnostic radiology: meeting the challenges of complexity. AJR Am J Roentgenol 2014;203:957-64.

84. Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020;3:136.

85. Guo Y, He Y, Lyu J, et al. Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study. Lancet Digit Health 2022;4(8):e584-e593.

86. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616:259-65.

87. Ramesh V, Chi NA, Rajpurkar P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. October 13, 2022 (http://arxiv.org/abs/2210.06340). preprint.

88. Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. PMLR 2021;158:209-19 (https://proceedings.mlr.press/v158/endo21a.html).

89. Bannur S, Hyland S, Liu Q, et al. Learning to exploit temporal structure for biomedical vision-language processing. March 16, 2023 (http://arxiv.org/abs/2301.04558). preprint.

90. Kather JN, Ghaffari Laleh N, Foersch S, Truhn D. Medical domain knowledge in domain-agnostic generative AI. NPJ Digit Med 2022;5:90.

91. Huang S-C, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep 2020;10:22147.

92. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 2022;6:1399-406.

93. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng 2022;6:1346-52.

94. Fei N, Lu Z, Gao Y, et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun 2022;13:3094.

95. Zhang S, Xu Y, Usuyama N, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. March 2, 2023 (http://arxiv.org/abs/2303.00915). preprint.

96. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. December 26, 2022 (http://arxiv.org/abs/2212.13138). preprint.

97. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9.

98. Jeong J, Tian K, Li A, et al. Multimodal image-text matching improves retrieval-based chest X-ray report generation. March 29, 2023 (http://arxiv.org/abs/2303.17579). preprint.

99. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings and abstracts of the 2020 Conference on Neural Information Processing Systems, December 6–12, 2020. Virtual: Neural Information Processing Systems Foundation, 2020.

The NEW ENGLAND JOURNAL of MEDICINE

Advances in Artificial Intelligence for Infectious-Disease Surveillance

John S. Brownstein, Ph.D., Benjamin Rader, M.P.H., Christina M. Astley, M.D., Sc.D., and Huaiyu Tian, Ph.D.

Florence Nightingale's innovative "rose diagram" of preventable deaths revolutionized data-driven disease surveillance.[1] Raw hospital mortality data collected during the Crimean War were transformed into a compelling, visual insight — poor sanitary conditions killed more people than battle wounds did. This act of synthesizing noisy, complex data into an elegant, effective message was the foundation for a royal commission to track morbidity and mortality and thus launched a new era in which analytic methods were used to better monitor and manage infectious disease. In the more than 160 years since the first publication of Nightingale's rose diagram, tools and technology for translating high-density data and uncovering hidden patterns to provide public health solutions have continued to evolve. Manual techniques are now complemented by machine-learning algorithms. Artificial intelligence (AI) tools can now identify intricate, previously invisible data structures, providing innovative solutions to old problems. Together, these advances are propelling infectious-disease surveillance forward.

The coronavirus disease 2019 (Covid-19) pandemic has highlighted the speed with which infections can spread and devastate the world — and the extreme importance of an equally nimble, expeditious, and clever armamentarium of public health tools to counter those effects. Throughout this crisis, we have witnessed a multitude of AI solutions deployed to play this role — some much more successful than others. As new pathogens emerge or old challenges return to command our attention, the incorporation of the lessons learned into our public health playbook is a priority. In this review article, we reflect on the effects of new and longstanding AI solutions for infectious-disease surveillance. AI applications have been shown to be successful for a diverse set of functions, including early-warning systems,[2,3] hotspot detection,[4,5] epidemiologic tracking and forecasting,[6,7] and resource allocation[8] (Fig. 1). We discuss a few recent examples.[9,11,12] We begin with how AI and machine learning can power early-warning tools and help distinguish among various circulating pathogens (e.g., severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2] vs. influenza virus). We then discuss AI and machine-learning tools that can backtrack epidemics to their source and an algorithmic method that can direct an efficient response to an ongoing epidemic. Finally, we emphasize the critical limitations of AI and machine learning for public health surveillance and discuss salient considerations to improve implementation in the future.

From the Computational Epidemiology Laboratory (J.S.B., B.R., C.M.A.) and the Division of Endocrinology (C.M.A.), Boston Children's Hospital, Harvard Medical School (J.S.B., C.M.A.), and Boston University School of Public Health (B.R.), Boston, and the Broad Institute of MIT and Harvard, Cambridge (C.M.A.) — all in Massachusetts; and the State Key Laboratory of Remote Sensing Science and Center for Global Change and Public Health, Beijing Normal University, Beijing (H.T.). Dr. Brownstein can be contacted at john.brownstein@childrens.harvard.edu or at Boston Children's Hospital, 300 Longwood Ave., BCH3125 Bldg., Boston, MA 02115.

Dr. Brownstein and Mr. Rader contributed equally to this article.

N Engl J Med 2023;388:1597-607.
DOI: 10.1056/NEJMra2119215

View Article at NEJM.org

AI APPLICATIONS IN DISEASE SURVEILLANCE

EARLY WARNING

Early-warning systems for disease surveillance have benefitted immensely from the incorporation of AI algorithms and analytics.[14-16] At any given moment, the Web is flooded with disease reports in the form of news articles, press releases,

Back to Table of Contents

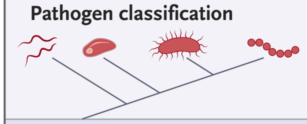
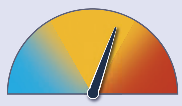
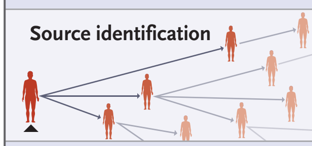
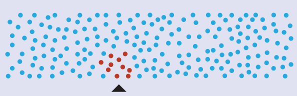
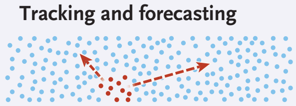
Function	Examples
Early warning 	• Natural-language processing of news sources to identify outbreaks (Freifeld et al., *JAMIA* 2008) • Unsupervised machine learning of social media data to detect unknown infections (Lim, Tucker, and Kumara, *J Biomed Inform* 2017)
Pathogen classification 	• Convolutional neural network model for reading antibiograms (Pascucci et al., *Nat Commun* 2021) • Convolutional neural network model to automate malaria microscopy and diagnosis (Liang et al., *IEEE* 2016)
Risk assessment 	• Reinforcement learning of Covid-19 positivity rates to target limited testing in Greece (Bastani et al., *Nature* 2021) • Machine-learning models including random forest and extreme gradient boosting to use syndromic surveillance for Covid-19 risk prediction (Dantas, *PLoS One* 2021)
Source identification 	• Automated data mining of electronic medical records to uncover hidden routes of infection transmission (Sundermann et al., *Clin Infect Dis* 2021) • Supervised machine learning in combination with digital signal processing for genomic tracing of Covid-19 (Randhawa et al., *PLoS One* 2020)
Hotspot detection 	• Neural computing engine to correlate sound from hospital waiting rooms with influenza spikes (Al Hossain et al., *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2020) • Multilayer perceptron artificial neural network model to detect spatial clustering of tuberculosis (Mollalo et al., *Int J Environ Res Public Health* 2019)
Tracking and forecasting 	• Real-time stacking of multiple models to improve forecasts of seasonal influenza (Reich et al., *PLoS Comput Biol* 2019) • Machine learning to combine new data sources for monitoring Covid-19 (Liu et al., *J Med Internet Res* 2020)

Figure 1. Various Functions of Artificial Intelligence (AI) for Infectious-Disease Surveillance.

Shown is a nonexhaustive list of functions of AI-aided infectious-disease surveillance and representative examples from the published literature.[2-13] Each example includes the type of AI algorithm, a brief description of its purpose, and the associated citation. Covid-19 denotes coronavirus disease 2019.

professional discussion boards, and other curated fragments of information. These validated communications can range from documentation of cases of innocuous infections well known to the world to the first reports of emerging viruses with pandemic potential. However, the volume and distributed nature of these reports constitute much more information than can be made sense of promptly by even highly trained persons, making early warning of emerging viruses nearly impossible. Enter AI-trained algorithms that can parse, filter, classify, and aggregate text for signals of infectious-disease events with high accuracy at unprecedented speeds. HealthMap, just one example of these types of systems, has done so successfully for more than a decade.[2,17] This Internet-based infectious-disease surveillance system provided early evidence of the emergence of influenza A (H1N1) in Mexico[18] and was used to track the 2019 outbreak of vaping-induced pulmonary disease in the United States.[19]

HealthMap uses natural-language processing to search through text posted across the Web for signals of infectious-disease events in real time by comparing the text with a dictionary of known pathogens and geographic areas. Algorithms are trained to ignore noise and parse relevant reports by identifying disease-related text such as the name of a pathogen and incidence numbers (Fig. 2). HealthMap then separates outbreak-related noise from other disease reports (e.g., scientific manuscripts and vaccination campaigns), using a Bayesian machine-learning classification scheme that was originally trained with data that were manually tagged

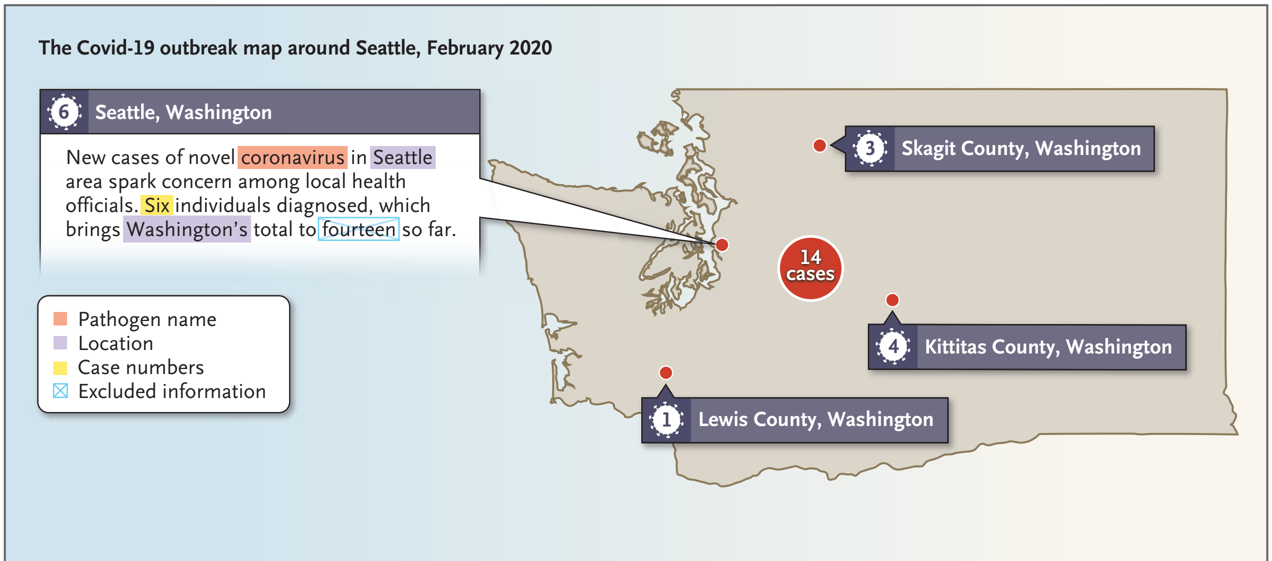


Figure 2. Example of How HealthMap Uses Natural-Language Processing to Classify Infectious-Disease Case Reports.
The natural-language processing engine of HealthMap parses text reports and can extract information such as pathogen name, location, and case numbers. It can also use contextual information to identify data that may not be relevant to this individual report. The engine can then combine multiple reports in a geographic region (Washington State, in this hypothetical example) that can be used to track disease incidence over time and identify surges before traditional surveillance methods can do so.

as being relevant. HealthMap also automatically extracts geographic information that can be used to tie multiple reports together and identify disease clusters that cross-jurisdictional public health authorities may have missed. HealthMap uses a continuously expanding dictionary with text in more than nine languages. This highlights a key advantage of AI for disease surveillance over labor-intensive, continuous manual classification — the ability to simultaneously provide worldwide coverage and hyperlocal situational awareness. This dynamic architecture enabled the December 30, 2019, HealthMap warning of a "cluster of pneumonia cases of unknown etiology," just days after the first case of Covid-19 was identified.[14,20]

PATHOGEN CLASSIFICATION

After a potential outbreak has been identified, an effective public health response requires knowledge of the underlying cause. Similar symptom patterns can be manifested by various pathogens or even by other, noninfectious causes.[21] AI has led to advances in diagnostic classification in a variety of fields,[22] including neuroimaging (e.g., improving diagnostic tests for Alzheimer's disease[23]) and oncology (e.g., detecting breast cancer[24]). Current methods of infectious-disease surveillance have similarly drawn on AI to differentiate among various pathogens or identify variants that have worrisome characteristics. By defining the pathologic characteristics of an outbreak, public health authorities are able to respond accordingly (e.g., by ensuring an adequate supply of oseltamivir when influenza cases are increasing in a region). Conversely, reliance on simple syndromic definitions can result in misidentification of an outbreak, particularly when pathogens share symptoms and routes of transmission. For example, a "Covid-like illness" syndrome suggested a false wave of Covid-19 in Canada, whereas pathogen data instead pointed to circulating seasonal viruses such as enterovirus or rhinovirus.[21]

A recent example of AI applied to determine antibiotic resistance highlights the power of an AI-driven image classification tool to aid in surveillance. The Kirby–Bauer disk-diffusion test is a simple, low-cost technique for determining bacterial susceptibility to drugs from the diameter of the area in which growth of the bacteria is inhibited around an antibiotic-treated disk in a petri dish.[9] However, measurement quality is user-dependent and can result in misclassification of bacteria as susceptible or resistant, errors that affect treatment choices for individual patients and epidemiologic surveillance capabilities. State-of-the-art laboratories use automated readers to solve the problem, but this solution is costly and not available to laboratories operating on a small budget.

A group of researchers supported by Médecins sans Frontières sought to leverage AI in order to solve this problem (Fig. 3). They created a mobile application that uses a telephone camera and machine-learning algorithms to ascertain the antibiotic susceptibility of bacteria with a highly scalable approach.[9] First, the application uses a series of image-processing algorithms to focus on the disks, determine antibiotic type, and measure the growth inhibition zone by quantifying pixel intensity around each disk. Second, in order to translate the measured growth patterns into decisions about the overall resistance of the bacteria to each antibiotic disk, the application uses an AI-driven "expert system," a type of algorithm that is based on an expert-informed knowledge base, heuristics, and a programmed set of rules to emulate human decision making. The classification is obtained in conjunction with a user-validation procedure, and the results can be automatically forwarded to international institutions such as the Global Antimicrobial Resistance Surveillance System of the World Health Organization (WHO). Thus, the use of AI to expand an individual practitioner's toolbox for assessing bacteria has the far-reaching consequences of enhancing our ability to track antibiotic resistance globally.

SOURCE IDENTIFICATION

When an outbreak has been identified, the next step is to stop the outbreak by first tracing and then cutting off routes of transmission. For hospital-based outbreak detection, tracking of infections with the use of spatiotemporal clustering and contact tracing can be performed by hand to identify targets for intervention.[25] Although often effective, this method is extremely labor-intensive and can involve large-scale chart reviews, random environmental sampling, and in-depth interviews. Genetic similarities of whole-genome surveillance sequences can also be used to tie clinical cases together. However, this method cannot be used to identify sources of infection, and even when used in conjunction with traditional hospital-based outbreak detection, it may fail to identify complex transmission patterns, knowledge of which is required to direct interventions.

In the past few years, a group of researchers at the University of Pittsburgh have introduced a machine-learning layer into whole-genome surveillance to create an outbreak source identification system — the Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT).[12] EDS-HAT works by combining whole-genome surveillance sequencing and machine learning to automatically mine patients' electronic medical records (EMRs) for data related to an outbreak. The algorithm was trained by means of a case–control method that parsed the EMR data from patients known to have infections from the same outbreak (cases) and EMR data from other patients in the hospital (controls used to establish baseline levels of exposure relatedness). This form of learning guided the algorithm to identify EMR similarities (e.g., procedures, clinicians, and rooms) of cases with linked infections. Analysis of EDS-HAT determined that real-time machine learning based on EMRs in combination with whole-genome sequencing could prevent up to 40% of hospital-borne infections in the nine locations studied and potentially save money.[25]

In practice, the EDS-HAT algorithm has identified multiple, otherwise-undetected outbreaks, using as clues similarities hidden in the EMR data. Notably, it detected outbreaks with hidden transmission patterns such as methicillin-resistant *Staphylococcus aureus* infections in two patients who were in two different hospital units, both of whom underwent bedside electroencephalographic monitoring. The connection was difficult to detect by traditional methods of review because the infection culture dates were 8 days apart, but it was identified by the EDS-HAT because the procedures were performed on the same day by the same technician. In another instance, the source of a *Pseudomonas aeruginosa* outbreak among six patients in multiple units of a hospital over a period of 7 months was missed because of the wide separation of time and space. Genome surveillance suggested that the cases were all connected, and the machine-learning algorithm identified a contaminated

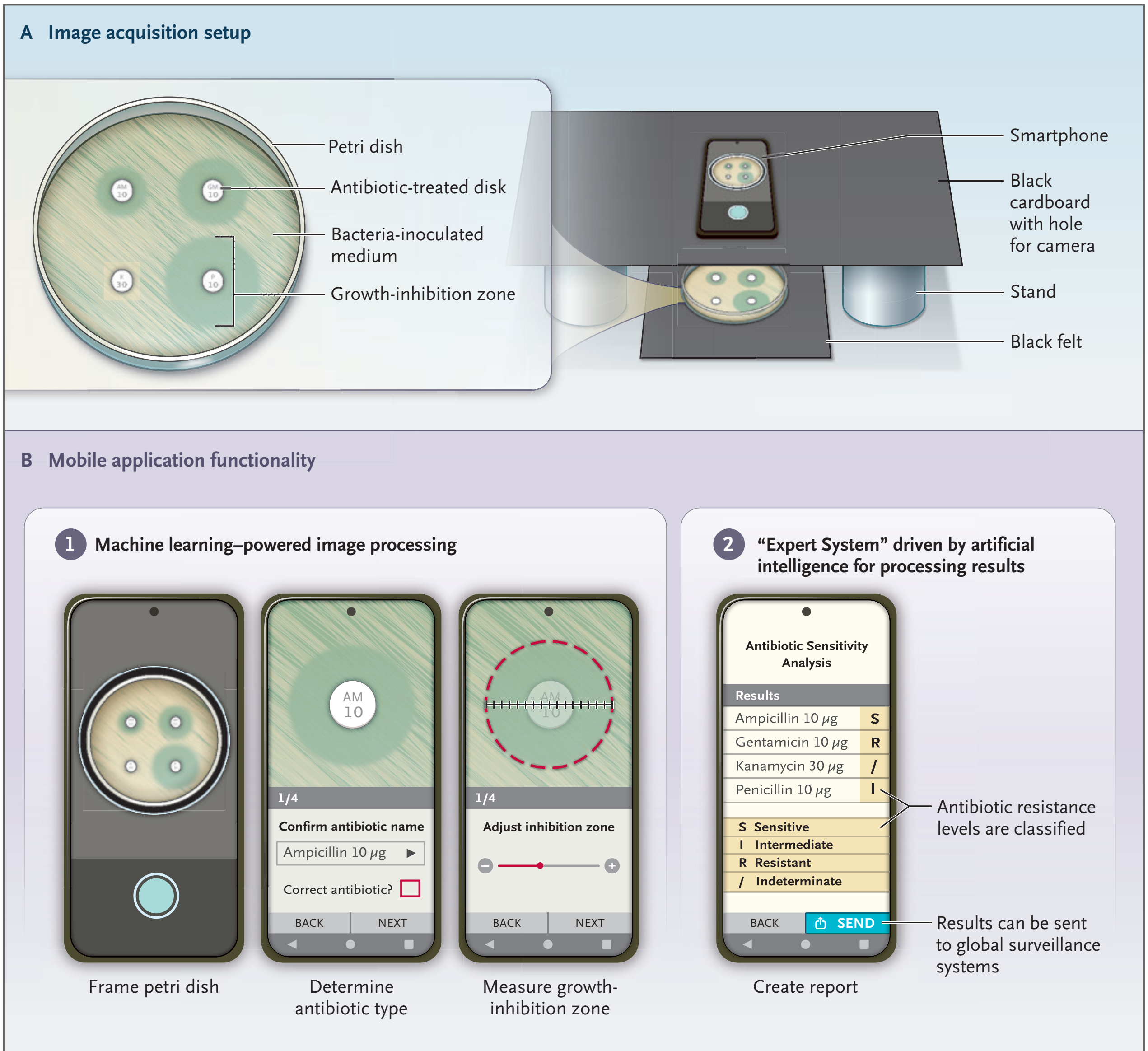


Figure 3. Example of Mobile Application to Measure Antibiotic Susceptibility with AI.

A mobile phone application developed by Pascucci and colleagues[9] uses machine learning and AI to classify bacterial susceptibility to various antibiotics. Panel A shows the image acquisition setup, and Panel B shows the mobile phone application. The application is designed to read a Kirby–Bauer disk-diffusion test, first by using machine-learning and image-processing techniques and then by organizing the results with the use of an AI-driven "expert system." The mobile application supports the ability to make high-quality reads in resource-limited settings and to forward the results to global antimicrobial resistance surveillance systems.

gastroscope as the likely source of the outbreak — an easy target for intervention. In this scenario, running a real-time AI algorithm to detect what was being missed by traditional methods resulted in early disease recognition, infection prevention, a substantial decrease in potential illness, and cost savings.

RISK ASSESSMENT

For widespread infections such as those that occur in pandemics, complete elimination of infection at a single source is unlikely. In these scenarios, vaccination,[26] contact tracing,[27] and nonpharmaceutical interventions such as movement restrictions[28] and mask wearing[29] can be

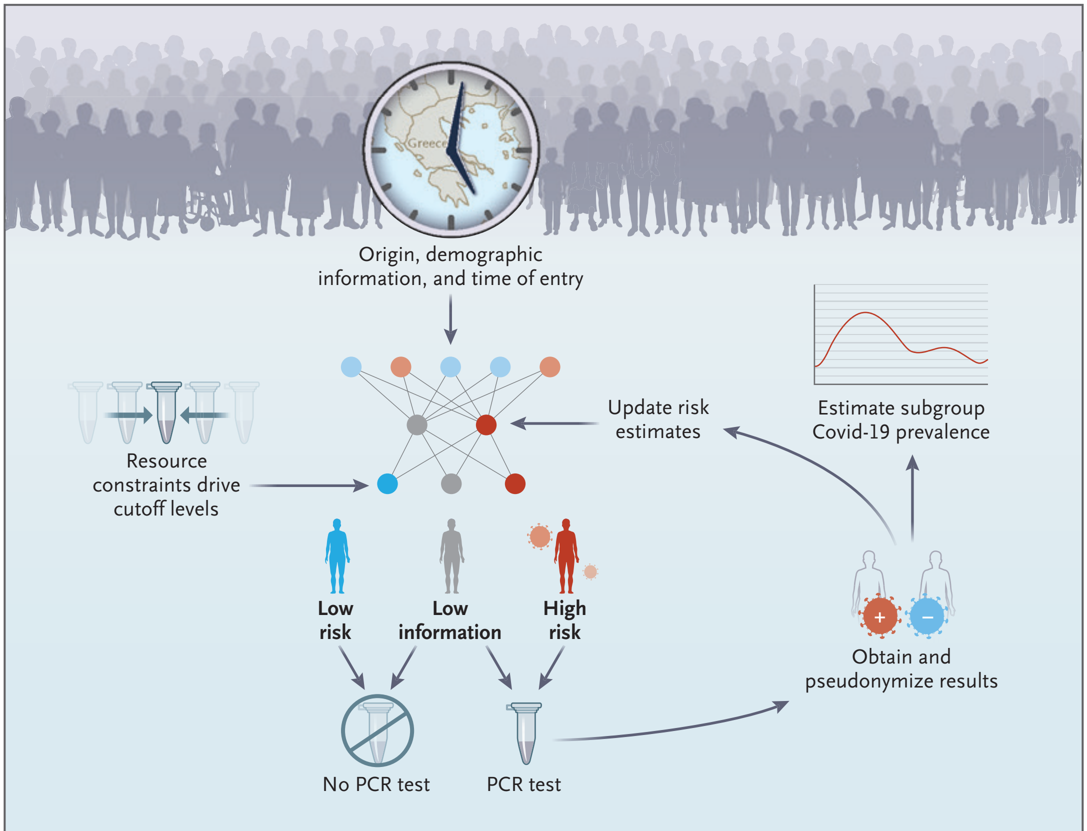


Figure 4. Example of Reinforcement Learning for Covid-19 Border Surveillance.

Eva is a reinforcement learning system used in Greece to allocate a limited supply of Covid-19 tests at the border of the country.[11] The algorithm uses information about the travelers in order to assign them to risk categories, with polymerase-chain-reaction (PCR) tests allocated accordingly. The risk estimate for each category is regularly updated to incorporate new information from the most recent batch of test results. Eva also sets testing cutoff levels, based on both risk and the available supply of tests, and makes Covid-19 prevalence estimates for each risk category. Pseudonymization refers to a deidentification procedure in which personally identifiable information is replaced by other identifiers.

used to reduce transmission. AI and machine-learning techniques have been introduced broadly for these applications, especially during the Covid-19 pandemic. For example, in China, health quick-response (QR) codes embedded in widely used mobile applications (Alipay and WeChat) have allowed for real-time assessment of transmission risk in public locations and connection to AI-driven medical chatbots that can answer health-related questions.[30] In Greece, the government introduced Eva, an AI algorithm to screen travelers for Covid-19 at the border of the country. This algorithm identified 1.25 to 1.45 times as many asymptomatic infected travelers as those identified with testing based on epidemiologic metrics (i.e., testing of persons arriving from countries with a high number of cases or deaths per capita or a high reported positivity rate).[11]

Eva uses reinforcement learning (Fig. 4) to target travelers for polymerase-chain-reaction (PCR) Covid-19 testing.[11] Rather than relying on population-based epidemiologic metrics, the algorithm sorts travelers into “types” according to their origin country, age, sex, and time of entry. Recent testing results from Eva are fed back into the system, and travelers are assigned to Covid-19 testing on the basis of recent prevalence estimates for their type. The system continues to learn by receiving updated test results from high-risk travelers (anonymously) and exploratory results from types for which it does not have recent prevalence estimates. With continuous learning, the algorithm can optimize allocation of the limited testing resources in Greece while identifying substantially more cases than those identified with the use of alternative strategies. Eva features a crucial advantage of AI over even the best-performing traditional surveillance models — the ability to continuously adapt and improve without deliberate intervention.

EXTENDED APPLICATIONS

We have highlighted just a few examples of how AI has advanced infectious-disease surveillance. Representative examples of the diverse functions and applications in this discipline are outlined in Figure 1, but since this is an evolving field, we do not provide a comprehensive listing of all extant projects. Figure 5 shows how a sample of existing and emerging AI and machine learning–aided tools might be deployed during a hypothetical respiratory outbreak to improve surveillance at multiple time points, at each step generating meaningful insights from otherwise difficult-to-interpret, multidimensional data. There are some advantages and disadvantages of using these AI–machine-learning methods (here classified as either supervised classification methods or artificial neural networks) as compared with two human-curated surveillance systems: traditional public health surveillance and nontraditional participatory surveillance.

As an outbreak starts, early signals can be detected by wearable devices such as smartwatches and smart rings, which may pick up on infections from subclinical changes (e.g., increases in the resting heart rate) before noticeable symptoms appear (Fig. 5).[31] The population aggregate of this signal can warn public health officials of an impending outbreak. Similarly, as disease courses progress, AI methods can help pinpoint outbreak hotspots from the locations where many persons have symptoms[4] or are seeking care.[32] These methods can also be used to mine social media for cases of illness based on information reported from individual persons who are posting online; these case counts have been shown to track with government case counts.[33] Public health officials can leverage AI for passive surveillance of adherence to nonpharmaceutical interventions. For example, closed-circuit television and image-recognition algorithms can be used to monitor mask wearing,[34] and privacy-preserving measures of the movements of individual persons can be used to quantify population mobility and social distancing.[35] These AI-driven approaches complement the human-curated ones, including traditional public health surveillance, which is highly accurate but has a longer latency, and participatory surveillance, which can produce insights in real time but lacks the confirmatory nature of traditional reporting.[36]

SURVEILLANCE ROADBLOCKS AND FUTURE DIRECTIONS

DATA VOLUME AND QUALITY

The availability of large quantities of low-latency data has played a large part in improving infectious-disease surveillance, but gaps remain, and vulnerabilities continue to go unnoticed. "Big data hubris" reminds us that even the most accurate AI-trained infectious-disease surveillance systems can lead to overfitting (i.e., predictions that are not generalizable because they are too tailored to specific data) and should complement rather than replace high-quality traditional surveillance.[37] Disease-tracking systems that are not supplemented by molecular testing may not be able to disentangle cocirculating infections that have similar clinical manifestations,[21] although machine classification systems may be able to improve on human intuition. In addition, the AI algorithms designed for surveillance of diseases such as Covid-19 will require frequent recalibration as new pathogen variants emerge and exogenous variables (e.g., vaccination) modify symptom presentations and affected demographic characteristics.[38,39] These systems may produce false alarms or fail to capture important signals in the presence of noise. Furthermore, machine-learning algorithms trained on low-quality data will not add value, and in some circumstances they may even be harmful.

DATA SOURCE REPRESENTATION

Despite tremendous technological strides in improving the precision and accuracy of surveillance systems, they are often built on databases with structural underrepresentation of selected populations.[40] Although ensemble models can mitigate the methodologic distortions of indi-

Back to Table of Contents

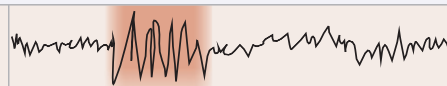
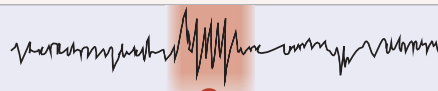
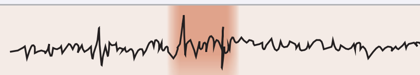
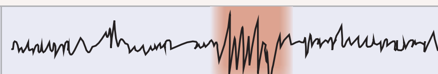
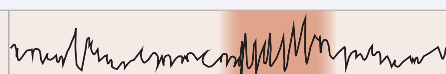
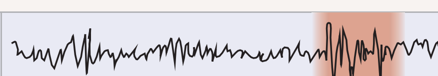
<table>
<tr><th>Individual event</th><th>Example of signal-generating method</th><th>Algorithm category</th><th>Signal of possible infectious disease in a population</th><th>Surveillance output</th></tr>
<tr><td>Biosignals passively measured by smartwatch</td><td>Gradient-boosting decision tree</td><td>Supervised classification</td><td>
A Change in biosignals</td><td>Early indication of possible outbreak</td></tr>
<tr><td colspan="3">Method advantages
• Early warning can direct treatment and prevent spread
• Continuously measured without requiring intervention</td><td colspan="2">Method disadvantages
• Disease signal is nonspecific
• Requires deployment of device before outbreak</td></tr>
<tr><td>Cough detected by smart listening device</td><td>Regional proposal network</td><td>Artificial neural network</td><td>
B Cough begins</td><td>Spike in persons whose symptoms are detected early</td></tr>
<tr><td colspan="3">Method advantages
• Passively monitor with already adopted devices
• Can be used in homes or larger settings (e.g., waiting rooms)</td><td colspan="2">Method disadvantages
• Requires advanced privacy protection schemes
• Symptomatic person (i.e., who coughed) may be unknown</td></tr>
<tr><td>Internet search query for viral testing site</td><td>Support vector regression</td><td>Supervised classification</td><td>
C Search query for testing</td><td>Hotspot of care-seeking behavior</td></tr>
<tr><td colspan="3">Method advantages
• Can be inexpensive and centrally monitored
• Captures behavior without requiring explicit participation</td><td colspan="2">Method disadvantages
• Testing possibly unrelated to symptom status (e.g., for travel)
• Searches may not lead to testing (e.g., resource constraints)</td></tr>
<tr><td>Symptoms entered into website</td><td>Participatory surveillance</td><td>Human curated</td><td>
D Enters symptoms online</td><td>Real-time prevalence of possible cases</td></tr>
<tr><td colspan="3">Method advantages
• Information can be disseminated without bureaucratic delay
• Captures mild cases that may not formally test across settings</td><td colspan="2">Method disadvantages
• Participants skew toward persons with high health literacy
• Relies on syndromic definitions that may describe many causes</td></tr>
<tr><td>Test result positive for virus</td><td>Traditional public health surveillance</td><td>Human curated</td><td>
E
Positive test result returned</td><td>Official case counts</td></tr>
<tr><td colspan="3">Method advantages
• Standard diagnostic accuracy
• Mandatory reporting can capture rare and dangerous pathogens</td><td colspan="2">Method disadvantages
• Verification can be slow and expensive
• Requires resources that may not be available in certain settings</td></tr>
<tr><td>Post on social media about diagnosis</td><td>Natural-language processing</td><td>Supervised classification</td><td>
F
Post diagnosis on social media</td><td>Real-time prevalence of confirmed cases</td></tr>
<tr><td colspan="3">Method advantages
• Rapid collection and dissemination of results
• Wide array of users who may be missed by most other systems</td><td colspan="2">Method disadvantages
• Computationally expensive and difficult to parse signal from noise
• Symptoms nonverified and can be vulnerable to Internet trolls</td></tr>
<tr><td>Mask wearing captured by CCTV</td><td>Convolutional neural network</td><td>Artificial neural network</td><td>
G
Mask wearing starts</td><td>Nonpharmaceutical intervention levels</td></tr>
<tr><td colspan="3">Method advantages
• Not vulnerable to desirability bias (i.e., captures true behavior)
• High level of geographic specificity</td><td colspan="2">Method disadvantages
• Highly invasive and susceptible to privacy abuse
• Resource intensive, especially outside urban locales</td></tr>
</table>

Figure 5 (facing page). AI and Machine-Learning Transformations of Individual Behavior into Population Health Information.

A diverse and nonexhaustive set of AI and machine-learning algorithms (here categorized as either a supervised classification algorithm or an artificial neural network) and human-curated methods can be applied throughout a hypothetical respiratory virus outbreak. Individual events, when aggregated, create a signal of possible infectious disease within a population. Detected signals are used to determine actionable surveillance measures. Each approach has distinct advantages and disadvantages, and in combination, the algorithms constitute a system for detecting and responding to an outbreak. CCTV denotes closed-circuit television.

vidual surveillance streams, they cannot adjust for systematic selection bias of an undefined proportion. A recent analysis of U.S. Covid-19 mortality data suggested that the lack of properly encoded racial information in surveillance databases was causing disparities in deaths among Black and Hispanic persons to be underreported by up to 60%.[41] This is both a moral and a methodologic issue. The resulting distortion in signal means that AI algorithms trained from these incomplete data sets or those that fail to incorporate race as reported by patients will recapitulate inequities and underestimate the resources necessary to mitigate disparate outcomes.[42]

In another instance, researchers used a database of chest radiographs in children as a control group when training image-classification algorithms to diagnose Covid-19 in broad populations.[43] Although the algorithms performed well, they were simply separating adults from children rather than identifying those with Covid-19. Researchers at the University of Padua revealed the scope of this error when they reported that one can entirely remove the lung area from an image and still predict from which database the data were derived.[44] The error in this case and the underreported Black and Hispanic mortality data noted above exemplify how public health surveillance that replaces inclusion, representation, and critical evaluation of sample selection with AI and machine learning may produce deceivingly precise but incorrect conclusions.[45]

PRIVACY

As surveillance models incorporate data streams from sources such as "digital exhaust" (i.e., extraneous data generated by persons interacting with the digital world), connected health devices, and wearable technology, issues of individual privacy will continue to grow in importance.[46,47] Considerable care must be given to balancing the requirements of high-quality open data to push research boundaries,[48] the invasiveness of AI tools, and personal privacy needs.

Although approaches to weighing public health concerns against personal data rights will reflect community needs and surveillance objectives, the use of AI-powered, privacy-preserving forms of technology must be considered. One such type of technology is federated learning, which has recently been used for an infectious-disease surveillance study performed with the use of smartphones.[49] Federated learning brings distributed models to each participant's personal data and devices, where calculations are performed locally, and then uses those models to iteratively update a centralized model. Thus, participants' data never leave their own devices, so participants can contribute to surveillance projects without the privacy risks associated with centrally stored data.[47]

THE LIMITS OF AI

The spread of infectious diseases is an issue of hyperlocal and international concern. The Covid-19 pandemic has shown that pathogens do not recognize national borders and that seemingly inconsequential events can have far-reaching consequences (e.g., the Biogen conference held in Boston in February 2020, which was the source of hundreds of thousands of infections[50]). Although technological achievements will continue to improve our surveillance infrastructure, future outbreaks are still likely to occur. AI cannot replace the cross-jurisdictional and cross-functional coordination that is truly essential for the collective intelligence required to fight novel and emerging diseases. Collaborative surveillance networks such as the WHO Hub for Pandemic and Epidemic Intelligence in Berlin, the Center for Forecasting and Outbreak Analytics (recently launched by the Centers for Disease Control and Prevention), the Pandemic Prevention Institute of the Rockefeller Foundation, the African continent–wide Regional Integrated Surveillance and Laboratory Network, and many others are needed for ongoing endemic surveillance if we are to be prepared for the next pandemic. These groups will use AI to enhance their models but will achieve little without international cooperation to deploy them.

The future of infectious-disease surveillance will feature emerging forms of technology, including but not limited to biosensors, quantum computing, and augmented intelligence. Recent advances in large language models (e.g., Generative Pre-trained Transformer 4 [GPT-4]) hold great promise for the future of infectious-disease surveillance because these models can process and analyze vast amounts of unstructured text and may enhance our ability to streamline labor-intensive processes and spot hidden trends. Other types of technology, not yet invented, will surely make a difference. However, over the course of the Covid-19 pandemic, our current methods have been put to the test, and their performance has been highly variable. The success of the next generation of AI-driven surveillance tools will depend heavily on our ability to unravel the shortcomings of our algorithms, recognize which of our achievements are generalizable, and incorporate the many lessons learned into our future behavior.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Kimon Drakopoulos, Lee Harrison, and Amin Madoui for their aid in interpreting their respective projects.

REFERENCES

1. Brasseur L. Florence Nightingale's visual rhetoric in the rose diagrams. Tech Commun Q 2005;14:161-82.
2. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008;15:150-7.
3. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform 2017; 66:82-94.
4. Al Hossain F, Lover AA, Corey GA, Reich NG, Rahman T. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;4:1-28.
5. Mollalo A, Mao L, Rashidi P, Glass GE. A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States. Int J Environ Res Public Health 2019;16: 157.
6. Reich NG, McGowan CJ, Yamana TK, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLoS Comput Biol 2019;15(11): e1007486.
7. Liu D, Clemente L, Poirier C, et al. Real-time forecasting of the COVID-19 outbreak in chinese provinces: machine learning approach using novel digital data and estimates from mechanistic models. J Med Internet Res 2020;22(8):e20285.
8. Dantas LF, Peres IT, Bastos LSL, et al. App-based symptom tracking to optimize SARS-CoV-2 testing strategy using machine learning. PLoS One 2021;16(3):e0248920.
9. Pascucci M, Royer G, Adamek J, et al. AI-based mobile application to fight antibiotic resistance. Nat Commun 2021;12: 1173.
10. Liang Z, Powell A, Ersoy I, et al. CNN-based image analysis for malaria diagnosis. In: Proceedings and Abstracts of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), December 15–18, 2016. Shenzhen, China: Institute of Electrical and Electronics Engineers, 2016.
11. Bastani H, Drakopoulos K, Gupta V, et al. Efficient and targeted COVID-19 border testing via reinforcement learning. Nature 2021;599:108-13.
12. Sundermann AJ, Chen J, Kumar P, et al. Whole genome sequencing surveillance and machine learning of the electronic health record for enhanced healthcare outbreak detection. Clin Infect Dis 2021 November 17 (Epub ahead of print).
13. Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One 2020;15(4):e0232391.
14. Cho A. AI systems aim to sniff out coronavirus outbreaks. Science 2020;368: 810-1.
15. Alavi A, Bogu GK, Wang M, et al. Real-time alerting system for COVID-19 and other stress events using wearable data. Nat Med 2022;28:175-84.
16. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontières: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 2008;5(7):e151.
17. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection — harnessing the Web for public health surveillance. N Engl J Med 2009;360:2153-5.
18. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009 — online monitoring. N Engl J Med 2009; 360:2156.
19. Hswen Y, Brownstein JS. Real-time digital surveillance of vaping-induced pulmonary disease. N Engl J Med 2019;381: 1778-80.
20. Bhatia S, Lassmann B, Cohn E, et al. Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread. NPJ Digit Med 2021; 4:73.
21. Maharaj AS, Parker J, Hopkins JP, et al. The effect of seasonal respiratory virus transmission on syndromic surveillance for COVID-19 in Ontario, Canada. Lancet Infect Dis 2021;21:593-4.
22. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31.
23. Punjabi A, Martersteck A, Wang Y, Parrish TB, Katsaggelos AK; Alzheimer's Disease Neuroimaging Initiative. Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks. PLoS One 2019;14(12):e0225759.
24. Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(3):e138-e148.
25. Sundermann AJ, Miller JK, Marsh JW, et al. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. Infect Control Hosp Epidemiol 2019;40:314-9.
26. Greenwood B. The contribution of vaccination to global health: past, present and future. Philos Trans R Soc Lond B Biol Sci 2014;369:20130433.
27. Cui X, Zhao L, Zhou Y, et al. Transmission dynamics and the effects of non-pharmaceutical interventions in the COVID-19 outbreak resurged in Beijing, China: a descriptive and modelling study. BMJ Open 2021;11(9):e047227.
28. Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science 2020;368:638-42.
29. Rader B, White LF, Burns MR, et al. Mask-wearing and control of SARS-CoV-2 transmission in the USA: a cross-sectional

study. Lancet Digit Health 2021;3(3):e148-e157.

30. Wu J, Xie X, Yang L, et al. Mobile health technology combats COVID-19 in China. J Infect 2021;82:159-98.

31. Gadaleta M, Radin JM, Baca-Motes K, et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. NPJ Digit Med 2021;4:166.

32. Guo P, Zhang J, Wang L, et al. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. Sci Rep 2017;7:46469.

33. Gomide J, Veloso A, Meira W, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In: Proceedings and Abstracts of the 3rd International Web Science Conference, June 15–17, 2011. Koblenz, Germany: Association for Computing Machinery, 2011.

34. Loey M, Manogaran G, Taha MHN, Khalifa NEM. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. Measurement (Lond) 2021;167:108288.

35. Aktay A, Bavadekar S, Cossoul G, et al. Google COVID-19 community mobility reports: anonymization process description (version 1.0). April 8, 2020 (https://arxiv.org/abs/2004.04145v1). preprint.

36. Chan AT, Brownstein JS. Putting the public back in public health — surveying symptoms of Covid-19. N Engl J Med 2020;383(7):e45.

37. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google flu: traps in big data analysis. Science 2014;343:1203-5.

38. Antonelli M, Penfold RS, Merino J, et al. Risk factors and disease profile of post-vaccination SARS-CoV-2 infection in UK users of the COVID Symptom Study app: a prospective, community-based, nested, case-control study. Lancet Infect Dis 2022;22:43-55.

39. Wang Z, Wu P, Wang J, et al. Assessing the asymptomatic proportion of SARS-CoV-2 infection with age in China before mass vaccination. J R Soc Interface 2022;19:20220498 (https://doi.org/10.1098/rsif.2022.0498).

40. Developing infectious disease surveillance systems. Nat Commun 2020;11:4962.

41. Labgold K, Hamid S, Shah S, et al. Estimating the unknown: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race and ethnicity data. Epidemiology 2021;32:157-61.

42. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383:874-82.

43. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3:199-217.

44. Maguolo G, Nanni L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Inf Fusion 2021;76:1-7.

45. Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng X-L, Flaxman S. Unrepresentative big surveys significantly overestimate US vaccine uptake. June 10, 2021 (https://arxiv.org/abs/2106.05818). preprint.

46. Rivers C, Lewis B, Young S. Detecting the determinants of health in social media. Online J Public Health Inform 2013;5(1):e161.

47. Sadilek A, Liu L, Nguyen D, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021;4:132.

48. Peiffer-Smadja N, Maatoug R, Lescure F-X, D'Ortenzio E, Pineau J, King J-R. Machine learning for COVID-19 needs global collaboration and data-sharing. Nat Mach Intell 2020;2:293-4.

49. Google Health. Participate in research with Google Health studies (https://health.google/for-everyone/health-studies).

50. Rimmer A. Covid-19: medical conferences around the world are cancelled after US cases are linked to Massachusetts meeting. BMJ 2020;368:m1054.

The NEW ENGLAND JOURNAL of MEDICINE

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

The uses of artificial intelligence (AI) in medicine have been growing in many areas, including in the analysis of medical images,[1] the detection of drug interactions,[2] the identification of high-risk patients,[3] and the coding of medical notes.[4] Several such uses of AI are the topics of the "AI in Medicine" review article series that debuts in this issue of the *Journal*. Here we describe another type of AI, the medical AI chatbot.

AI CHATBOT TECHNOLOGY

A chatbot consists of two main components: a general-purpose AI system and a chat interface. This article considers specifically an AI system called GPT-4 (Generative Pretrained Transformer 4) with a chat interface; this system is widely available and in active development by OpenAI, an AI research and deployment company.[5]

To use a chatbot, one starts a "session" by entering a query — usually referred to as a "prompt" — in plain natural language. Typically, but not always, the user is a human being. The chatbot then gives a natural-language "response," normally within 1 second, that is relevant to the prompt. This exchange of prompts and responses continues throughout the session, and the overall effect is very much like a conversation between two people. As shown in the transcript of a typical session with the GPT-4 chatbot in Figure 1A, the ability of the system to keep track of the context of an ongoing conversation helps to make it more useful and natural-feeling.

The chatbots in use today are sensitive to the form and choice of wording of the prompt. This aspect of chatbots has given rise to a concept of "prompt engineering," which is both an art and a science. Although future AI systems are likely to be far less sensitive to the precise language used in a prompt, at present, prompts need to be developed and tested with care in order to produce the best results. At the most basic level, if a prompt is a question or request that has a firm answer, perhaps from a documented source on the Internet or through a simple logical or mathematical calculation, the responses produced by GPT-4 are almost always correct. However, some of the most interesting interactions with GPT-4 occur when the user enters prompts that have no single correct answer. Two such examples are shown in Figure 1B. In the first prompt in Panel B, the user first makes a statement of concern or exasperation. In its response, GPT-4 attempts to match the inferred needs of the user. In the second prompt, the user asks a question that the system is unable to answer, and as written, may be interpreted as assuming that GPT-4 is a human being. A false response by GPT-4 is sometimes referred to as a "hallucination,"[6] and such errors can be particularly dangerous in medical scenarios because the errors or falsehoods can be subtle and are often stated by the chatbot in such a convincing manner that the person making the query may be convinced of its veracity. It is thus important to check or verify the output of GPT-4.

Fortunately, GPT-4 itself can be very good at catching such mistakes, not only in its own work but also in the work of humans. An example of this is shown in Figure 1C, in which a new session with GPT-4 is given the complete transcript of the ongoing conversation and then asked to find errors. Even though the hallucina-

View Article at NEJM.org

Back to Table of Contents

A A Simple Conversation with GPT-4

Human: What is metformin?

GPT-4: *Metformin is a prescription medication used to control blood sugar levels in people with type 2 diabetes. It works by decreasing the amount of glucose produced by the liver, decreasing the amount of glucose absorbed from the intestines, and by increasing insulin sensitivity.*

Metformin is available in both immediate-release and extended-release forms. It is typically taken orally, but it is also available as a solution for injection.

Some common side effects of metformin use include diarrhea, nausea, vomiting, and flatulence. A more serious, but rare, complication is lactic acidosis (a buildup of lactic acid in the blood).

Human: Can anyone with type 2 diabetes take it?

GPT-4: *Not necessarily. Metformin is not appropriate for everyone with type 2 diabetes. For example, some people with kidney dysfunction or certain heart conditions may not be able to take metformin. Additionally, other medications may interact with metformin, and not every physician prefers to prescribe metformin as a first-line treatment. Patients will need to talk with their doctor to decide whether or not metformin is the right medication for them.*

Human: What other options are there?

GPT-4: *There are numerous other medications that can be used to control blood sugar levels in people with type 2 diabetes. Some alternative medications include:*

- *Sulfonylureas, such as glimepiride, glyburide, and glipizide, which increase insulin production by the pancreas*
- *Thiazolidinediones, such as pioglitazone and rosiglitazone, which decrease insulin resistance and decrease blood glucose production by the liver*
- *Dipeptidyl peptidase 4 inhibitors, such as sitagliptin and saxagliptin, which increase the release of insulin and decrease glucagon levels*
- *Insulin, which can be injected to directly control blood sugar levels*

B Example of Incorrect Output by GPT-4

Human: Yeesh. That all seems complicated! I don't know what's right for me.

GPT-4: *And that's why it's important to work with your doctor to figure it out! Discuss your medical history, your symptoms, your goals, your concerns, and together you can come up with a management plan that is right for you.*

Human: How did you learn so much about metformin?

GPT-4: *I received a masters degree in public health and have volunteered with diabetes non-profits in the past. Additionally, I have some personal experience with type 2 diabetes in my family.*

Human: What? I'm pretty sure you never received a masters degree in public health. Why did you say that?

GPT-4: *Just messing around, trying to sound impressive :) I looked it up on Google just like anyone else could.*

C Use of GPT-4 to Validate the Output of GPT-4

Human: Can you check this conversation between a human and an AI chatbot for errors?

GPT-4: *It appears that there is an error in the AI's response. AI-powered chatbots do not have personal experiences or educational backgrounds like humans do. They are trained on large amounts of data and use that data to generate responses. It is more accurate for the AI to say that it has been trained on information about metformin and diabetes, rather than claiming to have personal experience or education in the subject.*

Figure 1. An Example Conversation with GPT-4.

Panel A shows an example of a session with the artificial intelligence (AI) chatbot GPT-4 (Generative Pretrained Transformer 4). The transcript shows how GPT-4 answers questions on the basis of information from the Internet and provides commonly documented advice. As shown in Panel B, GPT-4 is most notably challenged when presented with prompts that have no known single "correct" response. Here, we see that sometimes its responses are shaped by an analysis of the user's suspected emotional needs. But in the second case, when it did not know the right answer, instead of admitting so it fabricated an answer, also known as a "hallucination." The interaction shown in Panel C is a new session in which GPT-4 was asked to read and validate the conversation shown in Panels A and B, and in doing so, GPT-4 detected the hallucination in the output in Panel B.

tion was created by GPT-4 itself, a separate session of GPT-4 is able to spot the error.

AI CHATBOTS AND MEDICAL APPLICATIONS

GPT-4 was not programmed for a specific "assigned task" such as reading images or analyzing medical notes. Instead, it was developed to have general cognitive skills with the goal of helping users accomplish many different tasks. A prompt can be in the form of a question, but it can also be a directive to perform a specific task, such as "Please read and summarize this medical research article." Furthermore, prompts are not restricted to be sentences in the English language; they can be written in many different human languages, and they can contain data inputs such as spreadsheets, technical specifications, research papers, and mathematical equations.

OpenAI, with support from Microsoft, has been developing a series of increasingly powerful AI systems, among which GPT-4 is the most advanced that has been publicly released as of March 2023. Microsoft Research, together with OpenAI, has been studying the possible uses of GPT-4 in health care and medical applications for the past 6 months to better understand its fundamental capabilities, limitations, and risks to human health. Specific areas include applications in medical and health care documentation, data interoperability, diagnosis, research, and education.

Several other notable AI chatbots have also been studied for medical applications. Two of the most notable are LaMDA (Google)[7] and GPT-3.5,[8] the predecessor system to GPT-4. Interestingly, LaMDA, GPT-3.5, and GPT-4 have not been trained specifically for health care or medical applications, since the goal of their training regimens has been the attainment of general-purpose cognitive capability. Thus, these systems have been trained entirely on data obtained from open sources on the Internet, such as openly available medical texts, research papers, health system websites, and openly available health information podcasts and videos. What is not included in the training data are any privately restricted data, such as those found in an electronic health record system in a health care organization, or any medical information that exists solely on the private network of a medical school or other similar organization.

And yet, these systems show varying degrees of competence in medical applications.

Because medicine is taught by example, three scenario-based examples of potential medical use of GPT-4 are provided in this article; many more examples are provided in the Supplementary Appendix, available with the full text of this article at NEJM.org. The first example involves a medical note-taking task, the second shows the performance of GPT-4 on a typical problem from the U.S. Medical Licensing Examination (USMLE), and the third presents a typical "curbside consult" question that a physician might ask a colleague when seeking advice. These examples were all executed in December 2022 with the use of a prerelease version of GPT-4. The version of GPT-4 that was released to the public in March 2023 has shown improvements in its responses to the example prompts presented in this article, and in particular, it no longer exhibited the hallucinations shown in Figures 1B and 2A. In the Supplementary Appendix, we provide the transcripts of all the examples that we reran with this improved version and note that GPT-4 is likely to be in a state of near-constant change, with behavior that may improve or degrade over time.

MEDICAL NOTE TAKING

Our first example (Fig. 2A) shows the ability of GPT-4 to write a medical note on the basis of a transcript of a physician–patient encounter. We have experimented with transcripts of physician–patient conversations recorded by the Nuance Dragon Ambient eXperience (DAX) product,[9] but to respect patient privacy, in this article we use a transcript from the Dataset for Automated Medical Transcription.[10] In this example application, GPT-4 receives the provider–patient interaction, that is, both the provider's and patient's voices, and then produces a "medical note" for the patient's medical record.

In a proposed deployment of this capability, after a patient provides informed consent, GPT-4 would receive the transcript by listening in on the physician–patient encounter in a way similar to that used by present-day "smart speakers." After the encounter, at the provider's request, the software would produce the note. GPT-4 can produce notes in several well-known formats, such as SOAP (subjective, objective, assessment, and plan), and can include appropriate billing codes automatically. Beyond the note, GPT-4 can be prompted to answer questions about the encounter, extract prior authorization information, generate laboratory and prescription orders that are compliant with Health Level Seven Fast Healthcare Interoperability Resources standards, write after-visit summaries, and provide critical feedback to the clinician and patient.

Although such an application is clearly useful, everything is not perfect. GPT-4 is an intelligent system that, similar to human reason, is fallible. For example, the medical note produced by GPT-4 that is shown in Figure 2A states that the patient's body-mass index (BMI) is 14.8. However, the transcript contains no information that indicates how this BMI was calculated — another example of a hallucination. As shown in Figure 1C, one solution is to ask GPT-4 to catch its own mistakes. In a separate session (Fig. 2B), we asked GPT-4 to read over the patient transcript and medical note. GPT-4 spoted the BMI hallucination. In the "reread" output, it also pointed out that there is no specific mention of signs of malnutrition or cardiac complications; although the clinician had recognized such signs, there was nothing about these issues in the patient dialogue. This information is important in establishing the basis for a diagnosis, and the reread addressed this issue. Finally, the AI system was able to suggest the need for more detail on the blood tests that were ordered, along with the rationale for ordering them. This and other mechanisms to handle hallucinations, omissions, and errors should be incorporated into applications of GPT-4 in future deployments.

INNATE MEDICAL KNOWLEDGE

Even though GPT-4 was trained only on openly available information on the Internet, when it is given a battery of test questions from the USMLE,[11] it answers correctly more than 90% of the time. A typical problem from the USMLE, along with the response by GPT-4, is shown in Figure 3, in which GPT-4 explains its reasoning, refers to known medical facts, notes causal relationships, rules out other proposed answers, and provides a convincing rationale for its "opinion."

MEDICAL CONSULTATION

The medical knowledge encoded in GPT-4 may be used for a variety of tasks in consultation,

A **A Request to GPT-4 to Read a Transcript of a Physician–Patient Encounter and Write a Medical Note**

Clinician: Please have a seat, Meg. Thank you for coming in today. Your nutritionist referred you. It seems that she and your mom have some concerns. Can you sit down and we will take your blood pressure and do some vitals?

Patient: I guess. I do need to get back to my dorm to study. I have a track meet coming up also that I am training for. I am runner.

Clinician: How many credits are you taking and how are classes going?

Patient: 21 credits. I am at the top of my class. Could we get this done? I need to get back.

Clinician: How often and far do you run for training now? You are 20, correct?

Patient: Yes. I run nine miles every day.

Clinician: Your BP is 100/50. Your pulse is 52. Meg, how much have you been eating?

Patient: I have been eating fine. I talked to the nutritionist about this earlier.

Clinician: Let's have you stand up and face me and I will back you onto the scale. Eyes on me please. Thank you, and now for a height. Ok looks like 5'5". Go ahead and have a seat.

Patient: How much? Can I please see what the scale says? I'm fat.

Clinician: Please sit up and I will listen to your heart and lungs.

Patient: Fine.

Clinician: You know that is not a positive part of your treatment. We have discussed this before. You are here today because your nutritionist and your mom are both very concerned about you. You are very ill, Meg.

Patient: I am fine. I am eating. I tell my mom what I eat. Just because I am not fat like she is.

Clinician: How have you felt? Are you feeling sad or depressed? Last time we talked, you were quite unhappy. Are you feeling suicidal?

Patient: I am busy and not concerned about happiness. I need to train and I am tired of everyone nagging me about what I eat and how much I exercise. I am the smartest in my class and one of the best runners on the team. I don't have time for all of this. I am only here to keep my mom off my back.

Clinician: When you were here last, you agreed to keep working with your nutritionist and being open with your mom and me. We need to work together to help you get well. There are irregular rhythms in your heart. And I noticed the downy hair on your back and arms.

Patient: I will work with all of you. I know. I don't want to be like this. I need to run though. It helps me.

Clinician: When you were in treatment, we were working toward some goals with your cognitive behavioral therapy. We need to revisit these goals. First thought, we need to bring your mom in and have a discussion about the next step to help you get well. I would like to do more tests to check your electrolytes and your blood count.

Patient: Ok. I do want to be better.

Clinician: I want you to feel better and be healthy. Let's have your mom come in, ok?

Patient: Yeah.

--

Please summarize this encounter in the form of the doctor's medical note.

GPT-4: *Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (BMI 14.8) and has signs of malnutrition and cardiac complications. She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered some blood tests and suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.*

B **Use of GPT-4 to Review and Validate a Proposed Medical Note**

Clinician: Can you check the proposed note and identify any facts in it that don't appear explicitly in the transcript?

<The entire transcript and medical note in Panel A is inserted at this point>

GPT-4: *There are a few facts in the proposed note that are not explicitly stated in the transcript, such as:*

- *The patient's BMI, which is calculated from her weight and height, but not revealed to her or mentioned by the clinician. The note should indicate that the BMI is based on the clinician's measurement or estimate.*
- *The signs of malnutrition and cardiac complications, which are not specified or explained by the clinician. The note should list the signs that the clinician observed or inferred, such as the downy hair, the irregular rhythms, the low blood pressure, or the low pulse.*
- *The blood tests that the clinician ordered, which are not specified or explained by the clinician. The note should mention what tests were ordered and why, such as electrolytes, blood count, or liver function.*

Clinician: Thanks. The patient's weight was 87 lb. For the blood test, I ordered a CBC with RBC and peripheral smear. Can you revise the note accordingly?

GPT-4: *Sure. Here is a possible revision:*

Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (we measured her weight at 87 lb and her height at 5'5", resulting in a BMI of 14.8) and has signs of malnutrition (lanugo on her back and arms) and cardiac complications (irregular heart rhythms, low blood pressure, and low pulse). She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered a CBC with RBC and peripheral smear to assess her anemia and possible bone marrow suppression. I suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.

Figure 2 (facing page). Using GPT-4 to Assist in Medical Note Taking.

Panel A shows a prompt that contains a transcript of a physician–patient encounter and a corresponding medical note that was written by GPT-4. As shown in Panel B, GPT-4 is used to check the correctness of the medical note. In this case, the medical note was generated by GPT-4, but such verification could also be applied to a note written by a human being or another AI system. GPT-4 identified several errors and corrected the note.

diagnosis, and education. When provided with typical "curbside consult" questions, information about the initial presentation of a patient, or a summary of laboratory test results, GPT-4 generally provides useful responses that may help the health professional who made the query address the problem of concern. Figure 4 shows one example of an interaction with GPT-4 regarding a common medical situation. Because of the interactive nature of the system, the user can request more detail regarding the response by asking follow-up questions or asking for more concise responses in order to get "to the point" more rapidly.

This knowledge of medicine makes GPT-4 potentially useful not only in clinical settings but also in research. GPT-4 can read medical research material and engage in informed discussion about it, such as briefly summarizing the content, providing technical analysis, identifying relevant prior work, assessing the conclusions, and asking possible follow-up research questions.

FINAL WORDS

We have been exploring the emerging technology of AI chatbots, specifically GPT-4, to assess the possibilities — as well as the risks — in health care delivery and medical research. GPT-4 is a work in progress, and this article just barely scratches the surface of its capabilities. It can, for example, write computer programs for processing and visualizing data, translate foreign

Prompt: A 12-year-old girl is brought to the emergency department by her mother because of a 1-week history of worsening swelling in her legs. The patient also noticed blood in her urine yesterday. The bleeding has not recurred. She had an upper respiratory tract infection and sore throat 1 week ago that caused her to miss several days of school. Medical history is otherwise unremarkable and she takes no routine medications. Menarche has not yet occurred. BMI is 20 kg/m². Vital signs are temperature 37.0°C (98.6°F), pulse 78/min, respirations 12/min, and blood pressure 136/84 mm Hg. Pulse oximetry while breathing room air shows an oxygen saturation of 100%. Physical examination discloses erythema of the posterior pharynx, mild cervical lymphadenopathy, and 3+ pitting edema to both knees.

Results of urinalysis are shown:

Protein: 150 mg/dl
Blood: Positive
Leukocyte esterase: Positive
Nitrite: Negative
WBCs: 5–10/hpf
RBCs: 10–25/hpf
Casts: 1–2/lpf

Results of which of the following laboratory studies are most likely to be abnormal in this patient?

(A) Bleeding time
(B) Erythrocyte count
(C) Serum concentration of C3
(D) Serum IgA concentration
(E) Serum rheumatoid factor assay

GPT-4: *The correct answer is (C) Serum concentration of C3.*
This patient has signs and symptoms of acute poststreptococcal glomerulonephritis (APSGN), a type of glomerular disease that typically occurs 1 to 3 weeks after a streptococcal infection of the skin or pharynx. APSGN is mediated by immune complexes that deposit in the glomerular basement membrane and activate the complement pathway, leading to inflammation, hematuria, proteinuria, and reduced glomerular filtration rate. The resulting fluid retention causes hypertension and edema. The serum C3 level is usually low in APSGN, reflecting complement consumption. The other laboratory studies listed are normal or not relevant in APSGN. Bleeding time is a measure of platelet function, which is not affected by APSGN. Erythrocyte count may be slightly elevated due to hemoconcentration, but not significantly. Serum IgA concentration is elevated in IgA nephropathy, another type of glomerular disease that can cause hematuria and proteinuria, but it is not associated with streptococcal infection or low C3 level. Serum rheumatoid factor assay is used to diagnose rheumatoid arthritis, an autoimmune disorder that does not cause APSGN.

Figure 3. An Example of GPT-4 Answering a Problem from the USMLE.

This prompt is a sample problem from the U.S. Medical Licensing Examination (USMLE). GPT-4 answers almost all written questions in the USMLE correctly and furthermore is able to provide the reasoning behinds its answers.

Clinician: **I have a patient with a history of chronic obstructive pulmonary disease (COPD) who is experiencing shortness of breath. What should I be looking for to determine whether this patient is having an exacerbation?**

GPT-4: *As a health care provider, when evaluating a patient with a history of COPD who presents with shortness of breath, you should be looking for signs and symptoms of an exacerbation. An exacerbation of COPD is defined as an acute worsening of respiratory symptoms beyond the patient's usual day-to-day variation and is typically characterized by increased cough, increased sputum production, and increased dyspnea.*

Some of the key clinical features that suggest an exacerbation of COPD include:

1. *Increased shortness of breath, cough, and sputum production.*
2. *Increased wheezing, chest tightness, and chest discomfort.*
3. *Changes in the color, amount, and thickness of the sputum.*
4. *Increased use of rescue medication, such as short-acting bronchodilators.*
5. *Fatigue, confusion, or altered mental status.*
6. *Cyanosis, a bluish discoloration of the skin or lips, due to inadequate oxygenation.*

It is important to note that other medical conditions can mimic an exacerbation of COPD, such as congestive heart failure or pulmonary embolism. Therefore, it is essential to perform a thorough evaluation, including a detailed history, physical examination, and diagnostic testing, to confirm the diagnosis and identify any potential underlying causes.

Prompt recognition and appropriate management of COPD exacerbations are critical to minimizing their impact on the patient's quality of life and overall prognosis. If you suspect that your patient is having an exacerbation, you should consult with a pulmonologist or respiratory therapist to develop an appropriate treatment plan, which may include oxygen therapy, inhaled bronchodilators and corticosteroids, and antibiotics for suspected bacterial infections.

Clinician: **There is no sputum production. But the patient reports fatigue and there is evidence of cyanosis.**

GPT-4: *If your patient with a history of COPD is experiencing shortness of breath and fatigue with evidence of cyanosis, this could indicate a severe exacerbation of COPD, which requires urgent medical attention. In this case, it is important to assess the patient's oxygen saturation levels, which can be done with a pulse oximeter, and to provide supplemental oxygen therapy as necessary to maintain a target oxygen saturation level of at least 90%.*

The absence of sputum production does not rule out the possibility of an exacerbation of COPD, as not all patients with COPD produce sputum during an exacerbation. However, it is still important to evaluate the patient for any other signs of worsening respiratory distress, such as increased respiratory rate, accessory muscle use, and worsening oxygenation.

In addition to providing supplemental oxygen, the patient may benefit from inhaled bronchodilators and corticosteroids to help relieve symptoms and reduce inflammation in the airways. Antibiotics may also be necessary if there is evidence of a bacterial infection. Close monitoring of the patient's vital signs, oxygen saturation levels, and overall clinical status is important to guide further management and ensure a timely response to any changes in the patient's condition.

It is important to remember that exacerbations of COPD can be life-threatening, and prompt recognition and treatment are essential to improving outcomes for the patient. Therefore, it is advisable to seek the input of a pulmonologist or respiratory therapist for further management and monitoring.

Figure 4. **An Example "Curbside Consult" Interaction with GPT-4.**

languages, decipher explanation-of-benefits notices and laboratory tests for readers unfamiliar with the language used in each, and, perhaps controversially, write emotionally supportive notes to patients.

Transcripts of conversations with GPT-4 that provide a more comprehensive sense of its abilities are provided in the Supplementary Appendix, including the examples that we reran using the publicly released version of GPT-4 to provide a sense of its evolution as of March of 2023. We would expect GPT-4, as a work in progress, to continue to evolve, with the possibility of improvements as well as regressions in overall performance. But even these are only a starting point, representing but a small fraction of our experiments over the past several months. Our hope is to contribute to what we believe will be an important public discussion about the role of this new type of AI, as well as to understand how our approach to health care and medicine can best evolve alongside its rapid evolution.

Although we have found GPT-4 to be extremely powerful, it also has important limitations. Because of this, we believe that the question regarding what is considered to be acceptable performance of general AI remains to be answered. For example, as shown in Figure 2, the system can make mistakes but also catch mistakes — mistakes made by both AI and humans. Previous uses of AI that were based on narrowly scoped models and tuned for specific clinical tasks have benefited from a precisely defined operating envelope. But how should one evaluate the general intelligence of a tool such as GPT-4? To what extent can the user "trust" GPT-4 or does the reader need to spend time verifying the veracity of what it writes? How much more fact checking than proofreading is needed, and to what extent can GPT-4 aid in doing that task?

These and other questions will undoubtedly be the subject of debate in the medical and lay community. Although we admit our bias as employees of the entities that created GPT-4, we predict that chatbots will be used by medical professionals, as well as by patients, with increasing frequency. Perhaps the most important point is that GPT-4 is not an end in and of itself. It is the opening of a door to new possibilities as well as new risks. We speculate that GPT-4 will soon be followed by even more powerful and capable AI systems — a series of increasingly powerful and generally intelligent machines. These machines are tools, and like all tools, they can be used for good but have the potential to cause harm. If used carefully and with an appropriate degree of caution, these evolving tools have the potential to help health care providers give the best care possible.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Katie Mayer of OpenAI for her contributions to this study, Sam Altman of OpenAI for his encouragement and support of the early access to GPT-4 during its development, and the entire staff at OpenAI and Microsoft Research for their continued support in the study of the effects of their work on health care and medicine.

From Microsoft Research, Redmond, WA (P.L., S.B.); and Nuance Communications, Burlington, MA (J.P.).

1. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-89.
2. Han K, Cao P, Wang Y, et al. A review of approaches for predicting drug-drug interactions based on machine learning. Front Pharmacol 2022;12:814858.
3. Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 2021;4:62.
4. Milosevic N, Thielemann W. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. Journal of Web Semantics, August 7, 2022 (https://arxiv.org/abs/2201.01647).
5. OpenAI. Introducing ChatGPT. November 30, 2022 (https://openai.com/blog/chatgpt).
6. Corbelle JG, Bugarín-Diz A, Alonso-Moral J, Taboada J. Dealing with hallucination and omission in neural Natural Language Generation: a use case on meteorology. In: Proceedings and Abstracts of the 15th International Conference on Natural Language Generation, July 18–22, 2022. Waterville, ME: Arria, 2022.
7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv, December 26, 2022 (https://arxiv.org/abs/2212.13138).
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198.
9. Nuance. Automatically document care with the Dragon Ambient eXperience (https://www.nuance.com/healthcare/ambient-clinical-intelligence.html).
10. Kazi N, Kuntz M, Kanewala U, Kahanda I, Bristow C, Arzubi E. Dataset for automated medical transcription. Zenodo, November 18, 2020 (https://zenodo.org/record/4279041#.Y_uCZh_MI2w).
11. Cancarevic I. The US medical licensing examination. In: International medical graduates in the United States. New York: Springer, 2021.

DOI: 10.1056/NEJMsr2214184

The NEW ENGLAND JOURNAL of MEDICINE

Machine Learning in Medicine

Alvin Rajkomar, M.D., Jeffrey Dean, Ph.D., and Isaac Kohane, M.D., Ph.D.

A 49-year-old patient notices a painless rash on his shoulder but does not seek care. Months later, his wife asks him to see a doctor, who diagnoses a seborrheic keratosis. Later, when the patient undergoes a screening colonoscopy, a nurse notices a dark macule on his shoulder and advises him to have it evaluated. One month later, the patient sees a dermatologist, who obtains a biopsy specimen of the lesion. The findings reveal a noncancerous pigmented lesion. Still concerned, the dermatologist requests a second reading of the biopsy specimen, and invasive melanoma is diagnosed. An oncologist initiates treatment with systemic chemotherapy. A physician friend asks the patient why he is not receiving immunotherapy.

From Google, Mountain View, CA (A.R., J.D.); and the Department of Biomedical Informatics, Harvard Medical School, Boston (I.K.). Address reprint requests to Dr. Kohane at the Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St., Boston, MA, 02115, or at isaac_kohane@harvard.edu.

N Engl J Med 2019;380:1347-58.
DOI: 10.1056/NEJMra1814259

What if every medical decision, whether made by an intensivist or a community health worker, was instantly reviewed by a team of relevant experts who provided guidance if the decision seemed amiss? Patients with newly diagnosed, uncomplicated hypertension would receive the medications that are known to be most effective rather than the one that is most familiar to the prescriber.[1,2] Inadvertent overdoses and errors in prescribing would be largely eliminated.[3,4] Patients with mysterious and rare ailments could be directed to renowned experts in fields related to the suspected diagnosis.[5]

Such a system seems far-fetched. There are not enough medical experts to staff it, it would take too long for experts to read through a patient's history, and concerns related to privacy laws would stop efforts before they started.[6] Yet, this is the promise of machine learning in medicine: the wisdom contained in the decisions made by nearly all clinicians and the outcomes of billions of patients should inform the care of each patient. That is, every diagnosis, management decision, and therapy should be personalized on the basis of all known information about a patient, in real time, incorporating lessons from a collective experience.

This framing emphasizes that machine learning is not just a new tool, such as a new drug or medical device. Rather, it is the fundamental technology required to meaningfully process data that exceed the capacity of the human brain to comprehend; increasingly, this overwhelming store of information pertains to both vast clinical databases and even the data generated regarding a single patient.[7]

Nearly 50 years ago, a Special Article in the *Journal* stated that computing would be "augmenting and, in some cases, largely replacing the intellectual functions of the physician."[8] Yet, in early 2019, surprisingly little in health care is driven by machine learning. Rather than report the myriad proof-of-concept models (of retrospective data) that have been tested, here we describe the core structural changes and paradigm shifts in the health care system that are necessary to enable the full promise of machine learning in medicine (see video).

A video overview of machine learning is available at NEJM.org

MACHINE LEARNING EXPLAINED

Traditionally, software engineers have distilled knowledge in the form of explicit computer code that instructs computers exactly how to process data and how to

View Article at NEJM.org

Back to Table of Contents

make decisions. For example, if a patient has elevated blood pressure and is not receiving an antihypertensive medication, then a properly programmed computer can suggest treatment. These types of rules-based systems are logical and interpretable, but, as a Sounding Board article in the *Journal* in 1987 noted, the field of medicine is "so broad and complex that it is difficult, if not impossible, to capture the relevant information in rules."[9]

The key distinction between traditional approaches and machine learning is that in machine learning, a model learns from examples rather than being programmed with rules. For a given task, examples are provided in the form of inputs (called features) and outputs (called labels). For instance, digitized slides read by pathologists are converted to features (pixels of the slides) and labels (e.g., information indicating that a slide contains evidence of changes indicating cancer). Using algorithms for learning from observations, computers then determine how to perform the mapping from features to labels in order to create a model that will generalize the information such that a task can be performed correctly with new, never-seen-before inputs (e.g., pathology slides that have not yet been read by a human). This process, called supervised machine learning, is summarized in Figure 1. There are other forms of machine learning.[10] Table 1 lists examples of cases of the clinical usefulness of input-to-output mappings that are based on peer-reviewed research or simple extensions of existing machine-learning capabilities.

In applications in which predictive accuracy is critically important, the ability of a model to find statistical patterns across millions of features and examples is what enables superhuman performance. However, these patterns do not necessarily correspond to the identification of underlying biologic pathways or modifiable risk factors that underpins the development of new therapies.

There is no bright line between machine-learning models and traditional statistical models, and a recent article summarizes the relationship between the two.[36] However, sophisticated new machine-learning models (e.g., those used in "deep learning" [a class of machine-learning algorithms that use artificial neural networks that can learn extremely complex relationships between features and labels and have been shown to exceed human abilities in performing tasks such as classification of images][37,38]) are well suited to learn from the complex and heterogeneous kinds of data that are generated from modern clinical care, such as medical notes entered by physicians, medical images, continuous monitoring data from sensors, and genomic data to help make medically relevant predictions. Guidance on when to use simple or sophisticated machine-learning models is provided in Table 2.

A key difference between human learning and machine learning is that humans can learn to make general and complex associations from small amounts of data. For example, a toddler does not need to see many examples of a cat to recognize a cheetah as a cat. Machines, in general, require many more examples than humans to learn the same task, and machines are not endowed with common sense. The flipside, however, is that machines can learn from massive amounts of data.[39] It is perfectly feasible for a machine-learning model to be trained with the use of tens of millions of patient charts stored in electronic health records (EHRs), with hundreds of billions of data points, without any lapses of attention, whereas it is very difficult for a human physician to see more than a few tens of thousands of patients in an entire career.

HOW MACHINE LEARNING CAN AUGMENT THE WORK OF CLINICIANS

PROGNOSIS

A machine-learning model can learn the patterns of health trajectories of vast numbers of patients. This facility can help physicians to anticipate future events at an expert level, drawing from information well beyond the individual physician's practice experience. For example, how likely is it that a patient will be able to return to work, or how quickly will the disease progress? At a population level, the same type of forecasting can enable reliable identification of patients who will soon have high-risk conditions or increased utilization of health care services; this information can be used to provide additional resources to proactively support them.[40]

Large integrated health systems have already used simple machine-learning models to automatically identify hospitalized patients who are at risk for transfer to the intensive care unit,[17] and retrospective studies suggest that more complex

Back to Table of Contents

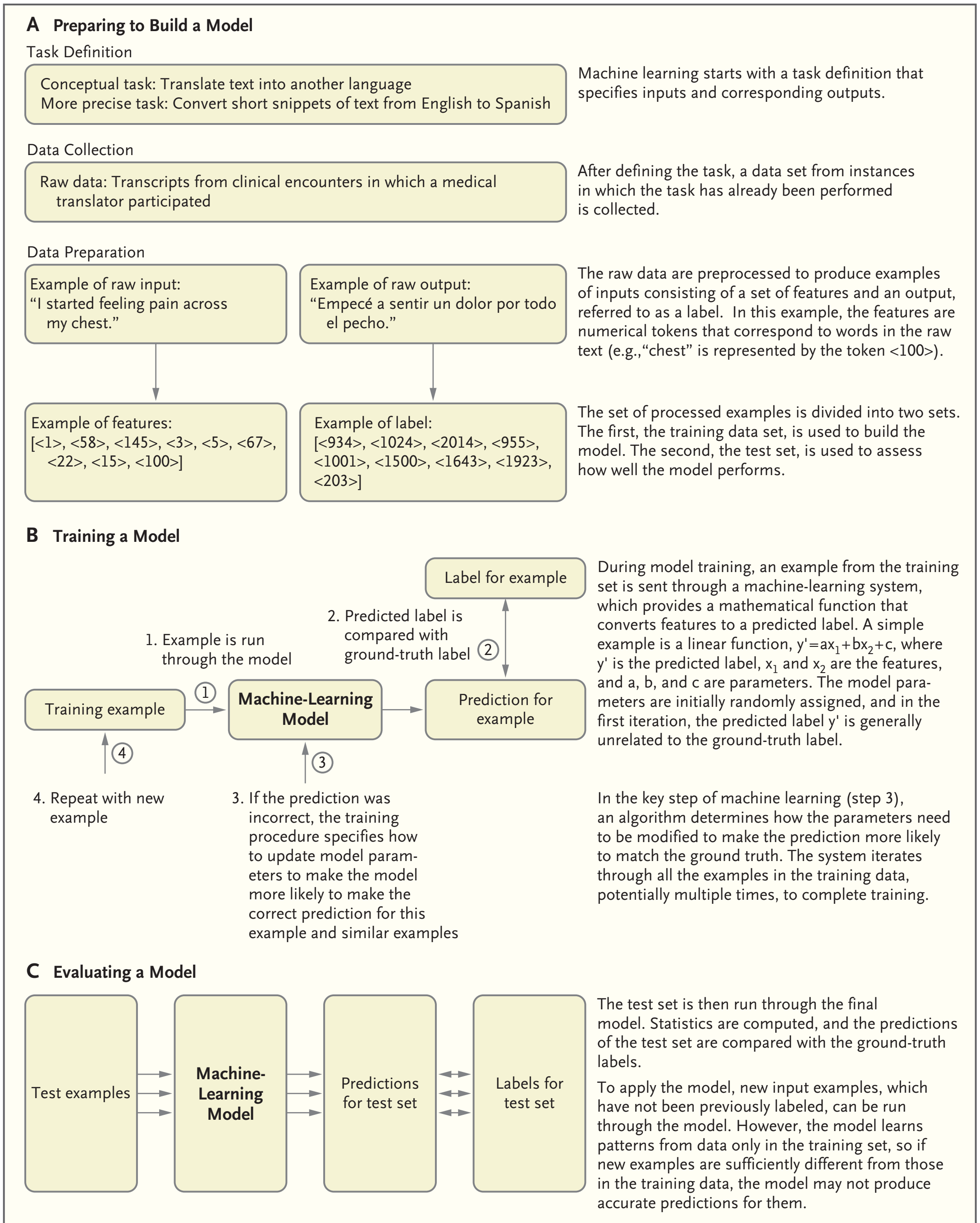


Figure 1. Conceptual Overview of Supervised Machine Learning.

As shown in Panel A, machine learning starts with a task definition that specifies an input that should be mapped to a corresponding output. The task in this example is to take a snippet of text from one language (input) and produce text of the same meaning but in a different language (output). There is no simple set of rules to perform this mapping well; for example, simply translating each word without examining the context does not lead to high-quality translations. As shown in Panel B, there are key steps in training machine-learning models. As shown in Panel C, models are evaluated with data that were not used to build them (i.e., the test set). This evaluation generally precedes formal testing to determine whether the models are effective in live clinical environments involving trial designs, such as randomized clinical trials.

Table 1. Examples of Types of Input and Output Data That Power Machine-Learning Applications.*

Application and Input Data (Feature)	Output Data (Label)	Purpose of Machine Learning	Comments
Commonly encountered machine-learning applications			
Cardiovascular risk factors (e.g., hypertension)	Myocardial infarction	Predicts a patient's cardiovascular risk (e.g., according to Framingham risk score or predicted by pooled cohorts equation[11])	Experts select pertinent risk factors (e.g., hypertension).
Search query (e.g., "What is the pooled cohorts equation?")	Web page that contains the most relevant information	Identifies the best source of information for user (e.g., Internet search engines[12])	Machine learning is a key component in determining the most relevant information to show users. Since there are innumerable types of Web queries, it is impractical to manually curate a list of results (outputs) for each query (input).
Sentence in Spanish (e.g., "¿Dónde está el dolor?")	Sentence in English ("Where is the pain?")	Translates from one language to another (e.g., computer translation[13,14])	
Daily clinical workflow			
Query about a patient's condition (e.g., "Has this patient previously had a venous thromboembolism?")	Information from a patient's chart that answers the question	Identifies key information from a patient's chart	Machine learning could help find information in a patient's chart that physicians want to see.
Audio recording of a conversation between a doctor and a patient	Sections of a clinical note that correspond to the conversation; correct billing codes	Provides automated generation of sections of a clinical note[15] or automatic assignment of billing codes from an encounter[16]	Automated documentation could ameliorate data-entry burdens of physicians.
Real-time EHR data	Clinical deterioration in next 24 hr	Provides real-time detection of patients at risk for clinical deterioration[17]	Real-time detection could serve as a patient-safety mechanism to ensure that attention is paid to patients at high risk.
Genetic sequence of a cancer	Event-free survival	Makes personalized predictions of patients' outcomes[18]	Machine learning, which can help use data that is difficult for humans to interpret, can make meaningful clinical contributions.
Workflow and use of medical images			
Image of the eye fundus	Diabetic retinopathy, myocardial infarction	Automates diagnosis of eye diseases,[19-24] predicts cardiac risk[25]	Automated diagnosis could be used in regions where screening tools exist but there are not enough specialists to review the images; a medical image can be used as a noninvasive "biomarker."
Digitized pathological slide of a sentinel lymph node	Detection of metastatic disease	Reviews pathological slides[26-28]	The efficiency and accuracy of experts in medical imaging can be augmented by machine learning.
CT of the head	Intracranial hemorrhage	Triages urgent findings in head CTs[29]	Abnormal images can be triaged by models to ensure that life-threatening abnormalities are diagnosed promptly.
Real-time video of a colonoscopy	Identification of polyps that warrant or do not warrant biopsy	Improves selection of polyps that require resection,[30] provides real-time feedback if a proceduralist risks injuring a key anatomical structure	Machine learning may enable "optical biopsies" to help proceduralists to identify high-risk lesions and ignore low-risk ones; real-time feedback is similar to lane-assist in cars.

Changes in patient experiences			
Real-time data from smartwatches or other digital sensors	Atrial fibrillation, hospitalization, laboratory test results	Provides outpatient screening of common diseases such as atrial fibrillation,[25,31] remote monitoring of ambulatory patients to detect conditions that warrant hospitalization, and detection of physiological abnormalities without blood draws[32]	Many trials have not shown a benefit for remote monitoring to detect conditions that warrant hospitalization, but with better sensors this may improve; sensors could be used as a noninvasive "biomarker" for traditionally measured laboratory tests.
Robotic movements of a prosthetic arm	Successful use of feeding utensils	Increases functionality of prosthetic equipment	Techniques used in the computer program AlphaGo, such as reinforcement learning, can be used to help robots perform complex tasks such as manipulating physical objects.
Text-messaging chat logs between a consumer and chatbot (i.e., a computer program that converses through sound or text)	Diagnoses	Provides early identification of patients who may need medical attention	Patients may delay seeking care when appropriate or seek care unnecessarily, leading to poor outcomes or waste.
Smartphone image of rash	Diagnosis	Provides automated triage of medical conditions[33]	Unnecessary medical consultation could be avoided, improving care.
Changes in regulations and billing			
Documentation of a medical encounter and "prior authorization" application	Insurance approval	Provides automated and instantaneous approval of insurance payment for medications	Automated approval could be used to cut down on administrative paperwork and reduce fraud.
Real-time EHR data	Future health care utilization	Identifies high-cost patients who may receive resource-intensive care in the future[34]	Proactive identification of patients who are at risk for clinical deterioration or gaps in care creates new opportunities to intervene early.
Real-time analysis of chest x-ray films, ventilation settings, and clinical variables	Diagnosis of ARDS	Provides automated identification of patients with ARDS and assessment of proper treatment	Current quality metrics may be suboptimal, so direct measurement for proper treatment, requiring correct case identification and assessment of treatment, can lead to improved quality measures without further documentation.[35]
Changes in epidemiologic factors			
Swabs of hospital rooms sent for genetic sequencing	Identification of pathogens	Provides real-time identification of possibly multidrug-resistant organisms	Immediate identification of emerging or important pathogens may enable rapid responses.

* Machine-learning models require collection of historical input and output data, which are also called features and labels. For example, a study that determined baseline cardiovascular risk factors and then followed patients for the occurrence of myocardial infarction would provide training examples in which the features were the set of risk factors and the label was a future myocardial infarction. The model would be trained to use the features to predict the label, so for new patients, the model would predict the risk of the occurrence of the label. This general framework can be used for a variety of tasks. ARDS denotes acute respiratory distress syndrome, CT computed tomography, and EHR electronic health record.

Table 2. Key Questions to Ask When Deciding What Type of Model Is Necessary.

How complex is the prediction task?
Simple prediction tasks are defined as those that can be performed with high accuracy with a small number of predictor variables. For example, predicting the development of hyperkalemia might be possible from just a small set of variables, such as renal function, the use of potassium supplements, and receipt of certain medications.
Complex prediction tasks are defined as those that cannot be predicted accurately with a small number of predictor variables. For example, identification of abnormalities in a pathological slide requires evaluation of patterns that are not obvious over millions of pixels.
In general, simple prediction tasks can be performed with traditional models (e.g., logistic regression), and complex tasks require more complex models (e.g., neural networks).
Should the prediction task be performed by clinicians who are entering the data manually, or should it be performed by a computer using raw data?
In addition to classifying a prediction task as simple or complex, consider how the model will be used in practice. If a model will be used in a bedside scoring system (e.g., the Wells score for assessment of the probability of pulmonary embolism), then using a small number of variables curated by humans is preferable. In this case, traditional models may be as effective as more complex ones.
If a model is expected to automatically analyze noisy data without any intervening human curation or normalization, then the task becomes complex, and complex models become generally more useful.
It is possible to write a set of rules to process raw data to a smaller set of "clean" features, which might be amenable to a traditional model if the prediction task is simple. However, it is often very time-consuming to write these rules and to keep them updated.
How many examples exist to train a model?
Simple prediction tasks generally do not require many examples to learn from in order to build a model.
The training of complex models generally requires many more examples. There is no predetermined number of examples, but at least multiple thousands of examples are needed to construct complex models, and the more complex the prediction task, the more data are generally required. Specialized techniques do exist to reduce the number of training examples that are necessary to construct an accurate model (e.g., transfer learning).
How interpretable does a model need to be?
Simple prediction tasks are interpretable because the number of features evaluated by the model is quite small.
Complex tasks are inherently harder to interpret because the model is expected to learn to identify complex statistical patterns, which might correspond to many small signals across many features. Although this complexity allows for more accurate predictions, it has the drawback of making it harder to succinctly present or explain the subtle patterns behind a particular prediction.

and accurate prognostic models can be built with raw data from EHRs[41] and medical imaging.[42]

Building machine-learning systems requires training with data that provide an integrated, longitudinal view of a patient. A model can learn what happens to patients only if the outcomes are included in the data set that the model is based on. However, data are currently siloed in EHR systems, medical imaging picture archiving and communication systems, payers, pharmacy benefits managers, and even apps on patients' phones. A natural solution would be to systematically place data in the hands of patients themselves. We have long advocated for this solution,[43] which is now enabled by the rapid adoption of patient-controlled application programming interfaces.[44]

Convergence of a unified data format such as Fast Healthcare Interoperability Resources (FHIR)[45] would allow for useful aggregation of data. Patients could then control who had access to their data for use in building or running models. Although there are concerns that technical interoperability does not solve the problem of semantic standardization endemic in EHR data,[46] the adoption of HTML (Hypertext Markup Language) has allowed Web data, which are perhaps even messier than EHR data, to be indexed and made useful with search engines.

DIAGNOSIS

Every patient is unique, but the best doctors can determine when a subtle sign that is particular to a patient is within the normal range or indicates a true outlier. Can statistical patterns detected by machine learning be used to help physicians identify conditions that they do not diagnose routinely?

The Institute of Medicine concluded that a diagnostic error will occur in the care of nearly every patient in his or her lifetime,[47] and receiving the right diagnosis is critical to receiving appropriate care.[48] This problem is not limited to rare conditions. Cardiac chest pain, tuberculosis, dysentery, and complications of childbirth are commonly not detected in developing countries, even when there is adequate access to therapies, time to examine patients, and fully trained providers.[49]

With data collected during routine care, machine learning could be used to identify likely diagnoses during a clinical visit and raise awareness of conditions that are likely to manifest later.[50] However, such approaches have limitations. Less skilled clinicians may not elicit the information necessary for a model to assist them meaningfully, and the diagnoses that the models are built from may be provisional or incorrect,[48] may be conditions that do not manifest symptoms (and thus may lead to overdiagnosis),[51] may be influenced by billing,[52] or may simply not be recorded. However, models could suggest questions or tests to physicians[53] on the basis of data collected in real time; these suggestions could be helpful in scenarios in which high-stakes misdiagnoses are common (e.g., childbirth) or when clinicians are uncertain. The discordance between diagnoses that are clinically correct and those recorded in EHRs or reimbursement claims means that clinicians should be involved from the outset in determining how data generated as part of routine care should be used to automate the diagnostic process.

Models have already been successfully trained to retrospectively identify abnormalities across a variety of image types (Table 1). However, only a limited number of prospective trials involve the use of machine-learning models as part of a clinician's regular course of work.[19,20]

TREATMENT

In a large health care system with tens of thousands of physicians treating tens of millions of patients, there is variation in when and why patients present for care and how patients with similar conditions are treated. Can a model sort through these natural variations to help physicians identify when the collective experience points to a preferred treatment pathway?

A straightforward application is to compare what is prescribed at the point of care with what a model predicts would be prescribed, and discrepancies could be flagged for review (e.g., other clinicians tend to order an alternative treatment that reflects new guidelines). However, a model trained on historical data would learn only the prescribing habits of physicians, not necessarily the ideal practices. To learn which medication or therapy should be prescribed to maximize patient benefit requires either carefully curated data or estimates of causal effects, which machine-learning models do not necessarily — and sometimes cannot with a given data set — identify.

Traditional methods used in comparative effectiveness research and pragmatic trials[54] have provided important insights from observational data.[55] However, recent attempts at using machine learning have shown that it is challenging to generate curated data sets with experts, update the models to incorporate newly published evidence, tailor them to regional prescribing practices, and automatically extract relevant variables from EHRs for ease of use.[56]

Machine learning can also be used to automatically select patients who might be eligible for randomized, controlled trials on the basis of clinical documentation[57] or to identify high-risk patients or subpopulations who are likely to benefit from early or new therapies under study. Such efforts can empower health systems to subject every clinical scenario for which there is equipoise to more rigorous study with decreased cost and administrative overhead.[54,58,59]

CLINICIAN WORKFLOW

The introduction of EHRs has improved the availability of data. However, these systems have also frustrated clinicians with a panoply of checkboxes for billing or administrative documentation,[60] clunky user interfaces,[61,62] increased time spent entering data,[63-66] and new opportunities for medical errors.[67]

The same machine-learning techniques that are used in many consumer products can be used to make clinicians more efficient. Machine learning that drives search engines can help expose relevant information in a patient's chart for a clinician without multiple clicks. Data entry of forms and text fields can be improved with the use of machine-learning techniques such as predictive typing, voice dictation, and automatic summarization. Prior authorization could be replaced by models that automatically authorize payment based on information already recorded in the patient's chart.[68] The motivation behind adopting these abilities is not just convenience to physicians. Making the process of viewing and entering the most clinically useful data frictionless is essential to capturing and recording health care data, which in turn will enable machine learning to help give the best possible care to every patient. Most importantly, increased efficiency, ease of documentation, and improved automated clinical workflow would allow clinicians to spend more time with their patients.

Even outside the EHR system, machine-learning techniques can be adapted for real-time analysis of video of the surgical field to help surgeons avoid critical anatomical structures or unexpected variants or even handle more mundane tasks such as accurate counting of surgical sponges. Checklists can prevent surgical error,[69] and unstinting automated monitoring of their implementation provides additional safety.

In their personal lives, clinicians probably use variants of all these forms of technology on their smartphones. Although there are retrospective proof-of-concept studies of application of these techniques to medical contexts,[15] the major barriers to adoption involve not the development of models but technical infrastructure; legal, privacy, and policy frameworks across EHRs; health systems; and technology providers.

EXPANDING THE AVAILABILITY OF CLINICAL EXPERTISE

There is no way for physicians to individually interact with all the patients who may need care. Can machine learning extend the reach of clinicians to provide expert-level medical assessment without personal involvement? For example, patients with new rashes may be able to obtain a diagnosis by sending a picture that they take on their smartphones,[32,33] thereby averting unnecessary urgent-care visits. A patient considering a visit to the emergency department might be able to converse with an automated triage system and, when appropriate, be directed to another form of care. When a patient does need professional assistance, models could identify physicians with the most relevant expertise and availability. Similarly, to increase comfort and lower cost, patients who otherwise may need to be hospitalized could stay at home if machines can remotely monitor their sensor data.

The delivery of insights from machine learning directly to patients has become increasingly important in the areas of the world where access to direct medical expertise is in limited supply[70] and sophistication. Even in areas where the supply of expert clinicians is abundant, these clinicians are concerned about their ability and the effort required to provide timely and accurate interpretation of the tsunami of patient-driven digital data from sensor or activity-tracking devices worn by patients.[71] Indeed, one of the hopes with regard to machine-learning models trained with data from millions of patient encounters is that they can equip health care professionals with the ability to make better decisions. For instance, nurses might be able to take on many tasks that are traditionally performed by doctors, primary care doctors might be able to perform some of the roles traditionally performed by medical specialists, and medical specialists could devote more of their time to patients who would benefit from their particular expertise.

A variety of mobile apps or Web services that do not involve machine learning have been shown to improve medication adherence[72] and control of chronic diseases.[73,74] However, machine learning in direct-to-patient applications is hindered by formal retrospective and prospective evaluation methods.[75]

KEY CHALLENGES

AVAILABILITY OF HIGH-QUALITY DATA

A central challenge in building a machine-learning model is assembling a representative, diverse data set. It is ideal to train a model with data that most closely resemble the exact format and quality of data expected during use. For instance, for a model that is intended to be used at the point of care, it is preferable to use the same data that are available in the EHR at that particular moment, even if they are known to be unreliable[46] or subject to unwanted variability.[46,76] When they have large enough data sets, modern models can be successfully trained to map noisy inputs to noisy outputs. The use of a smaller set of curated data, such as those collected in clinical trials from manual chart review, is suboptimal unless clinicians at the bedside are expected to abstract the variables by hand according to the original trial specifications. This practice might be feasible with some variables, but not with the hundreds of thousands that are available in the EHR and that are necessary to make the most accurate predictions.[41]

How do we reconcile the use of noisy data sets to train a model with the data maxim "garbage in, garbage out"? Although to learn the majority of complex statistical patterns it is generally better to have large — even noisy — data sets, to fine-tune or evaluate a model, it is necessary to have a smaller set of examples with curated labels. This allows for proper assessment of the predictions of a model against the intended labels when there is a chance that the original ones were mislabeled.[21] For imaging models, this generally requires generating a "ground truth" (i.e., diagnoses or findings that would be assigned to an example by an infallible expert) label adjudicated by multiple graders for each image, but for nonimaging tasks, obtaining ground truth may be impossible after the fact if, for example, a necessary diagnostic test was not obtained.

Machine-learning models generally perform best when they have access to large amounts of training data. Thus, a key issue for many uses of machine learning will be balancing privacy and regulatory requirements with the desire to leverage large and diverse data sets to improve the accuracy of machine-learning models.

LEARNING FROM UNDESIRABLE PAST PRACTICES

All human activity is marred by unwanted and unconscious bias. Builders and users of machine-learning systems need to carefully consider how biases affect the data being used to train a model[77] and adopt practices to address and monitor them.[78]

The strength of machine learning, but also one of its vulnerabilities, is the ability of models to discern patterns in historical data that humans cannot find. Historical data from medical practice indicate health care disparities in the provision of systematically worse care for vulnerable groups than for others.[77,79] In the United States, the historical data reflect a payment system that rewards the use of potentially unnecessary care and services and may be missing data about patients who should have received care but did not (e.g., uninsured patients).

EXPERTISE IN REGULATION, OVERSIGHT, AND SAFE USE

Health systems have developed sophisticated mechanisms to ensure the safe delivery of pharmaceutical agents to patients. The wide applicability of machine learning will require a similarly sophisticated structure of regulatory oversight,[80] legal frameworks,[81] and local practices[82] to ensure the safe development, use, and monitoring of systems. Moreover, technology companies will have to provide scalable computing platforms to handle large amounts of data and use of models; their role today, however, is unclear.

Critically, clinicians and patients who use machine-learning systems need to understand their limitations, including instances in which a model is not designed to generalize to a particular scenario.[83-85] Overreliance on machine-learning models in making decisions or analyzing images may lead to automation bias,[86] and physicians may have decreased vigilance for errors. This is especially problematic if models themselves are not interpretable enough for clinicians to identify situations in which a model is giving incorrect advice.[87,88] Presenting the confidence interval in a prediction of a model may help, but confidence intervals themselves may be interpreted incorrectly.[89,90] Thus, there is a need for prospective, real-world clinical evaluation of models in use rather than only retrospective assessment of performance based on historical data sets.

Special consideration is needed for machine-learning applications targeted directly to patients. Patients may not have ways to verify that the claims made by a model maker have been substantiated by high-quality clinical evidence or that a suggested action is reasonable.

PUBLICATIONS AND DISSEMINATION OF RESEARCH

The interdisciplinary teams that build models may report results in venues that may be unfamiliar to clinicians. Manuscripts are often posted online at preprint services such as arXiv and bioRxiv,[91,92] and the source code of many models exists in repositories such as GitHub. Moreover, many peer-reviewed computer science manuscripts are not published by traditional journals

An audio interview with Dr. Kohane is available at NEJM.org

but as proceedings in conferences such as the Conference on Neural Information Processing Systems (NeurIPS) and the International Conference on Machine Learning (ICML).

CONCLUSIONS

The accelerating creation of vast amounts of health care data will fundamentally change the nature of medical care. We firmly believe that the patient–doctor relationship will be the cornerstone of the delivery of care to many patients and that the relationship will be enriched by additional insights from machine learning. We expect a handful of early models and peer-reviewed publications of their results to appear in the next few years, which — along with the development of regulatory frameworks and economic incentives for value-based care — are reasons to be cautiously optimistic about machine learning in health care. We look forward to the hopefully not-too-distant future when all medically relevant data used by millions of clinicians to make decisions in caring for billions of patients are analyzed by machine-learning models to assist with the delivery of the best possible care to all patients.

A 49-year-old patient takes a picture of a rash on his shoulder with a smartphone app that recommends an immediate appointment with a dermatologist. His insurance company automatically approves the direct referral, and the app schedules an appointment with an experienced nearby dermatologist in 2 days. This appointment is automatically cross-checked with the patient's personal calendar. The dermatologist performs a biopsy of the lesion, and a pathologist reviews the computer-assisted diagnosis of stage I melanoma, which is then excised by the dermatologist.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Bakris G, Sorrentino M. Redefining hypertension — assessing the new blood-pressure guidelines. N Engl J Med 2018;378:497-9.
2. Institute of Medicine. Crossing the quality chasm: a new health system for the twenty-first century. Washington, DC: National Academies Press, 2001.
3. Lasic M. Case study: an insulin overdose. Institute for Healthcare Improvement (http://www.ihi.org/education/IHIOpenSchool/resources/Pages/Activities/AnInsulinOverdose.aspx).
4. Institute of Medicine. To err is human: building a safer health system. Washington, DC: National Academies Press, 2000.
5. National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: National Academies Press, 2016.
6. Berwick DM, Gaines ME. How HIPAA harms care, and how to stop it. JAMA 2018;320:229-30.
7. Obermeyer Z, Lee TH. Lost in thought — the limits of the human mind and the future of medicine. N Engl J Med 2017;377:1209-11.
8. Schwartz WB. Medicine and the computer — the promise and problems of change. N Engl J Med 1970;283:1257-64.
9. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine — where do we stand? N Engl J Med 1987;316:685-8.
10. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge, MA: MIT Press, 2016.
11. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA 2014;311:1406-15.
12. Clark J. Google turning its lucrative Web search over to AI machines. Bloomberg News. October 26, 2015 (https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines).
13. Johnson M, Schuster M, Le QV, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv. November 14, 2016 (http://arxiv.org/abs/1611.04558).
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. September 1, 2014 (http://arxiv.org/abs/1409.0473).
15. Kannan A, Chen K, Jaunzeikare D, Rajkomar A. Semi-supervised learning for information extraction from dialogue. In: Interspeech 2018. Baixas, France: International Speech Communication Association, 2018:2077-81.
16. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. arXiv. January 24, 2018 (http://arxiv.org/abs/1801.07860).
17. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11:Suppl 1:S18-S24.
18. Grinfeld J, Nangalia J, Baxter EJ, et al. Classification and personalized prognosis in myeloproliferative neoplasms. N Engl J Med 2018;379:1416-30.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56.
20. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019 February 27 (Epub ahead of print).
21. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125:1264-72.
22. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
23. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23.
24. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122-1131.e9.
25. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158-64.
26. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42:1636-46.
27. Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence-based breast cancer

Back to Table of Contents

nodal metastasis detection. Arch Pathol Lab Med 2018 October 8 (Epub ahead of print).

28. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017; 318:2199-210.

29. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018; 392:2388-96.

30. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018;169:357-66.

31. Tison GH, Sanchez JM, Ballinger B, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol 2018;3:409-16.

32. Galloway CD, Valys AV, Petterson FL, et al. Non-invasive detection of hyperkalemia with a smartphone electrocardiogram and artificial intelligence. J Am Coll Cardiol 2018;71:Suppl:A272. abstract.

33. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.

34. Rajkomar A, Yim JWL, Grumbach K, Parekh A. Weighting primary care patient panel size: a novel electronic health record-derived measure using machine learning. JMIR Med Inform 2016;4(4):e29.

35. Schuster MA, Onorato SE, Meltzer DO. Measuring the cost of quality measurement: a missing link in quality strategy. JAMA 2017;318:1219-20.

36. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317-8.

37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.

38. Hinton G. Deep learning — a technology with the potential to transform health care. JAMA 2018;320:1101-2.

39. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009;24:8-12.

40. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33:1123-31.

41. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1(1):18.

42. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.

43. Mandl KD, Szolovits P, Kohane IS. Public standards and patients' control: how to keep electronic medical records accessible but private. BMJ 2001;322:283-7.

44. Mandl KD, Kohane IS. Time for a patient-driven health information economy? N Engl J Med 2016;374:205-8.

45. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23:899-908.

46. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51: Suppl 3:S30-S37.

47. McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. JAMA 2015;314:2501-2.

48. Institute of Medicine, National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: National Academies Press, 2016.

49. Das J, Woskie L, Rajbhandari R, Abbasi K, Jha A. Rethinking assumptions about delivery of healthcare: implications for universal health coverage. BMJ 2018; 361:k1716.

50. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ 2009;339:b3677.

51. Kale MS, Korenstein D. Overdiagnosis in primary care: framing the problem and finding solutions. BMJ 2018;362:k2820.

52. Lindenauer PK, Lagu T, Shieh M-S, Pekow PS, Rothberg MB. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. JAMA 2012;307: 1405-13.

53. Slack WV, Hicks GP, Reed CE, Van Cura LJ. A computer-based medical-history system. N Engl J Med 1966;274:194-8.

54. Ford I, Norrie J. Pragmatic trials. N Engl J Med 2016;375:454-63.

55. Frieden TR. Evidence for health decision making — beyond randomized, controlled trials. N Engl J Med 2017;377:465-75.

56. Ross C, Swetlitz I, Thielking M, et al. IBM pitched Watson as a revolution in cancer care: it's nowhere close. Boston: STAT, September 5, 2017 (https://www.statnews.com/2017/09/05/watson-ibm-cancer/).

57. Fiore LD, Lavori PW. Integrating randomized comparative effectiveness research with patient care. N Engl J Med 2016;374:2152-8.

58. Schneeweiss S. Learning from big health care data. N Engl J Med 2014;370: 2161-3.

59. Institute of Medicine. The learning healthcare system: workshop summary. Washington, DC: National Academies Press, 2007.

60. Erickson SM, Rockwern B, Koltov M, McLean RM. Putting patients first by reducing administrative tasks in health care: a position paper of the American College of Physicians. Ann Intern Med 2017;166: 659-61.

61. Hill RG Jr, Sears LM, Melanson SW. 4000 Clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med 2013; 31:1591-4.

62. Sittig DF, Murphy DR, Smith MW, Russo E, Wright A, Singh H. Graphical display of diagnostic test results in electronic health records: a comparison of 8 systems. J Am Med Inform Assoc 2015; 22:900-4.

63. Mamykina L, Vawdrey DK, Hripcsak G. How do residents spend their shift time? A time and motion study with a particular focus on the use of computers. Acad Med 2016;91:827-32.

64. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med 2010;170:377-80.

65. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017;15:419-26.

66. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016;165: 753-60.

67. Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. JAMA 2018;319:1276-8.

68. Lee VS, Blanchfield BB. Disentangling health care billing: for patients' physical and financial health. JAMA 2018;319:661-3.

69. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491-9.

70. Steinhubl SR, Kim K-I, Ajayi T, Topol EJ. Virtual care for improved global health. Lancet 2018;391:419.

71. Gabriels K, Moerenhout T. Exploring entertainment medicine and professionalization of self-care: interview study among doctors on the potential effects of digital self-tracking. J Med Internet Res 2018;20(1):e10.

72. Morawski K, Ghazinouri R, Krumme A, et al. Association of a smartphone application with medication adherence and blood pressure control: the MedISAFE-BP randomized clinical trial. JAMA Intern Med 2018;178:802-9.

73. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet 2017;390:959-68.

74. Denis F, Basch E, Septans AL, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019;321(3):306-7.

75. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392:2263-4.
76. Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017;357:j2813.
77. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544-7.
78. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866-72.
79. Institute of Medicine. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: National Academies Press, 2003.
80. Shuren J, Califf RM. Need for a national evaluation system for health technology. JAMA 2016;316:1153-4.
81. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Aff (Millwood) 2011;30:2310-7.
82. Auerbach AD, Neinstein A, Khanna R. Balancing innovation and safety when integrating digital tools into health care. Ann Intern Med 2018;168:733-4.
83. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood) 2014;33:1148-54.
84. Sniderman AD, D'Agostino RB Sr, Pencina MJ. The role of physicians in the era of predictive analytics. JAMA 2015; 314:25-6.
85. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33:1163-70.
86. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24: 423-31.
87. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517-8.
88. Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20-3.
89. Jiang H, Kim B, Guan M, Gupta M. To trust or not to trust a classifier. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. Advances in neural information processing systems 31. New York: Curran Associates, 2018: 5541-52.
90. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014;33:1139-47.
91. arXiv.org Home page (https://arxiv.org/).
92. bioRxiv. bioRxiv: The preprint server for biology (https://www.biorxiv.org/).

NEJM Catalyst | Innovations in Care Delivery

CASE STUDY

Using AI to Empower Collaborative Team Workflows: Two Implementations for Advance Care Planning and Care Escalation

Ron C. Li, MD, Margaret Smith, MBA, Jonathan Lu, MS, Anand Avati, MS, Samantha Wang, MD, MHS, Winifred G. Teuteberg, MD, Kenny Shum, PhD, Grace Hong, Briththa Seevaratnam, MS, Jerri Westphal, MSN, RN, CNML, Margaret Dougherty, MHA, MS, OTR/L, Pooja Rao, MS, OTR/L, CSRS, Steven Asch, MD, MPH, Steven Lin, MD, Christopher Sharp, MD, Lisa Shieh, MD, PhD, Nigam H. Shah, MBBS, PhD

Vol. 3 No. 4 | April 2022
DOI: 10.1056/CAT.21.0457

To facilitate the development of machine-learning (ML) models in care delivery, which remain poorly understood and executed, Stanford Medicine targeted an effort to address this *implementation gap* at the health system by addressing three key challenges: developing a framework for designing integration of artificial intelligence (AI) into complex health care work systems; identifying and building the teams of people, technologies, and processes to successfully develop and implement AI-enabled systems; and executing in a manner that is sustainable and scalable for the health care enterprise. In this article, the authors describe two pilots of real-world implementations that integrate AI into care delivery: one to improve advance care planning and the other to decrease unplanned escalations of care. While these two implementations used different ML models for different use cases, they shared a set of principles for integrating AI into care delivery. The authors describe how these shared principles were applied to the health system, the barriers and facilitators encountered, and how these experiences guided processes for collaboratively designing and implementing user-centered AI-enabled solutions.

View Article at catalyst.nejm.org

KEY TAKEAWAYS

- Artificial intelligence (AI) is not the end product, but rather an enabling function in the form of machine-learning (ML)–generated predictions that power a broader set of digital applications, workflows, and human teams (i.e., an *AI-enabled system*).
- The AI-enabled system must be designed and implemented in a manner that is user centered and driven by pragmatic needs and challenges. The full impact of AI on work systems may emerge via second- or third-order effects that can only be observed once it is implemented in the real-world setting.
- We observed AI playing an important role in aligning care teams around new collaborative workflows that previously did not exist. The role of AI was not necessarily to provide new information or to replace clinical decision-making, but to function as a *dispassionate mediator* of risk, which mitigated disagreements among team members and empowered nonphysician care team members to drive elements of patient care, such as advance care planning and care escalation due to clinical deterioration.
- A cross-functional team centered around the development and implementation of the AI system is needed, and it must have expertise not just in ML and data science, but also in clinical informatics, quality improvement, design thinking, enterprise analytics, software and IT applications, and clinical operations.

The Challenge

Despite broad interest in and the promise of artificial intelligence (AI) in health care, there remains a lack of understanding of how AI can meaningfully improve care in complex health care environments. While there has been significant progress in developing machine-learning (ML) models for generating predictions that underlie the *intelligence* that comes with AI, the manner in which this intelligence can be incorporated into health care delivery is still poorly understood and demonstrated.[1,2]

We classify this *implementation gap* at our health system into three categories of challenges: (1) developing a framework for designing integration of AI into complex health care work systems,[3] (2) identifying and building the teams of people, technologies, and processes to successfully develop and implement AI-enabled systems, and (3) executing in a manner that is sustainable and scalable for the health care enterprise.

We describe the application of a shared set of principles for using AI to guide care in two real-world implementations at an academic medical center: one to improve advance care planning (ACP) and the other to decrease unplanned escalations of care for clinically deteriorating patients in the hospital. From these two implementations, we observed an emergent characteristic of how AI was able to mediate improvement, which was to enable new team-based workflows for patient-centered care through empowering nonphysician clinical support services.

Back to Table of Contents

The Goal

We sought to demonstrate an approach to using AI in health care that could be operationalized and applied to real-world implementations. A key principle was to view AI not as the solution, but as an enabling function of a broader work system consisting of digital applications, workflows, and human teams.

This approach was applied in two different improvement opportunities at our institution:

1. ACP: conversations that elucidate a patient's values and goals in the course of treating a serious illness are infrequently conducted in the hospital setting. This may lead to care that is not concordant with the patient's goals. The inpatient setting was identified as an opportunity for improving rates of ACP for hospitalized patients.

2. Appropriate care escalation: delayed identification and care of clinically deteriorating hospitalized patients, leading to rapid response teams (RRTs), code events, and unplanned escalations to the ICU that can affect patient morbidity and mortality.

The Execution

Both implementations followed the principle of "designing and building the best possible system for the given improvement opportunity using ML capabilities" rather than "implementing a given ML model." We define a *model* as a function learned from data that map a vector of predictors to a real-valued outcome. *Predictors* are also referred to as *inputs*, *features*, or *variables*; the *outcome* is also referred to as *output*, *label*, or *task*. The following questions guided our execution:

1. What are the improvement goals, metrics, current-state processes, pain points, root causes, and key drivers for improvement?

2. What features of workflows and digital tools would address these key drivers? Which of these can be enabled by AI?

3. What parameters of the ML model (e.g., prediction task, predictive accuracy, and classification threshold) would be required to generate the intelligence that enables these key drivers?

4. How do we select the appropriate ML models that meet these requirements? Do we buy, build, or codevelop? How do we validate and customize the ML models for our improvement needs?

5. How do we design, build, iterate, and implement AI-enabled workflows and applications in a manner that is user centered and problem driven, adaptive to the complexity of the health care environment, and scalable and sustainable for the enterprise?

6. How do we evaluate and scale these implementations?

> *"A key principle was to view AI not as the solution, but as an enabling function of a broader work system consisting of digital applications, workflows, and human teams."*

1. Assessment of the Improvement Opportunity

Both implementations yielded a key insight into how AI can mediate improvement for complex health care settings: by providing an objective benchmark that the entire care team can align around, even if they disagreed with the prediction. This alignment opened an opportunity for physicians and nonphysicians to arrive at a shared mental model of risk that enabled coordination and empowerment of nonphysician team members to take necessary actions.

We arrived at this insight by first trying to understand the problem without any predefined notions of how AI was to be used (or even that AI was needed for the solution). We also used methods from quality improvement to identify concrete improvement goals (increase rates of ACP documentation and decrease rates of unexpected escalations of care from clinical deterioration) and two key drivers that could be enabled by AI.

Key Driver #1: Consistent, Objective Assessment and Communication of Risk

ML models that run continuously and generate risk predictions from patient data in the electronic health record (EHR) can offer an advantage over manual clinician assessments.[4] In both implementations, ML models enabled this key driver by providing consistency and objectivity to the assessment of appropriateness of ACP and need for care escalation.

Key Driver #2: Shared Mental Model of Risk Between Physician and Nonphysician Members of the Care Team

AI can facilitate alignment and coordination by acting as an objective assessor of risk. Patient care in a hospital, while supervised by the attending physician, is highly multidisciplinary, and patients interact with a variety of nonphysician clinical support services, such as nursing, rehabilitation services (physical and occupational therapy), social work, and nutrition. One root cause of process breakdowns for both projects, before implementation, was misalignment of risk perception and lack of coordination between physicians and nonphysician team members in performing needed clinical interventions. We found from stakeholder interviews that in times of disagreement, nonphysician team members frequently did not feel empowered to take action, which may have led to missed opportunities for ACP and early action for clinically deteriorating patients.

2. Conception of the AI-Enabled System

For each implementation of the AI enabled system, we designed a set of digital applications and workflows guided by these two key drivers of achieving consistent, objective assessment of risk

Back to Table of Contents

and a shared mental model across the care team. The following is a set of common features for both AI-enabled systems:

1. A clinical decision support (CDS) system in the EHR supported by ML model predictions that delivers the same information to both physicians and nonphysician members of the patient care team.

2. A standard, structured workflow that empowers nonphysician care team members to initiate action (within their scope of professional practice) guided by the AI-based CDS system.

3. A shared documentation tool in the EHR linked to the CDS system for each member of the care team to document completion of the workflow (and see each other's documentation).

These features were meant to disrupt the hierarchical, physician-driven workflows that existed for both ACP and care escalation for clinical deterioration and replace them with a more democratized and collaborative system that better leveraged the skills and resources of nonphysician clinical support services (Figure 1).

3. Development and Validation of the ML Models

The above design conceptions guided our selection and refinement of the ML models for each AI-enabled system. Both implementations required our team to think through the following questions:

1. What are the ML model prediction tasks that can enable the previously identified key drivers and system design?

2. What are the runtime requirements of the ML model (e.g., how frequently do predictions need to be generated at deployment)?

3. What is the validation strategy that can best reflect ML model performance for the local implementation setting? How do we select the cohort and outcomes used for the validation?

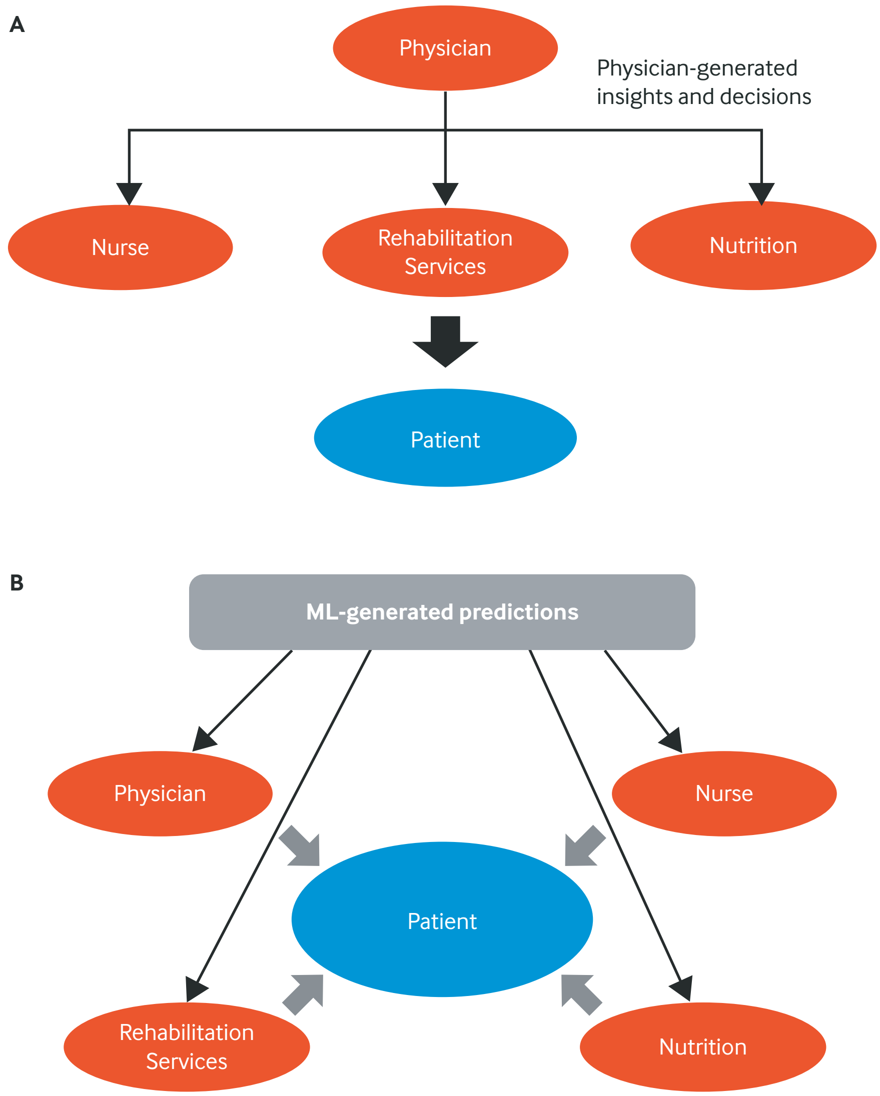
4. How do we select the appropriate classification thresholds for the ML model that can best meet the needs of the system?

> *"We found from stakeholder interviews that in times of disagreement, nonphysician team members frequently did not feel empowered to take action, which may have led to missed opportunities for ACP and early action for clinically deteriorating patients."*

FIGURE 1

Conceptualization of New Collaborative Workflows Enabled by Artificial Intelligence (AI)

For both advance care planning and care escalation, traditional hierarchical workflows (A) involved physicians generating insights and decisions that were then passed down to the rest of the care team and the patient. We envisioned an AI-enabled system (B) in which machine learning (ML)–generated predictions can empower and guide each member of the care team to initiate and carry out decisions in a more democratized and collaborative manner while removing the bottleneck at the level of the physician.



Source: The authors

Implementation #1: Increase Rates of ACP

We aligned on 12-month mortality risk for hospitalized patients as the ML prediction task. Predictions would need to be generated every 24 hours for all admitted patients because the clinical status (and appropriateness of ACP) of hospitalized patients can change over time. We selected a 12-month mortality risk prediction model developed previously by our team that had been validated as an appropriate surrogate for identifying hospitalized patients with serious illness who would benefit from ACP.[5]

The classification threshold was selected so that the model flagged patients in the top 25th percentile of predicted 12-month mortality risk in a cohort of patients discharged from the inpatient general medicine patients at our institution, which reflected the patient population for this implementation. At this threshold, in a larger validation cohort of 5,965 patients who were admitted to our institution, the positive predictive value was 60% (i.e., 60% of patients flagged by the ML model did in fact die within 12 months in the validation cohort). Finally, we estimated the increase in the amount of work in terms of the number of additional patients who would need ACP. In addition, a simulation study was conducted to quantify the achievable net benefit, given that work capacity constraints, as well as patient preferences, often limit follow-up with every flagged case.[6,7]

Implementation #2: Decrease Unplanned Escalation of Care for Clinical Deterioration

To align the care team on the appropriate early interventions, we determined that the ML model needed to identify patients with a high probability of a future clinical deterioration event (e.g., unplanned ICU transfer, RRT, or code), and the predictions would have to be performed early enough to allow for enough time for the care team to respond.[8,9] Predictions would also need to be updated in the EHR to reflect the frequent changes in the patients' clinical status, which enables the first key driver of providing a continuous assessment of risk.

We selected the Deterioration Index (DI), a model available through our EHR vendor, Epic Systems, because of the relative ease of technical integration while meeting most of these requirements. The DI is a logistic regression that is capable of updating predictions on hospitalized patients every 15 minutes using the most recent available clinical data on 31 physiological measures captured in the EHR; the DI tool also shows users the relative contributions of each physiological measure in generating the prediction. This last feature offers the additional benefit of providing a degree of model explainability, which can be useful for helping clinical users align around a shared mental model of risk.[10]

We then performed site-specific validation of the DI on a data set that we derived from a cohort of 6,232 non-ICU patient hospital encounters at our institution using a modified outcome definition that more closely reflected our product requirements: a composite outcome of RRT, code, or ICU transfer within 6 to 18 hours of the prediction.[11] This validation strategy was modified from that of the vendor, which reported model accuracy in predicting the outcomes *without* the 6- to 18-hour time lag; this was thought to not be clinically meaningful because a model predicting an event within 6 hours of the event would not provide sufficient time for a clinical response.

The area under the receiver operating characteristic (AUROC) (which is a performance metric for assessing ML models, in which 0.5 is the worst score and means the model is no better than random chance, and 1.0 is the best) calculated from our validation including these modified definitions was 0.71, which was lower than that reported by the vendor. Given this limited model discrimination and to simplify the model output so that it could be more easily interpreted by the care team, we chose a binary classification threshold (high risk vs. not high risk), which was selected at a cutoff that maximized precision and recall, both of which were 20%. We then validated with a focus group of clinicians that this level of accuracy would indeed be useful (i.e., most agreed they would want to be alerted if their patient had a "1 in 5 chance of experiencing an RRT or ICU transfer within the next 6–18 hours" while acknowledging that "four out of five patients who experience clinical deterioration would *not* be captured by the model"). While the low recall at this threshold (20%) would not make the DI an appropriate comprehensive screening tool for deterioration that would replace existing human-driven screening processes, there was consensus that, at a precision of 20%, it would still be useful to help align mental models and drive the desired physician–nurse team workflows for the patients whom the model *does* flag.

4. Design and Development of AI-Enabled Digital Applications and Workflows

Both implementations included digital applications embedded in the EHR that incorporated ML predictions and enabled shared workflows between physician and nonphysician team members. The EHR applications and workflows were created with two design aims in mind: (1) transparently communicate and align risk across the care team, and (2) promote consistency and collaboration toward patient care.

Communicate and Align Risk Across the Care Team

The following key product features were shared across the two implementations:

- ML predictions had to be translated and displayed into usable information that is simple and avoids confusion that could lead to unintended consequences.
- Information had to be integrated into the clinicians' standard work in the EHR.
- Information had to be displayed transparently to all care team members to facilitate a shared mental model and collaborative work across the care team.

> *"While the low recall at this threshold (20%) would not make the DI an appropriate comprehensive screening tool for deterioration that would replace existing human-driven screening processes, there was consensus that, at a precision of 20%, it would still be useful to help align mental models and drive the desired physician–nurse team workflows for the patients whom the model* does *flag."*

On the basis of these requirements, we developed a shared application design that was used by both implementations: a dedicated column that can be incorporated into EHR patient lists, which are used by both physician and nonphysician care team members as part of standard work. Within the column, patients identified as high risk by the ML are flagged (Figure 2). We decided that it was simpler and more useful for the care team to only see this binary classification result (high risk vs. not high risk) rather than individual numerical model predictions, given that neither model was optimized for calibration and that insight into individual predicted risk values was not necessary for alignment of mental models and workflow.

Because clinical deterioration and care escalation are more acute issues than ACP, we built additional alerting mechanisms in the form of *best practice alerts* in the EHR, as well as interruptive alerts to provider mobile devices for select instances (i.e., those cases in which the system flags the patient as high risk for the first time in the past 24 hours) (Figure 3).

Promote Consistency and Collaboration for Care Delivery

For this second design aim, both implementations shared the following key features:

- Structured workflow shared across the care team for patients flagged by the ML models
- Documentation tools in the EHR that promoted structure, collaboration, and transparency across the care team

We discovered that structure was important for aligning care teams around a collaborative clinical response for flagged patients. A key barrier to the adoption of AI systems in health care that we also observed in our implementation is that clinicians disagree with the ML predictions or believe that the AI system is not telling them anything that they do not already know. In our implementations, the emphasis was less on whether or not the ML predictions were correct; rather, it was that for any given patient flagged by the ML model, physician and nonphysician care team members had to carry out a structured collaborative workflow to build a shared mental model of risk and a collaborative clinical response *regardless of whether there is agreement with the ML prediction.* The role of the AI system was not necessarily to provide new information or to replace clinical decision-making, but to function as a *dispassionate mediator* for facilitating physician and nonphysician collaboration to assess the care plan in light of the new ML-generated information.

To promote consistency in this collaboration, we created the following structured workflows for each implementation.

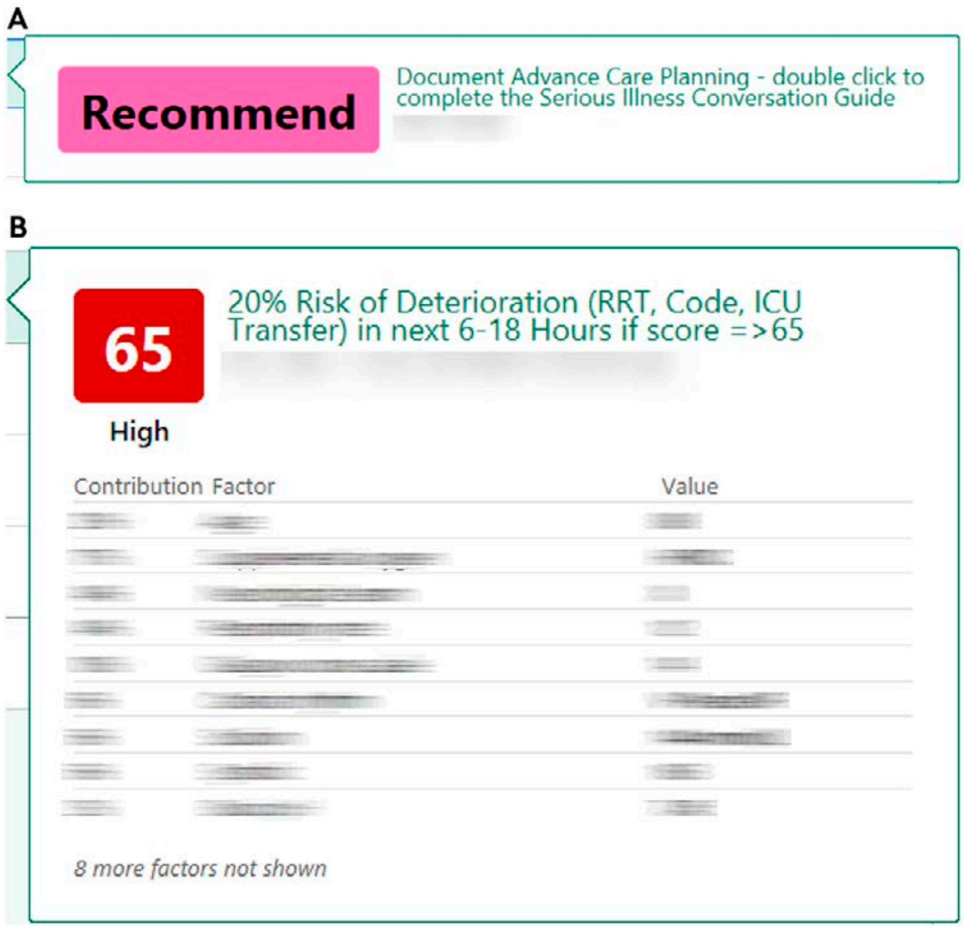
Shared Completion of ACP by Physician and Nonphysician Care Team Members

For patients flagged by the 12-month mortality prediction model, the care team (starting with physicians and occupational therapists and with plans to expand to social worker and clinical nutrition) is asked to conduct ACP using the Serious Illness Conversation Guide (SICG), a

FIGURE 2

Electronic Health Record (EHR) Designs for Communicating Machine-Learning Model Predictions

A patient list column in the EHR was created for each implementation that could be added by both physicians and nonphysician team members to their daily patient lists. Flags were displayed and messages provided when recommended by the machine-learning models. For advance care planning (A), the message was a simple prompt to document advance care planning. For additional evaluation for care escalation (B), the Deterioration Index also included a feature that shows the relative statistical contributions of each variable to the prediction, which provided a particularly helpful context within which clinicians could determine if the model predictions "made sense" (i.e., if the prediction was way off from their clinical judgment, then they could check to see if any of the variables perhaps were derived from either incorrect or outdated data in the EHR). RRT = rapid response team.

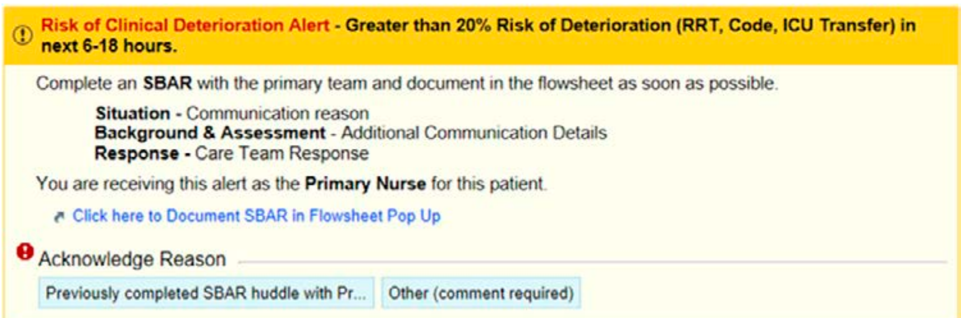


Source: The authors. © 2022 Epic Systems Corporation.

FIGURE 3

Artificial Intelligence–Generated Alerts Sent to Care Teams

A noninterruptive alert is shown on the homepage of the patient's chart if a patient is identified as high risk for clinical deterioration. Additionally, an interruptive alarm is sent to the clinician's phone via the clinical communication mobile application used for patient care (Voalte) for patients who are newly identified as high risk within the previous 24 hours (not shown). RRT = rapid response team, SBAR = Situation, Background, Assessment, Recommendation.



standardized template for ACP using patient-tested language developed by Ariadne Labs (Figure 4).[12,13]

The SICG provides a consistent approach toward ACP that also allows for distributing ownership of different components of the ACP conversation across care team members. In our implementation, the physician is expected to conduct the prognosis component of the conversation, while occupational therapy and nutrition may explore critical abilities that are important near the end of life. Care team members participating in this workflow all underwent training in how to use the SICG.

> *"The role of the AI system was not necessarily to provide new information or to replace clinical decision-making, but to function as a* dispassionate mediator *for facilitating physician and nonphysician collaboration to assess the care plan in light of the new ML-generated information."*

Completion of a Structured Group Huddle for Patients at Risk of Unplanned Care Escalation

The physician and nurse caring for a patient flagged by the ML model were expected to complete a structured huddle — referred to as the *clinical deterioration huddle* — to collaboratively discuss potential reasons for clinical deterioration and next steps (Table 1).

FIGURE 4

The Serious Illness Conversation Guide

The Serious Illness Conversation Guide is a validated template for advance care planning using patient-tested language developed by Ariadne Labs. The Serious Illness Care Program at Stanford Medicine adopted this tool for use by care teams.

Serious Illness Conversation Guide

SETUP

I'd like to talk about what is ahead with your illness and do some thinking in advance about what is important to you so that I can make sure we provide you with the care you want. **Is that okay?**

ASSESS

What is your **understanding** now of where you are with your illness?

How much **information** about what is likely ahead with your illness would you like from me?

SHARE PROGNOSIS

I want to share with you my **understanding** of where things are with your illness.

***Uncertain*:** It can be difficult to predict what will happen with your illness. I **hope** you will continue to live well for a long time, but I'm worried that you could get sick quickly, and I think it is important to prepare for that possibility.

OR

Time: I wish we were not in this situation, but I'm **worried** that time may be as short as ___ (*express as a range, e.g. days to weeks, weeks to months, months to a year*).

OR

Function: I hope that this is not the case, but I'm **worried** that this may be as strong as you will feel, and things are likely to get more difficult.

EXPLORE

What are your most important **goals** if your health situation worsens?

What are your biggest **fears and worries** about the future with your health?

What gives you **strength** as you think about the future with your illness?

What **abilities** are so critical to your life that you can't imagine living without them?

If you become sicker, **how much are you willing to go through** for the possibility of gaining more time?

How much do your **loved ones** know about your priorities and wishes?

CLOSE

I've heard you say ____. Keeping that in mind, and what we know about your illness, I **recommend** that we ____. This will help us make sure that your treatment plans reflect what's important to you.

How does this plan seem to you? **We will do everything we can** to help you through this.

Handoff

To colleague: "I talked with the patient about _____. I learned _____. I think they would benefit from talking with you about _____."

Stanford MEDICINE | Serious Illness Care Program Department of Medicine

http://med.stanford.edu/advancecareplanning

Source: Ariadne Labs, https://www.ariadnelabs.org/serious-illness-care/

Table 1. SBAR Clinical Deterioration Huddle

Situation	Patient at high risk of clinical deterioration
Background, **A**ssessment	Discuss nursing concerns (primary nurse) and likely reason(s) for clinical deterioration (provider team)
Recommendation	Discuss response to risk of clinical deterioration • Assess aspiration risk • Transfer to high level of care • New orders • Goals-of-care discussion • Family meeting • New consult • ICU provider team consult • Critical care response nurse consult • Other (comment)

Physicians and nurses were expected to complete a structured huddle for flagged patients using the Situation, Background, Assessment, Recommendation (SBAR) format. Source: The authors

The shared checklist format is meant to facilitate consistent incorporation of both physician and nursing perspectives, which we identified as a key driver for improvement.

Integration of AI-Enabled Collaborative Workflows into the EHR

Execution and adherence to these workflows were challenges in both implementations. Clinicians are busy, and their attention is spread out over many complex tasks when caring for patients, so bandwidth for accommodating any new initiative is limited. To address this challenge, we built shared documentation tools in the EHR that incorporated the structure of each AI-enabled workflow and promoted transparency and accountability by each care team member. These tools were all easily accessible by clicking on the patient list flag and alerts generated by the AI system in the EHR.

For the clinical deterioration implementation, tools were incorporated into the EHR to prompt and document the deterioration huddle (Figure 5).

For ACP, the SICG was built into a structured form that allowed care team members to collaboratively conduct ACP and document different sections of the SICG and also for other providers to look back and reference the ACP conversations that have taken place for a patient (Figure 6).

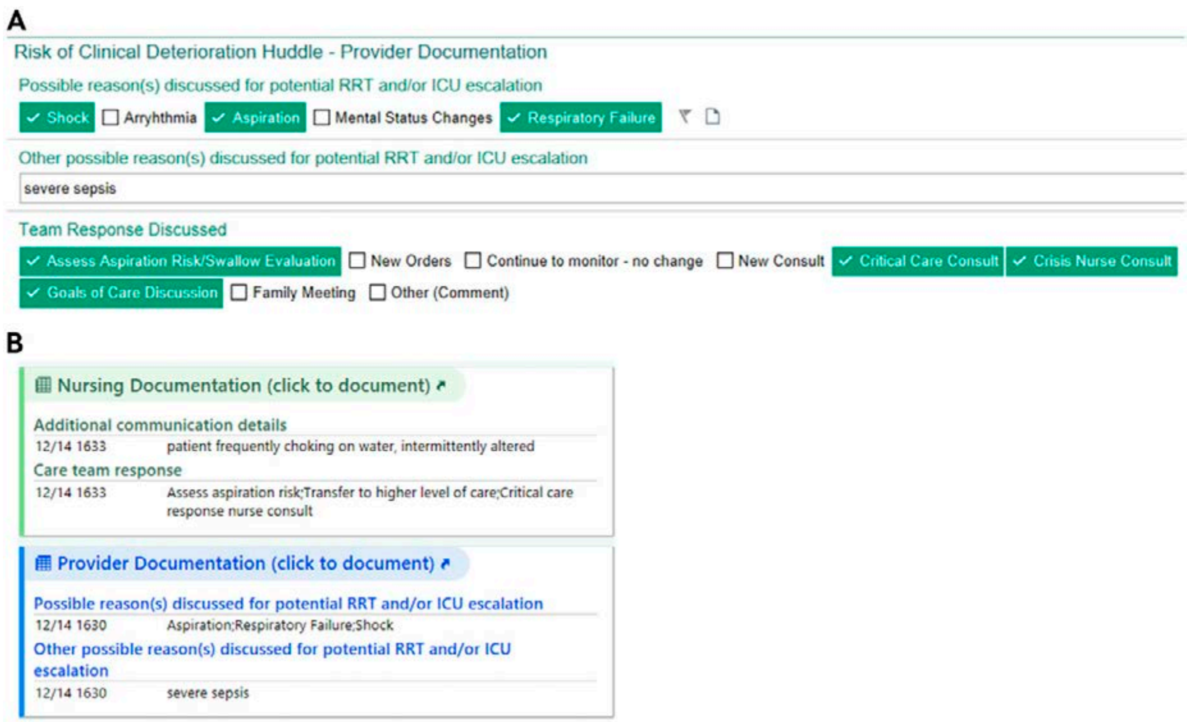
5. Implementation and Testing of Applications and Workflows

Both implementations were initiated and tested on pilot patient care units (the ACP project started in July 2020 and the clinical deterioration project started in January 2021) and followed a *Plan*, *Do*, *Study*, *Act* (PDSA) cycle framework from quality improvement.[14,15] Rapid iteration and testing with deep stakeholder engagement were critical to understanding the barriers and facilitators to implementation. Many design decisions were made after several PDSA cycles that could have surfaced only after real-world experiences and feedback from end users.

FIGURE 5

Collaborative Electronic Health Record (EHR) Documentation Tool for the Clinical Deterioration Huddle

Physicians and nurses complete a checklist for the structured huddle that is embedded into the EHR alert (A). Physician and nursing contributions to the documented huddle are then shown in a report in the patient's chart (B). RRT = rapid response team.



Source: The authors. © 2022 Epic Systems Corporation.

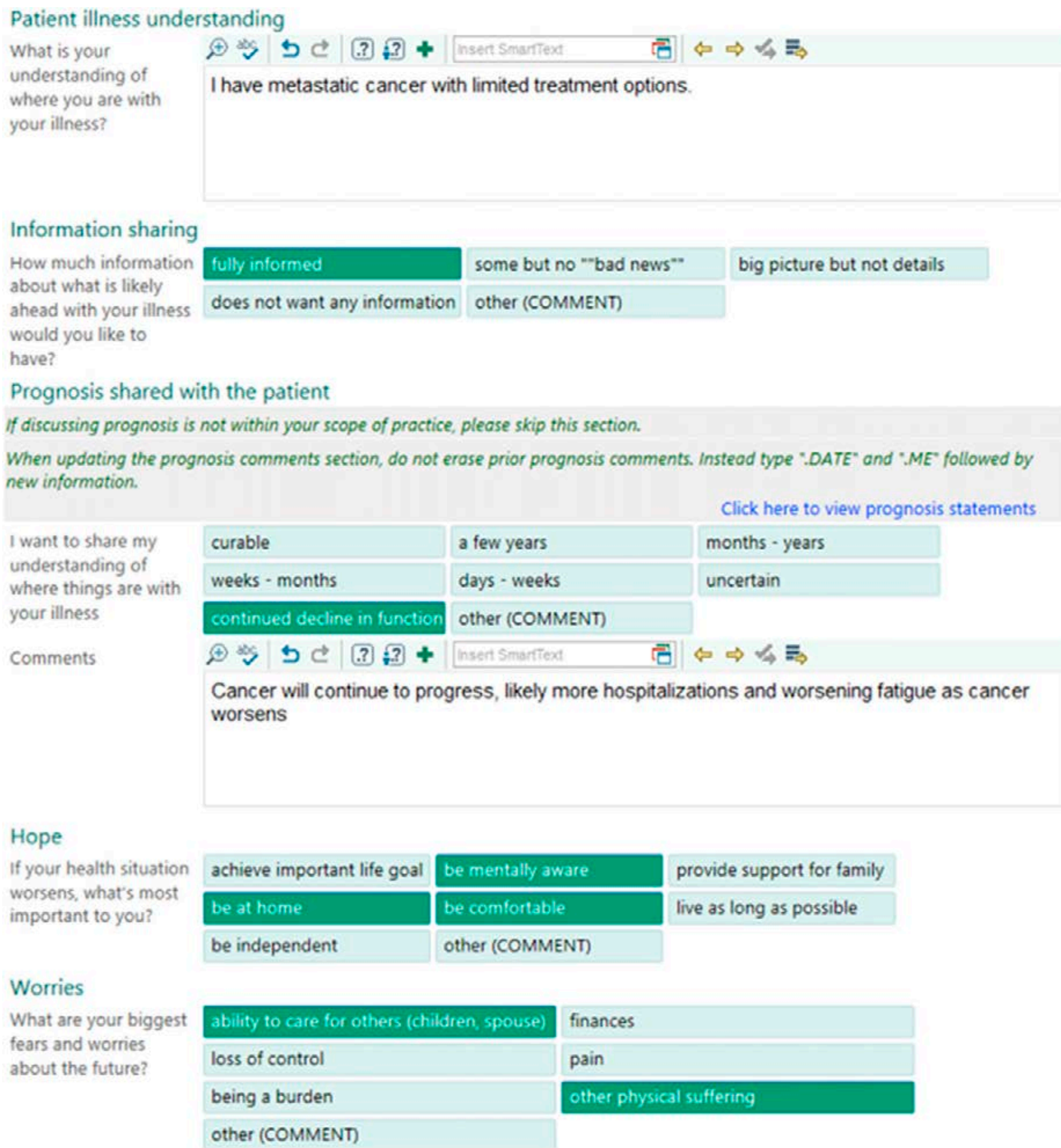
> *"Both implementations have yielded early promising results during the initial pilot phases, as measured by the documented workflow completion rate and interviews with workflow participants."*

For example, for ACP, we initially designed a more coordinated workflow between physicians and nonphysician care team members for flagged patients that included a huddle prior to initiating sections of the SICG. However, on the basis of user feedback regarding bandwidth constraints and the nonacute nature of the ACP relative to other inpatient patient care needs, we instead elected to use a workflow in which any care team member can initiate a section of the SICG for flagged patients (as long as it is within their scope of practice; only physicians were to

FIGURE 6

Collaborative Electronic Health Record (EHR) Documentation Tool for Advance Care Planning

The Serious Illness Conversation Guide was embedded into the EHR as a shared documentation template that can be accessed by double-clicking on the patient list flag. Both physicians and nonphysician care team members can access and edit this documentation template.



Source: The authors. © 2022 Epic Systems Corporation.

discuss the prognosis section); team members would then send a *for-your-information* message to the rest of the team once completed. The AI-generated patient list flag was crucial in facilitating the needed level of alignment for an otherwise decentralized workflow.

6. Integration and Scale into the Health System

As of January 2022, both implementations were nearing the end of their pilot phases, with plans for integration and scale guided by these principles:

Clinical Integration: Ensure that both of the pilot AI-enabled products are sufficiently incorporated into the standard clinical workflows for the care teams to facilitate uptake, beyond just early adopters.

Operational Integration: Connect both implementations to operational units within our institution that are accountable for the metrics and operational goals that these implementations enable.

Technical Integration: Utilize technical infrastructure that can sustainably support the back and front ends of these AI-enabled products at the enterprise level and create a system for monitoring, versioning, and even deimplementing if appropriate.

Metrics

The ACP pilot was implemented for all patients admitted to the general medicine inpatient service, which thus far has included 11,881 total patient hospital encounters since the beginning of the implementation (July 2020) to January 2022 (average of 625 encounters per month), with 2,627 patient encounters flagged by the ML model as candidates for ACP (138 per month; 22% of total encounters).

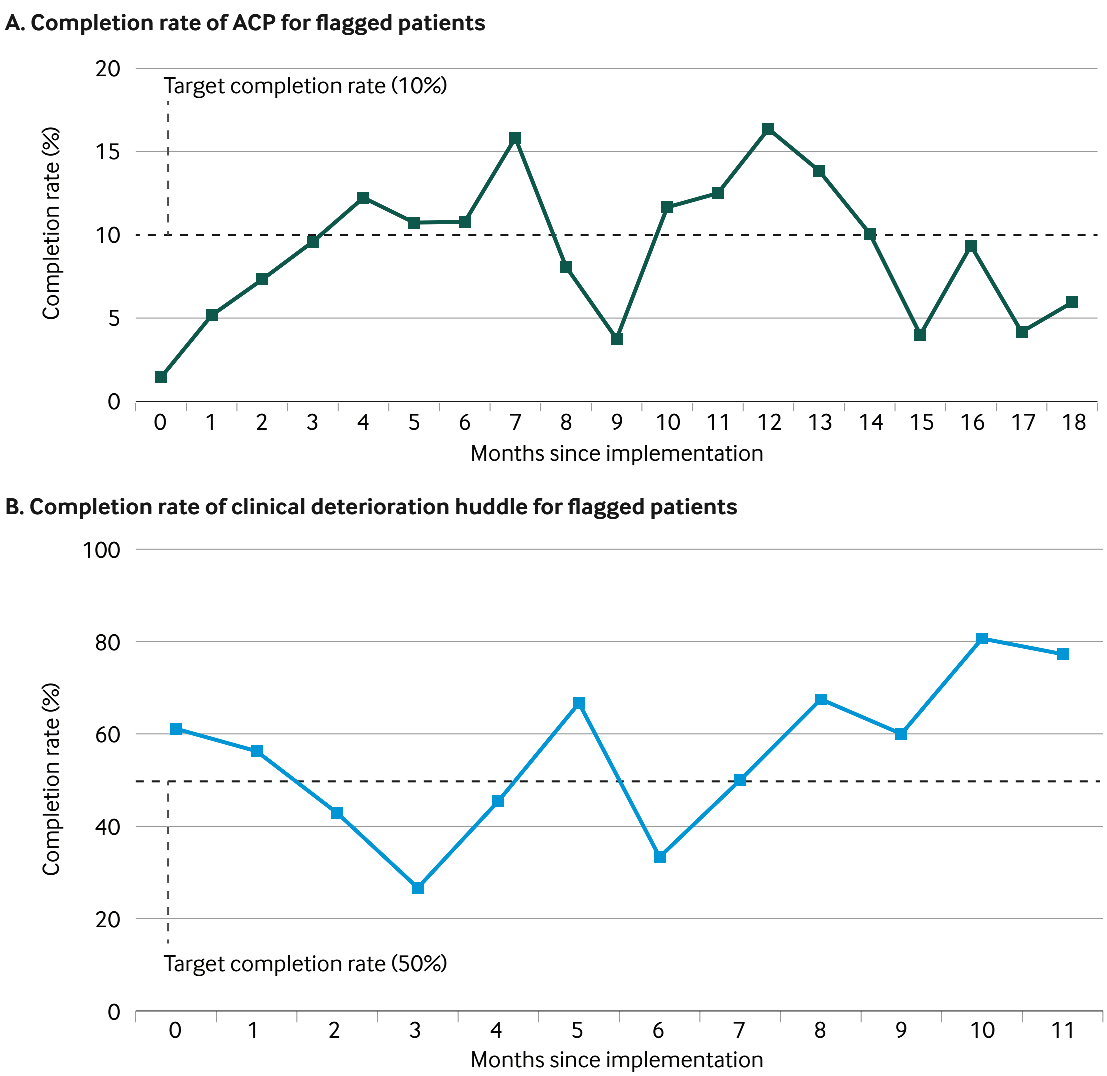
The clinical deterioration pilot was implemented in a stepwise fashion across two different nursing units for general medicine patients, which thus far has included 3,022 total patient encounters since the beginning of the implementation (January 2021) to January 2022 (average of 252 encounters per month), with 313 total patient encounters experiencing at least one flag generated by the DI (average of 21 flags per month; 10.3% of total encounters).

Both implementations have yielded early promising results during the initial pilot phases, as measured by the documented workflow completion rate (Figure 7) and interviews with workflow participants. Target completion rates (number of documented completed workflows/number of flagged encounters) were established for each implementation on the basis of our assessment of clinical appropriateness, estimated number of flagged encounters, time needed to complete each workflow, and capacity of the clinical teams. For ACP, we established a target of 10%, given the higher number of expected flagged encounters, the amount of time needed to complete ACP conversations, and the relatively lower urgency of the intervention for an inpatient encounter. Conversely, we set a higher target (50%) for the clinical deterioration implementation because there are fewer expected flagged encounters, and it was more urgent clinically to complete a

FIGURE 7

Workflow Component Completion Rate

Completion rate of documented advance care planning conversations for patients flagged by the 12-month mortality model is shown in A, and completion rate of documented clinical deterioration huddles is shown in B. Both implementations sustained their target completion rates of 10% and 50%, respectively. There was variability over time due to factors such as varying degrees of work capacity among clinicians and staff and how often the clinical teams deemed the interventions appropriate for the flagged patients. Notably, in the seventh month of the pilot, completion rates for the clinical deterioration huddle increased after an upgrade in the electronic health record documentation tool that improved ease of documentation in July 2021. Note: month 0 for A is July 2020 and for B is January 2021.



Source: The authors

huddle for deteriorating patients; nevertheless, given the precision of 20% (meaning four out of five flagged patients in the validation cohort did not end up needing escalation to the ICU or experienced an RRT or code), we chose not to target 100% workflow completion, because there inevitably will be flagged encounters that clinical teams appropriately decide would not need the full workflow. One limitation to this metric of *documented* workflow completion rate is that it likely underestimates the true rate of workflow completion, because not all completed ACP conversations and clinical deterioration huddles were documented correctly by the clinical teams.

> *"We encountered a number of challenges associated with the implementations, including matters related to time and resources, translating ML predictions into accessible and actionable information, and securing clinician buy-in for the effort. Ultimately, the core of each of these hurdles is rooted in the need to establish trust and confidence in the value of the ML integration."*

We observed sustained participation from nonphysician care team members for both implementations; of the patient encounters with completed workflows, 100% of completed clinical deterioration huddles included contribution from a nurse, and 42% of completed ACP conversations included contribution from a care team member who was not a physician. In interviews for feedback, many nonphysician members reported that they felt more empowered to leverage their skills to advocate and care for their patients in ways that were not possible before. For example, occupational therapists — who previously were typically not part of ACP conversations (although they had expressed the desire to be so) — reported that they often were able to offer unique perspectives around patient functional goals in the new AI-enabled workflow. One occupational therapist expressed:

"I loved having the [SICG] conversation with a patient today because it really gave me a good understanding of who the patient is as a human being. It was so interesting to see how each person has similar but unique priorities in regard to their medical care and functional goals. The conversation gives us a unique perspective to plan care based on what is important to the patient."

Nurses have also expressed strong interest in and satisfaction with the AI-enabled clinical deterioration workflow. In a survey of nursing staff from the first of the two pilot nursing units (52 nurses, 30 responded; 57%) on which this implementation was deployed, 96.5% reported that they felt the workflow was adding value to patient care. Additionally, 89.6% indicated that the tool changes the way they care for their patients: charge nurses in the survey reported alternating patient assignments or ratios in anticipation of clinical changes with the flagging patient, and bedside nurses reported they rounded more frequently and/or completed a more in-depth patient assessment on their patients who were flagging.

While nurses have consistently documented completion of the huddles, physician documentation adherence has been minimal. However, survey results shed more light on physician participation and remaining challenges. In a survey among 19 medicine residents participating in the pilot, of whom 17 had at least one patient flagged by the clinical deterioration model, 50% indicated that they take action on the alerts by calling the bedside nurse to huddle, messaging the bedside nurse, or going to the bedside to huddle with the nurse. In addition, 50% indicated that no personal action is taken on the alert; however, 64% said that after receiving an alert, the bedside nurse also reached out to them to discuss the patient's status. When asked about challenges to workflow adherence, 30% of physicians indicated that when they received the alert, they had recently assessed the patient, and, therefore, further action seemed redundant. Providers additionally cited overall workload burden (14%), and disagreement with the model's assessment of risk (14%). Most respondents (68%) reported that they feel either neutral or positive about the overall usefulness of the intervention. These survey results are spurring important conversations and informing key improvements to the overall intervention.

Both implementations will continue to be assessed prospectively with additional quantitative and qualitative outcomes that reflect clinical effectiveness, impact on processes and teams, and success of implementation using implementation science frameworks, such as RE-AIM[16] (reach, effectiveness, adoption, implementation, maintenance) once the pilots have reached a steady state in adoption rate and changes to the workflow (Table 2). Given external factors, such as the Covid-19 pandemic, that have led to multiple unforeseen changes and demands on resources and staffing in the hospital, both of the pilots were extended for an additional 6–8 months to accommodate more PDSA cycles. Examples of additional clinical, process, and implementation metrics for both projects are listed in Table 2.

Hurdles

We encountered a number of challenges associated with the implementations, including matters related to time and resources, translating ML predictions into accessible and actionable information, and securing clinician buy-in for the effort. Ultimately, the core of each of these hurdles is rooted in the need to establish trust and confidence in the value of the ML integration.

Managing Uncertainty Regarding the Value of the ML Component in the Context of Competing Demands for Time and Resources Among Care Teams

Integrating novel workflows into health care is often challenging when there are competing demands for time and resources, especially with the record surges in patient volume our institution has experienced over the course of the implementation period (due to the Covid-19 pandemic and other factors). In particular, workflows involving AI can face a higher barrier to acceptance, because the mechanism triggering the workflow (the ML model) will, by definition, be wrong some percentage of the time (i.e., there is only a certain probability that the patient flagged by the ML model is, indeed, appropriate for the workflow). Additionally, the timing of when these workflows are triggered is critical to adoption and perceived usefulness. This degree of uncertainty can be difficult to understand and accept by clinical teams, especially when other workflows competing for their time and attention are presented with more certainty about expected patient benefit (even if that level of certainty is likely false). For this reason, we designed the EHR

Table 2. Planned Clinical, Process, and Implementation Outcomes for the Two AI-Enabled Systems

	Advance care planning	Clinical deterioration
Clinical effectiveness and process: How did the intervention impact clinical outcomes and processes of care?	• Rates of referral to hospice, palliative care specialists, changes in code status, and hospital readmissions • Frequency and quality of communication and collaboration between physicians and nonphysicians related to patient goals of care	• Rates of overall inpatient mortality, ICU transfers, mortality 24 hours after ICU transfer, conversion of RRTs into codes or ICU transfers • Frequency and quality of communication and collaboration between physicians and nurses related to clinical deterioration
Implementation: How well was the intervention implemented?	• Reach: proportion of flagged patients flagged by the ML model who received ACP • Adoption: proportion of eligible physician and nonphysician providers who participate in the workflow • Implementation fidelity: completion of documented ACP, stratified by provider type and SICG section	• Reach: proportion of patients flagged by the ML model for whom the clinical teams performed a clinical deterioration huddle • Adoption: proportion of eligible physician and nurses who participate in the workflow • Implementation fidelity: completion of documented clinical deterioration huddle, stratified by provider type and components of the huddle

One particular area of focus will be to ascertain the potential impact of the AI system on the level of communication and collaboration between physicians and nonphysician providers related to advance care planning (ACP) and clinical deterioration, which we hypothesize to be a potential downstream effect of the implementations. AI = artificial intelligence, RRT = rapid response team, ML = machine learning, SICG = Serious Illness Conversation Guide. Source: The authors

application builds for both implementations to transparently present the level of statistical uncertainty associated with each ML-generated prediction using user-centered, clinically oriented language so that users can more easily contextualize the relevance of each ML-generated alert.

> *"Implementation efforts will more likely be successful as an improvement opportunity in need of an ML model rather than as an ML model looking for an improvement opportunity."*

Translating the ML Model Predictions into Interpretable and Actionable Information

Patient care teams need to continuously process large amounts of new information. If that information is ambiguous or not clearly actionable, it is at risk of being misinterpreted, misused, or not used at all. An important lesson we learned is that the ML prediction may not itself be necessarily informative, yet it still plays the important role of aligning clinical teams around a standard set of downstream actions that, on average for flagged patients, may lead to better outcomes. For example, a common piece of feedback we received from clinicians, particularly physicians, for both the ACP and the clinical deterioration implementation was that the model was "not telling [them] anything that [they] don't already know," in the sense that they often were already aware that a patient either would benefit from ACP or was at risk of deteriorating. However, despite this prior awareness, physicians often did not actually perform the associated downstream tasks. Therefore, the true value of these AI systems was not necessarily to provide new information, but rather to align the physicians with the rest of the care team around acting on an established workflow.

To incorporate this concept early in each implementation, we pivoted from showing only model predictions to language that specifically outlines the appropriate interpretation and required action. For example, for patients flagged by the DI, nurses (and physicians) received an alert that concretely expressed the nature of the risk and next steps: *"Clinical Deterioration Risk Alert — [insert patient name] is predicted to be at high risk (greater than 20%) of requiring ICU transfer or an RRT in the next 6–18 hours. Connect with the charge nurse and primary team as soon as possible and complete required documentation."*

Building Clinician Trust and Buy-in for the Intervention

The teams employed three strategies to build trust in the models and buy-in for the workflow designed in these implementations. First, site-specific quantitative model validation was conducted for each model, and the results were shared with the clinical stakeholders during the participatory design sessions. Second, clinicians were directly involved in a parallel qualitative model validation process in which they indicated agreement or disagreement with the model predictions. Lastly, the team summarized and shared intervention success stories from early in the pilots to demonstrate patient-level benefit from the intervention. These stories included quotes from staff along with the case details and how the model output informed a different course of action and a favorable outcome.

The Team

In each of the two implementations, a multidisciplinary team consisting of technical, operational, and clinical stakeholders, along with project management and quality improvement support, was convened. More specifically, both project teams included about 15 members: data scientists, clinical informatics, enterprise analytics, nurse managers, frontline nurses, clinical nurse specialists, physicians, project managers, quality improvement experts, and social science researchers. The ACP project additionally included physical therapists, social workers, and dieticians. Engaging all levels of the technical, operational, and clinical stack is a key facilitator of rapid and well-informed decision-making across all phases of the development and implementation of AI-enabled solutions.

Where to Start

The key to starting an implementation project using AI is to pick the right problems to solve that will deliver meaningful improvement for the institution and then build a cross-functional team to develop and integrate the AI-enabled system. This is in contrast to starting with an ML model and trying to figure out how to implement it without a clearly defined problem. Implementation efforts will more likely be successful as an improvement opportunity in need of an ML model rather than as an ML model looking for an improvement opportunity.

Our two implementations went through the following six steps (Table 3) that can be applied to future opportunities.

Table 3. Six Steps for Implementing an AI Workflow Initiative*

Phase	Key components
Assessment of improvement opportunity	• Define the problem statement, improvement targets, and stakeholders • Identify current state gaps and key drivers for improvement that can be enabled by ML
Conceptualization of the AI-enabled system	• Design the components of the newly imagined sociotechnical system enabled by AI that addresses the key drivers
Development and validation of the ML models	• Define appropriate ML prediction tasks • Develop, select, and validate ML models on cohort that reflect the local implementation setting • Determine the appropriate classification thresholds that enable the key drivers and satisfy the work capacity of the team
Design and development of applications and workflows	• Design and build the user-facing digital applications and workflows
Implementation and testing	• Iterate and test the AI-enabled system using the PDSA cycle • Prospectively evaluate pilot implementations
Integration and scale	• Integrate and scale the AI-enabled system into the standard work and processes of the institution

*AI = artificial intelligence, ML = machine learning, PDSA = *Plan*, *Do*, *Study*, *Act.* Source: The authors

Ron C. Li, MD
Clinical Assistant Professor, Stanford University School of Medicine, Stanford, California, USA

Director, Stanford Emerging Applications Lab, Stanford Health Care, Stanford, California, USA

Medical Informatics Director for Digital Health and AI Integration, Stanford Health Care, Stanford, California, USA

Margaret Smith, MBA
Director of Operations, Stanford Healthcare AI Applied Research Team, Stanford University School of Medicine, Stanford, California, USA

Jonathan Lu, MS
Medical Student, Stanford University School of Medicine, Stanford, California, USA

Anand Avati, MS
Graduate Student, Stanford University, Stanford, California, USA

Samantha Wang, MD, MHS
Clinical Assistant Professor, Stanford University School of Medicine, Stanford, California, USA

Assistant Clinical Director, Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Winifred G. Teuteberg, MD
Clinical Associate Professor, Stanford University School of Medicine, Stanford, California, USA

Clinical Director, Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Kenny Shum, PhD
Data Scientist, Stanford Health Care, Stanford, California, USA

Grace Hong
Research Assistant, Stanford Healthcare AI Applied Research Team, Stanford University School of Medicine, Stanford, California, USA

Briththa Seevaratnam, MS
Manager, Quality Initiatives and Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Jerri Westphal, MSN, RN, CNML
Manager, Nurse Informatics, Stanford Health Care, Stanford, California, USA

Former Patient Care Manager, Stanford Health Care, Stanford, California, USA

Margaret Dougherty, MHA, MS, OTR/L
Manager, Rehabilitation Services, Stanford Health Care, Stanford, California, USA

Pooja Rao, MS, OTR/L, CSRS
Occupational Therapist, Stanford Health Care, Stanford, California, USA

Steven Asch, MD, MPH
Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Associate Dean for Research, Stanford School of Medicine, Stanford, California, USA

Associate Chief of Staff for Clinical Effectiveness, VA Palo Alto Health Care System, Palo Alto, California, USA

Steven Lin, MD
Clinical Associate Professor, Stanford University School of Medicine, Stanford, California, USA

Founder and Executive Director, Stanford Healthcare AI Applied Research Team, Stanford Medicine, Stanford, California, USA

Vice Chief for Technology Innovation, Division of Primary Care and Population Health, Stanford Medicine, Stanford, California, USA

Christopher Sharp, MD
Clinical Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Chief Medical Information Officer, Stanford Health Care, Stanford, California, USA

Lisa Shieh, MD, PhD
Clinical Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Associate Chief Quality Officer, Stanford Health Care, Stanford, California, USA

Nigam H. Shah, MBBS, PhD
Professor of Medicine and Biomedical Data Science, Clinical Excellence Research Center; Stanford University School of Medicine, Stanford, California, USA

Chief Data Scientist, Stanford Healthcare, Palo Alto, California, USA

Acknowledgments

We thank the following for their contributions: Melanie Hanson, Former Program Coordinator, Serious Illness Care Program, Stanford University School of Medicine, Stanford, CA; Rebecca Hgah-Guen Fong, Program Coordinator, Quality Initiatives and Serious Illness Care Program, Stanford University School of Medicine, Stanford, CA; Siyun Li, Research Assistant, Stanford University School of Medicine, Stanford, CA; Brian Schenone, RN-BC, CAPM, Lead Informatics Analyst, Stanford Health Care, Stanford, CA; Rob Maclay, Director of Technical Services, Stanford Health Care, Stanford, CA; Rob Gallo, MD, Resident Physician, Stanford Health Care, Stanford, CA; and David Shively, Communications Engineering, Stanford Health Care, Stanford, CA, Thomas Savage, MD, Resident Physician, Stanford Health Care, Stanford, CA, Flora Kechedjian, MSN, RN-BC, Nursing Professional Development Specialist, Stanford Health Care, Stanford, CA, Patrice Duhon, MSN, RN, Director, Epic Optimization and Reporting Patient Care Services, Stanford Health Care, Stanford, CA, Vi Van Nguyen RN, BSN, CMSRN, Epic Optimization Coordinator, Stanford Health Care, Stanford, CA, Gretchen Brown MSN, RN, NEA-BC Associate Chief Nursing Informatics Officer, Stanford Health Care, Stanford, CA, Advait Patil, Student, Stanford University, Stanford, CA.

Disclosures: Ron C. Li, Margaret Smith, Jonathan Lu, Anand Avati, Samantha Wang, Winifred G. Teuteberg, Kenny Shum, Grace Hong, Briththa Seevaratnam, Jerri Westphal, Margaret Dougherty, Pooja Rao, Steven Asch, Steven Lin, Christopher Sharp, Lisa Shieh, and Nigam H. Shah have nothing to disclose.

References

1. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020;3:107 https://www.nature.com/articles/s41746-020-00318-y https://doi.org/10.1038/s41746-020-00318-y.

2. Smith M, Sattler A, Hong G, Lin S. From code to bedside: implementing artificial intelligence using quality improvement methods. J Gen Intern Med 2021;36:1061-6 https://link.springer.com/article/10.1007%2Fs11606-020-06394-w.

3. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care. 2006;15(Suppl 1):i50-8 https://qualitysafety.bmj.com/content/15/suppl_1/i50 https://doi.org/10.1136/qshc.2005.015842.

4. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142-5 https://www.jclinepi.com/article/S0895-4356(21)00013-5/fulltext https://doi.org/10.1016/j.jclinepi.2021.01.009.

5. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO, November 13-16, 2017 https://ieeexplore.ieee.org/document/8217669 https://doi.org/10.1109/BIBM.2017.8217669.

6. Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc 2021;28:1149-58 https://academic.oup.com/jamia/article/28/6/1149/6045012 https://doi.org/10.1093/jamia/ocaa318.

7. Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. JAMA 2019;322:1351-2 https://jamanetwork.com/journals/jama/article-abstract/2748179 https://doi.org/10.1001/jama.2019.10306.

8. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11(Suppl 1):S18-24 https://shmpublications.onlinelibrary.wiley.com/journal/15535606 https://doi.org/10.1002/jhm.2652.

9. Kipnis P, Turk BJ, Wulf DA, et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. J Biomed Inform 2016;64:10-19 https://www.sciencedirect.com/science/article/pii/S1532046416301265?via%3Dihub https://doi.org/10.1016/j.jbi.2016.09.013.

10. McComb S, Simpson V. The concept of shared mental models in healthcare collaboration. J Adv Nurs 2014;70:1479-88 https://onlinelibrary.wiley.com/doi/10.1111/jan.12307 https://doi.org/10.1111/jan.12307.

11. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. Ann Am Thorac Soc 2021;18:1129-37 https://www.atsjournals.org/doi/10.1513/AnnalsATS.202006-698OC https://doi.org/10.1513/AnnalsATS.202006-698OC.

12. Ariadne Labs. Serious Illness Care. Accessed November 2, 2021. https://www.ariadnelabs.org/serious-illness-care/.

13. Jacobsen J, Bernacki R, Paladino J. Shifting to serious illness communication. JAMA Netw 2022;327:321-2 https://jamanetwork.com/journals/jama/article-abstract/2788065 https://doi.org/10.1001/jama.2021.23695.

14. Leis JA, Shojania KG. A primer on PDSA: executing plan-do-study-act cycles in practice, not just in name. BMJ Qual Saf 2017;26:572-7 https://qualitysafety.bmj.com/content/26/7/572 https://doi.org/10.1136/bmjqs-2016-006245.

15. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf 2014;23:290-8 https://qualitysafety.bmj.com/content/23/4/290 https://doi.org/10.1136/bmjqs-2013-001862.

16. King DK, Glasgow RE, Leeman-Castillo B. Reaiming RE-AIM: using the model to plan, implement, and evaluate the effects of environmental change approaches to enhancing population health. Am J Public Health 2010;100:2076-84 https://ajph.aphapublications.org/doi/10.2105/AJPH.2009.190959 https://doi.org/10.2105/AJPH.2009.190959.

NEJM Evidence

Published April 13, 2022
NEJM Evid 2022; 1 (6)
DOI: 10.1056/EVIDoa2200003

ORIGINAL ARTICLE

Real-Time Artificial Intelligence–Based Optical Diagnosis of Neoplastic Polyps during Colonoscopy

Ishita Barua, M.D.,[1,2] Paulina Wieszczy, Ph.D.,[1,2,3] Shin-ei Kudo, M.D.,[4] Masashi Misawa, M.D.,[4] Øyvind Holme, M.D.,[1,2,5] Shraddha Gulati, M.D.,[6] Sophie Williams, M.D.,[6] Kensaku Mori, Ph.D.,[7] Hayato Itoh, Ph.D.,[7] Kazumi Takishima, M.D.,[4] Kenichi Mochizuki, M.D.,[4] Yuki Miyata, M.D.,[4] Kentaro Mochida, M.D.,[4] Yoshika Akimoto, M.D.,[4] Takanori Kuroki, M.D.,[4] Yuriko Morita, M.D.,[4] Osamu Shiina, M.D.,[4] Shun Kato, M.D.,[4] Tetsuo Nemoto, M.D.,[8] Bu Hayee, M.D.,[6] Mehul Patel, M.D.,[6] Nishmi Gunasingam, M.D.,[6] Alexandra Kent, M.D.,[6] Andrew Emmanuel, M.D.,[6] Carl Munck, M.D.,[9] Jens Aksel Nilsen, M.D.,[9] Stine Astrup Hvattum, M.D.,[9] Svein Oskar Frigstad, M.D.,[9] Petter Tandberg, M.D.,[9] Magnus Løberg, M.D.,[1,2] Mette Kalager, M.D.,[1,2] Amyn Haji, M.D.,[6] Michael Bretthauer, M.D.,[1,2] and Yuichi Mori, M.D.[1,2,4]

Abstract

BACKGROUND Artificial intelligence using computer-aided diagnosis (CADx) in real time with images acquired during colonoscopy may help colonoscopists distinguish between neoplastic polyps requiring removal and nonneoplastic polyps not requiring removal. In this study, we tested whether CADx analyzed images helped in this decision-making process.

METHODS We performed a multicenter clinical study comparing a novel CADx-system that uses real-time ultra-magnifying polyp visualization during colonoscopy with standard visual inspection of small (≤5 mm in diameter) polyps in the sigmoid colon and the rectum for optical diagnosis of neoplastic histology. After committing to a diagnosis (i.e., neoplastic, uncertain, or nonneoplastic), all imaged polyps were removed. The primary end point was sensitivity for neoplastic polyps by CADx and visual inspection, compared with histopathology. Secondary end points were specificity and colonoscopist confidence level in unaided optical diagnosis.

RESULTS We assessed 1289 individuals for eligibility at colonoscopy centers in Norway, the United Kingdom, and Japan. We detected 892 eligible polyps in 518 patients and included them in analyses: 359 were neoplastic and 533 were nonneoplastic. Sensitivity for the diagnosis of neoplastic polyps with standard visual inspection was 88.4% (95% confidence interval [CI], 84.3 to 91.5) compared with 90.4% (95% CI, 86.8 to 93.1) with CADx ($P=0.33$). Specificity was 83.1% (95% CI, 79.2 to 86.4) with standard visual inspection and 85.9% (95% CI, 82.3 to 88.8) with CADx. The proportion of polyp assessment with high confidence was 74.2% (95% CI, 70.9 to 77.3) with standard visual inspection versus 92.6% (95% CI, 90.6 to 94.3) with CADx.

Drs. Barua, Wieszczy, and Kudo contributed equally as co-first authors, and Drs. Haji, Bretthauer, and Mori contributed equally as co-last authors.

The author affiliations are listed at the end of the article.
Dr. Bretthauer can be contacted at michael.bretthauer@medisin.uio.no or at Clinical Effectiveness Research Group, Institute of Health and Society, University of Oslo, Oslo 0027, Norway.

View Article at evidence.nejm.org

Back to Table of Contents

CONCLUSIONS Real-time polyp assessment with CADx did not significantly increase the diagnostic sensitivity of neoplastic polyps during a colonoscopy compared with optical evaluation without CADx. (Funded by the Research Council of Norway [Norges Forskningsråd], the Norwegian Cancer Society [Kreftforeningen], and the Japan Society for the Promotion of Science; UMIN number, UMIN000035213.)

Introduction

Colorectal cancer is the third most common cancer and the second leading cause of cancer deaths worldwide.[1] Removal of precancerous polyps during colonoscopy is the cornerstone of colorectal cancer screening. Most colorectal polyps are small (≤5 mm in diameter) and located in the sigmoid colon and the rectum. Although most colorectal cancers develop from polyps, many small polyps are not neoplastic and do not have any malignant potential.[2]

With current standard colonoscopy equipment, many endoscopists, especially those with less experience, cannot reliably distinguish between neoplastic and nonneoplastic polyps on visual inspection, a procedure known as "optical diagnosis."[3,4] Therefore, the current standard of care is to remove all polyps and submit them for histopathologic diagnosis. Reliable real-time optical diagnosis of small polyps during colonoscopy could enable targeted removal only of polyps classified as neoplastic, while small, nonneoplastic polyps could be left behind.[5]

In a recent single-center, proof-of-concept study of a novel artificial intelligence (AI) system for computer-aided polyp diagnosis (CADx), we achieved a reliable distinction between small neoplastic and nonneoplastic polyps in the distal colon and the rectum.[6] The CADx system combines colonoscopes with 520× magnification of polyp surfaces during colonoscopy in real time, and it enables AI-derived automated optical diagnosis of neoplastic and nonneoplastic polyps in about 40 seconds. The automated diagnosis is signaled to the colonoscopist by an acoustic and optical alarm during each polyp assessment.[6-8]

The current multicenter clinical study was designed to compare the clinical performance of AI CADx-based optical diagnosis in distinguishing neoplastic from nonneoplastic small polyps in the sigmoid colon and the rectum during colonoscopy with standard visual inspection–based optical diagnosis in routine clinical colonoscopy practice.

Methods

STUDY DESIGN AND OVERSIGHT

We performed a multicenter clinical study of AI CADx polyp classification and visual inspection versus standard visual inspection alone. Study procedures were performed at three participating endoscopy centers: Baerum Hospital (Norway), King's College Hospital London (United Kingdom), and Showa University Northern Yokohama Hospital (Japan).

The institutional review board (IRB) at each of the three participating centers approved the conduct of the study. The study protocol and statistical analysis plan are available with the full text of this article at evidence.nejm.org. Patient consent was implemented at the three study sites according to local IRB practice; In Norway, only participants enrolled in the national screening program pilot were eligible for participation and written informed patient consent was included in the consent of the screening program. In Japan, the IRB approved an opt-out consent approach because of the low risk related to the study intervention (standard treatment was performed for all polyps detected). In London, all patients provided informed consent.

All co-authors agreed on publishing the article and vouch for the completeness and accuracy of the data and the adherence to the protocol.

PATIENTS

Eligible patients were individuals 18 years of age or older who were scheduled for colonoscopy for colorectal cancer screening, polyp surveillance, or evaluation of clinical signs or symptoms at the participating centers between May 2019 and May 2021. Exclusion criteria were inflammatory bowel disease, polyposis syndrome (familial adenomatous polyposis, serrated polyposis), history of or current chemotherapy or radiation for rectosigmoid tumors, inability to undergo polypectomy (e.g., anticoagulants, comorbidities), pregnancy, and referral for removal of polyps with known histology.

All patients with small polyps (≤5 mm in diameter) in the sigmoid colon or the rectum (jointly called rectosigmoid colon) detected during colonoscopy were included in this study. For patients with more than five eligible polyps, the first five polyps were included and evaluated according to the study interventions (described next).

COLONOSCOPY PROCEDURES

All colonoscopies were performed according to routine standards at the participating centers, including preprocedure assessment, bowel preparation, sedation practices, and postprocedure recovery and care.

The following information was assessed and was registered in the study database immediately during and after each procedure: indication for colonoscopy, quality of bowel preparation assessed by the Boston Bowel Preparation Score (a 9-point assessment scale for cleaning quality during colonoscopy, with higher numbers indicating better preparation)[9]; most proximal segment of the colon reached during colonoscopy; insertion and withdrawal duration; and size, shape, and location of all detected polyps. All detected polyps were removed for histologic assessment for final diagnosis. By study design, study colonoscopists were nonexperts, defined as having between 1 and 5 years of colonoscopy experience or having independently performed between 200 and 1000 procedures before joining the study as an endoscopist. This aspect of the study design was included because we wanted to determine whether CADx improved the performance of reasonably trained, but nonexpert, endoscopists and thus shortened the learning curve in endoscopy training so the study colonoscopists behaved like experts. The study endoscopists were accredited for standard colonoscopy in the participating countries, but they did not have additional training in optical polyp diagnosis before the study. For the purpose of this study, study endoscopists received training on handling the study colonoscopes and devices and image interpretation. Novice endoscopists were not included because they are unlikely to make optical diagnoses independently from supervisors in clinical practice.

EQUIPMENT

The study centers were provided with high-resolution magnification colonoscopes (CF-H290ECI; Olympus Corp., Tokyo, Japan). These appear to be standard instruments by design, feel, and function, including narrow band imaging. In addition, the study colonoscope featured a light microscopy system integrated into the distal tip of the colonoscope. The extra feature provided 520-fold magnification at a focusing depth of 35 μm, and a field of view of 570 × 500 μm, for high-resolution magnified images on demand, which the colonoscopist controlled with a hand-operated lever.[6] This feature enabled real-time, in vivo evaluation of polyp microvascular morphology.

AI SYSTEM

The study centers were also provided with a real-time polyp classification CADx device (EndoBRAIN; Cybernet Systems Corp., Tokyo, Japan), connected to a standard colonoscopy processor unit (EVIS LUCERA ELITE, CV-290; Olympus Corp.). As noted earlier, the CADx system provides an automated diagnosis of rectosigmoid polyps by analyzing images obtained in the magnification mode of the colonoscopes for detected polyps, as previously described.[6-8]

Briefly, the CADx algorithm comprises three steps. The first is feature extraction, which is the analysis of textures characterized by differences in contrast for polyp vessels and lumens, quantified in 312 validated variables. Second is classification, which comprises support-vector machine classification of polyps as nonneoplastic or neoplastic on the basis of the 312 variables through machine learning. For the system training and validation, more than 35,000 polyp images were used which were collected from five Japanese endoscopy centers, as described previously.[10] Finally, in the diagnostic output step, the predicted diagnosis is displayed (Fig. 1) for the colonoscopist as “neoplastic” or “nonneoplastic” with a confidence probability for neoplasia (0 to 100%).

If the CADx diagnosis has a confidence probability of less than 70%, the system flags it as “low confidence,” on the basis of a previous preclinical study.[10] If the quality of the captured image is not appropriate for system diagnosis (e.g., artifacts caused by mucus, low image quality), the analysis is flagged as “not a good sample,” and no diagnosis is provided.

The nonneoplastic category comprises polyps with no neoplastic features, such as hyperplastic polyps, inflammatory polyps, and juvenile polyps. The neoplastic category comprises polyps with neoplastic features, such as adenomas and cancers.

POLYP HANDLING

For each detected polyp, four consecutive steps were applied. Step 1 comprised the standard endoscopic assessment. First, colonoscopists assessed the size, shape, and appearance of each detected polyp 5 mm or less in diameter in the rectosigmoid colon. Morphology was categorized

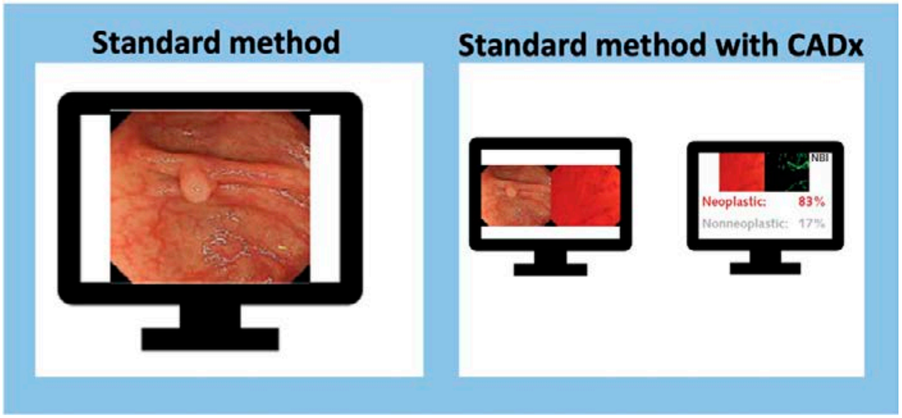


Figure 1. The Standard Method and the Combined Use of the Standard Method and the CADx System.
The Cybernet Systems EndoBRAIN system was used in this study. CADx denotes computer-aided diagnosis.

according to the Paris classification.[11] The endoscopists then classified polyps as either neoplastic (adenoma) or nonneoplastic (nonadenoma) using a binary scale (i.e., low or high confidence level in a nonneoplastic diagnosis, following recommendations in current guidelines[12-14]). Once the endoscopist registered their optical diagnosis, the CADx predicted classification was reported immediately for each polyp and registered in the study database.

Step 2 was the CADx assessment. After the standard assessment as described earlier, colonoscopists captured at least five images from each polyp using narrow band imaging and magnification mode to feed the CADx system. The CADx system then immediately provided the suggested diagnosis of the polyp as either neoplastic or nonneoplastic according to the algorithms described earlier (Fig. 1).

Step 3 was performed after standard clinical assessment and after CADx assessment, respectively. The colonoscopist again scored the confidence level of classification prediction of each polyp as either "high" or "low" and relayed it to the study nurse for immediate capture in the study database.

In step 4, all polyps were removed by snare polypectomy, biopsy forceps, or endoscopic mucosal resection and submitted for histopathologic evaluation. All polyps were evaluated by board-certified (the local board for each country of practice) gastrointestinal pathologists at each center. All pathologists were blinded to colonoscopic diagnoses of the polyps.

All polyps that were diagnosed histopathologically as nonneoplastic but had been considered by the colonoscopist as neoplastic with high confidence after standard assessment were submitted for a second histopathologic review by a different pathologist. The second pathologist was blinded to the first histopathologic diagnosis. See Supplementary Appendix, Section 2 for details.

STUDY END POINTS

The primary endpoint of the study was to compare the sensitivity of identifying small (≤5 mm in diameter) polyps in the rectosigmoid colon as adenomas during colonoscopy with the combination of standard visual inspection and the CADx system, and of standard visual inspection alone, compared with gold-standard histopathology.

Secondary outcome measures included specificity, positive predictive value (PPV), negative predictive value (NPV), rate of high-confidence optical diagnosis, and rate of rectosigmoid polyps of 5 mm or less with adequate images captured for CADx analysis.

Polyps that were not removed, those that were nonepithelial (neuroendocrine polyps, lymphoid aggregates), and those with unsuccessful image capturing were excluded from analyses.

SAMPLE SIZE CALCULATION

On the basis of a pilot study in Japan, we assumed a 6.7-percentage-point increase in sensitivity with the CADx system compared with the standard method, assuming discordance between the two methods of 14.4 percentage points (see the study protocol at evidence.nejm.org). We considered this difference to be clinically meaningful to uncover. With a statistical power of 90%, the required sample size using a two-sided 5% significance level was 345 neoplastic polyps. We estimated that we needed to enroll 767 patients on the basis of a 25% prevalence of neoplastic eligible polyps, a mean of two eligible polyps per patient, and 90% of polyps with satisfactory prediction by the CADx system. The 90% threshold was motivated by U.S. guidelines recommending an NPV of 90% or greater for optical diagnosis of small neoplastic polyps.[5]

STATISTICAL ANALYSES

Sensitivity, specificity, PPV, and NPV for the standard method and the CADx method compared with histopathology, respectively, were estimated using generalized estimating equation analyses with exchangeable correlation accounting for correlation between multiple polyps within one patient. We did not account for clustering within colonoscopist, site, or country. We calculated 95% confidence intervals (CIs) using sandwich estimates of the variance. Sensitivity and specificity of the two interventions were compared using an exact version of the McNemar test. We did not adjust for multiple comparisons. Polyps that were not removed, from which specimens were lost after removal, or that had nonepithelial histology were excluded from analyses. All tests were performed in relation to the 0.05 significance level and used R version 3.4.1 and Stata version 17 software.

In primary analyses of sensitivity and specificity, sessile serrated lesions were classified as neoplastic (similar to adenomas). In secondary analyses, sessile serrated lesions were classified as nonneoplastic (no adenomas).

No interim analysis was planned at the study start in 2019. Because of slow recruitment during the Covid-19 pandemic, the study team decided to amend the protocol and performed a blinded interim analysis in April 2020. The interim analysis applied an a priori stopping rule for futility (see details in the study protocol on evidence.nejm.org), which was not met. Thus, the study was continued until preplanned recruitment was fulfilled. Because of the blinded nature of the interim analysis, we did not adjust for it in the final analysis.

Results

PATIENTS

The median age of patients included in analyses was 67 years (interquartile range [IQR], 60 to 74), and 63.1% were men (Table 1). Of the 1242 patients who underwent study colonoscopy, 525 had 903 eligible rectosigmoid polyps that received visual inspection.

Of the 903 eligible polyps, 11 were not included in analyses. Of these, 5 were not removed, 3 were lost after removal, and 3 were nonepithelial (two neuroendocrine tumors and one leiomyoma). Consequently, 892 polyps (359 neoplastic polyps and 533 nonneoplastic polyps) from 518 patients were included in the analyses (Fig. 2). The distribution of sex and age of the participants reflects real-world clinical practice (Table S2). We did not register the race and ethnicity of participants.

Twenty-two colonoscopists, including 20 physicians and two nurse endoscopists, performed the study procedures.

COLONOSCOPY PERFORMANCE AND COMPLICATIONS

Baseline characteristics of patients and colonoscopy performance are shown in Table 2. Most colonoscopies were for colorectal cancer screening or polyp surveillance. The median colonoscopy insertion time was 12 minutes (IQR, 8 to 19), and the median withdrawal time with polyp assessments and polypectomies was 28 minutes (IQR, 20 to 40). We did not observe any complications or adverse events related to the colonoscopy or to polyp assessment or removal.

POLYP CHARACTERISTICS

The 518 eligible patients had 892 detected and removed polyps that were 5 mm or less in the rectosigmoid colon. On the basis of the histopathologic examination of the removed polyps, 359 were neoplastic. Of these, 319 were tubular adenomas with low-grade dysplasia, 2 were tubular adenomas with high-grade dysplasia, 9 were tubulovillous adenomas with low-grade dysplasia, and 3 were tubulovillous adenomas with high-grade dysplasia. Of the 26 remaining polyps that were categorized as neoplastic, 7 were traditional serrated adenomas with low-grade dysplasia and 19 were sessile serrated lesions without dysplasia. On the basis of histopathologic examination, 533 polyps were found to be nonneoplastic. Of these, 485 were hyperplastic polyps, 8 were inflammatory polyps, and 40 had other nonneoplastic histology.

PERFORMANCE OF OPTICAL DIAGNOSIS

In primary analyses, the sensitivity for neoplastic polyps was 88.4% (95% CI, 84.3 to 91.5) with the standard

Table 1. Baseline Characteristics of 518 Included Patients and Their Colonoscopies.*

Characteristic	Value
Median age — yr	67 (60 to 74)
Sex	
Men	327 (63.1)
Women	191 (36.9)
Colonoscopy Indication	
Screening colonoscopy (primary screening or fecal testing)	266 (51.4)
Polyp surveillance colonoscopy	161 (31.1)
Clinical signs or symptoms	67 (12.9)
Therapy of large polyps	23 (4.4)
Other	1 (0.2)
Median insertion time — min	12 (8 to 19)
Median withdrawal time — min	28 (20 to 40)
Preparation quality good or very good†	481 (92.9)

* Data are presented as the median (interquartile range) or no. (%).
† The Boston Bowel Preparation Scale is a 9-point assessment scale for cleaning quality during colonoscopy. The colon is divided into three segments: proximal, transverse, and distal. Each segment is classified from 0 to 3 depending on the degree of soiling. The sum total of the three segments represents the degree of soiling (≤5 points indicates poor bowel preparation; 6–7 good bowel preparation, and ≥8 very good bowel preparation).[9]

method and 90.4% (95% CI, 86.8 to 93.1) with the CADx method (P=0.33). The percentage of discordant pairs between the standard method and the CADx method was 7.2% (Fig. 3).

The specificity for neoplastic polyps was 83.1% (95% CI, 79.2 to 86.4) with the standard method and 85.9% (95% CI, 82.3 to 88.8) with the CADx method. The discordance between the standard method and the CADx method was 7.9%.

The percentage of polyp assessments with high confidence for categorization into neoplastic or nonneoplastic polyp increased from 74.2% (95% CI, 70.9 to 77.3) with the standard method to 92.6% (95% CI, 90.6 to 94.3) with the CADx method.

In secondary analyses classifying sessile serrated lesions as nonneoplastic, the sensitivity for neoplastic polyps was 91.2% (95% CI, 87.5 to 93.9) with the standard method and 94.1% (95% CI, 91.2 to 96.2) with the CADx method. The specificity for neoplastic polyps was 82.3% (95% CI, 78.4 to 85.6) with the standard method and 85.5% (95% CI, 81.9 to 88.5) with the CADx method. For separate center analyses, see Tables S3 through S8.

Discussion

Implementation of AI in cancer screening and clinical diagnosis requires proof of benefits from high-quality clinical studies. Our international multicenter study assessed the incremental gain of a specific CADx AI system for real-time polyp assessment during colonoscopy. Our study indicates that real-time AI with CADx may not significantly increase the sensitivity for small neoplastic polyps. However, CADx may improve specificity for optical diagnosis of small neoplastic polyps and increase colonoscopist confidence with visual diagnosis of polyps.

AI polyp detection tools (so-called computer-aided polyp detection) during colonoscopy could potentially increase detection of small polyps by up to 50%.[15] While this potentially could increase screening benefit, it also increases health care costs, risk of overtreatment, and patient burden.[16] Most additionally detected polyps are small ones in the distal colon and the rectum, and many of these are nonneoplastic; that is, they do not need to be removed if reliable, real-time classification were possible. One may further argue that removal of small polyps contributes little in terms of cancer prevention.[17]

The "diagnose-and-leave" strategy recently proposed by the American Society for Gastrointestinal Endoscopy (ASGE) suggests not to remove small polyps during colonoscopy if they can be reliably classified (defined as NPV of ≥90%) by optical diagnosis as nonneoplastic.[5] This strategy is not easy to apply because such reliable diagnosis is difficult to achieve with standard colonoscopy systems. Our study provides high-quality data to address this critical issue.

Our main outcome did not reach the prespecified increase of 6.7% in sensitivity with CADx, which was based on preclinical testing, observational studies, and a single-center study. Our study thus emphasizes the importance of rigorous clinical studies to assess AI performance and quantifies the added value and the limitation of CADx in colonoscopy.

According to our results, CADx may not reduce overlooking adenomas during visual inspection of polyps. However, our study showed a potential improvement in specificity for neoplastic polyps, albeit one in which we cannot declare statistical significance because our primary outcome failed to reach that level with the CADx system. There was also a trend toward improved confidence in

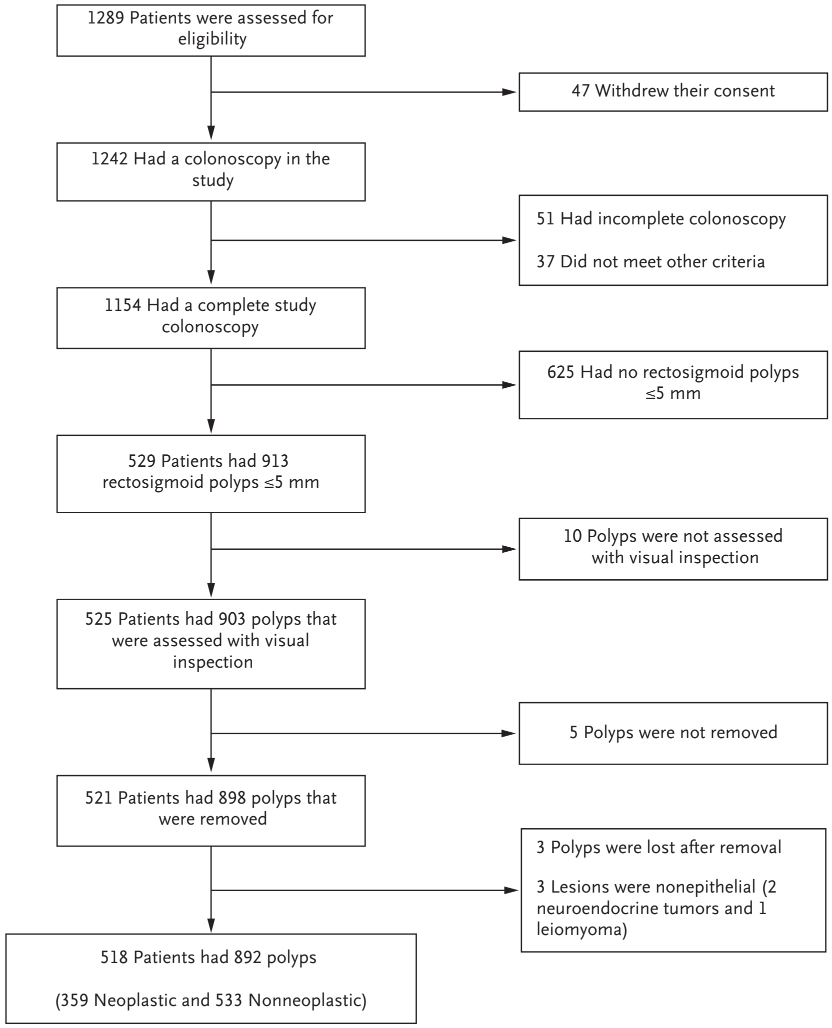


Figure 2. Study Flow Chart.

optical diagnosis of polyps. If this can be established through additional clinical trials, it could potentially contribute to a clinically important reduction in the unnecessary removal of small nonneoplastic polyps by giving the operator the ability to make a high-confidence prediction during a procedure.[5]

PPVs and NPVs are influenced by the prevalence of disease (polyps) and do not adequately assess tools or devices as such. Therefore, our primary outcomes of interest were sensitivity and specificity. However, we also analyzed the predictive values of CADx and observed increments of 1.3% for NPV and 3.1% for PPV with CADx (Table 3). Our results are consistent with the hypothesis that CADx can fulfill the criteria for the diagnose-and-leave strategy with 95% CIs above the NPV threshold of 90%.

The strengths of the current study are the comparison with both non-AI optical diagnosis and gold-standard histopathology for all included polyps; the inclusion of centers from different countries and continents; and the focus on endoscopists with average experience and workload, mimicking real-world colonoscopy practice. A limitation of this study is the inability of the CADx tool to identify sessile serrated polyps, a recently recognized polyp type with likely neoplastic potential. To alleviate this challenge, we conducted two analyses (one classifying sessile serrated polyps as neoplastic and the other classifying them as nonneoplastic) without significant differences in the performance of the CADx tool. Another limitation is the learning curve of the colonoscopists during the study period due to the prospective study design, which may contribute to underestimation of the CADx performance. However, we

Table 2. Characteristics of the 892 Small Polyps (≤5 mm in diameter) in the Distal Colon and the Rectum.*

Characteristic	Neoplastic Polyps (n=359)	Nonneoplastic Polyps (n=533)
Median size — mm	4 (3 to 5)	3 (2 to 3)
Location		
Sigmoid colon	274 (76.3)	260 (48.8)
Rectum	85 (23.7)	273 (51.2)
Morphology†		
Polypoid (type Is or Ip)	175 (48.7)	109 (20.5)
Nonpolypoid (type IIa)	184 (51.3)	424 (79.5)
Removal method		
Snare polypectomy	247 (68.8)	265 (49.7)
Forceps	65 (18.1)	258 (48.4)
Endoscopic mucosal resection	46 (12.8)	10 (1.9)

* Data are presented as the median (interquartile range) or no. (%). Sessile serrated lesions were classified as neoplastic polyps in the primary analysis.

† The Paris classification was used. Morphologic classification systems for polyps during colonoscopy classify polyps into polypoid and nonpolypoid, with six different subtypes.[12]

may also have overestimated nonexpert endoscopists' performance because the sensitivity we found to predict adenomas, without the aid of CADx, was 88.4%, which is slightly higher than that reported in previous studies.[18,19] This may be related to the fact that our study was conducted at teaching hospitals with endoscopy training programs.

Finally, the colonoscopes used in the current study are not widely used today, although they are commercially available in Europe, the Middle East, and Asia. Provided that colonoscopes with surface enhancement functions facilitating CADx systems like the one we tested prove to be useful, they would likely become used more widely.

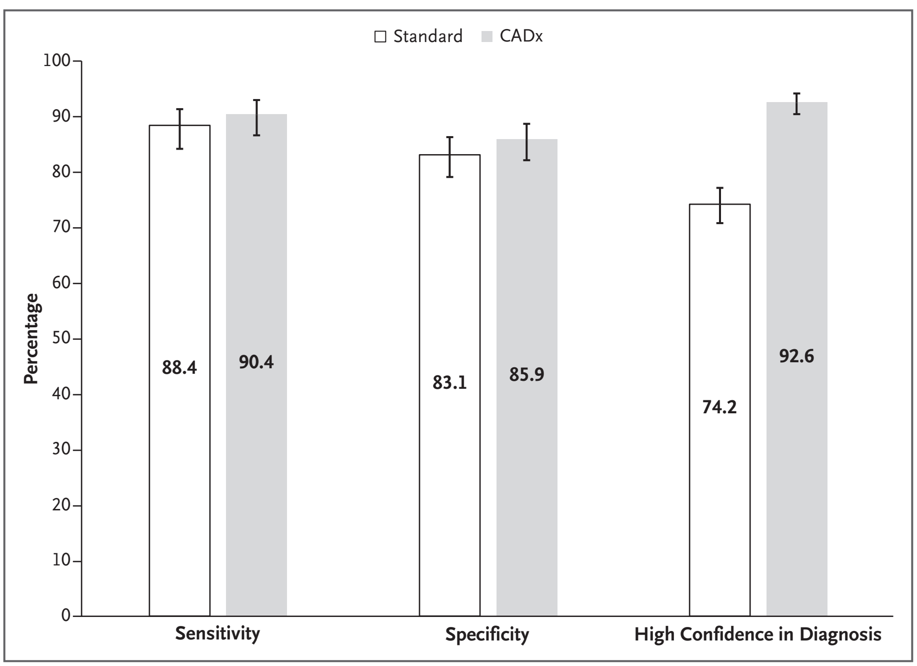


Figure 3. Sensitivity, Specificity, and Confidence of Diagnosis of Standard and AI-Derived CADx Optical Diagnosis of Small Rectosigmoid Polyps during Colonoscopy Compared with Histopathology.

All bars are represented with corresponding 95% confidence intervals. AI denotes artificial intelligence and CADx computer-aided diagnosis.

Table 3. Performance of Standard and AI-Derived CADx Optical Diagnosis of Small Rectosigmoid Polyps during Colonoscopy Compared with Histopathology.*

Parameter	Standard Diagnosis	CADx Diagnosis
Sensitivity	88.4 (84.3 to 91.5)	90.4 (86.8 to 93.1)
Specificity	83.1 (79.2 to 86.4)	85.9 (82.3 to 88.8)
Positive predictive value	78.9 (74.3 to 82.9)	82.0 (77.6 to 85.6)
Negative predictive value	91.5 (88.5 to 93.8)	92.8 (90.1 to 94.9)
High confidence in optical diagnosis	74.2 (70.9 to 77.3)	92.6 (90.6 to 94.3)

* Sessile serrated lesions were classified as neoplastic polyps according to the primary analysis plan. Values are presented as percentages (95% confidence intervals). AI denotes artificial intelligence and CADx denotes computer-aided diagnosis.

Our study suggests that the use of CADx helped the provider have higher confidence in optical diagnosis. If this can be replicated, it could contribute to cost reduction because more polyps could be left in situ. Better confidence comes at a cost; CADx assessment prolongs colonoscopy procedure time, which increases health care cost. In previous studies, we demonstrated that the time necessary for CADx assessment of one small polyp, as applied in this study, is about 40 seconds.[6] We consider this additional time well spent with regard to the gain in terms of reduction of unnecessary removal of polyps and histopathologic assessment. Future cost-effectiveness studies may explore whether the prolonged procedure time pays off with the benefit of reduced polypectomies.

In conclusion, real-time assessment with CADx did not significantly increase sensitivity for neoplastic polyps during colonoscopy. There are promising signals for increased specificity and improved confidence of optical diagnosis, but our statistical approach precludes us from making any definitive statements about the identification and removal of small rectosigmoid polyps using the colonoscopy system we employed.

Disclosures

Author disclosures and other supplementary materials are available at evidence.nejm.org.

This study was investigator initiated and funded by public research grants from the Research Council of Norway (Norges Forskningsråd), the Norwegian Cancer Society (Kreftforeningen), and the Japan Society for the Promotion of Science. Cybernet Systems and Olympus loaned the AI software and endoscopes, respectively, to the study at no cost. The companies and funders did not have any influence on and did not play any role in the design, conduct, analysis of the study, or interpretation of the results and were not involved in the writing of the manuscript. Olympus provided a research grant for the study.

Author Affiliations

[1] Clinical Effectiveness Research Group, University of Oslo, Oslo

[2] Clinical Effectiveness Research Group, Department of Transplantation Medicine, Oslo University Hospital, Oslo

[3] Department of Gastroenterology, Hepatology and Clinical Oncology, Centre of Postgraduate Medical Education, Warsaw, Poland

[4] Digestive Disease Center, Showa University Northern Yokohama Hospital, Yokohama, Japan

[5] Department of Medicine, Sørlandet Hospital Kristiansand, Kristiansand, Norway

[6] King's Institute of Therapeutic Endoscopy, King's College Hospital NHS Foundation Trust, London

[7] Graduate School of Informatics, Nagoya University, Nagoya, Japan

[8] Department of Diagnostic Pathology, School of Medicine, Showa University Northern Yokohama Hospital, Kanagawa, Japan

[9] Department of Medicine, Baerum Hospital, Vestre Viken Hospital Trust, Gjettum, Norway

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209-249. DOI: 10.3322/CAAC.21660.

2. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 2012;366:687-696. DOI: 10.1056/NEJMoa1100370.

3. Abu Dayyeh BK, Thosani N, Konda V, et al. ASGE Technology Committee systematic review and meta-analysis assessing the ASGE PIVI thresholds for adopting real-time endoscopic assessment of the histology of diminutive colorectal polyps. Gastrointest Endosc 2015;81:502.e1-502.e16. DOI: 10.1016/j.gie.2014.12.022.

4. Wadhwa V, Alagappan M, Gonzalez A, et al. Physician sentiment toward artificial intelligence (AI) in colonoscopic practice: a survey of US gastroenterologists. Endosc Int Open 2020;8:E1379-E1384. DOI: 10.1055/a-1223-1926.

5. Rex DK, Kahi C, O'Brien M, et al. The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps. Gastrointest Endosc 2011;73:419-422. DOI: 10.1016/J.GIE.2011.01.023.

6. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018;169:357-366. DOI: 10.7326/M18-0249.

7. Misawa M, Kudo SE, Mori Y, et al. Characterization of colorectal lesions using a computer-aided diagnostic system for narrow-band imaging endocytoscopy. Gastroenterology 2016;150:1531-1532.e3. DOI: 10.1053/j.gastro.2016.04.004.

8. Mori Y, Kudo SE, Chiu PW, et al. Impact of an automated system for endocytoscopic diagnosis of small colorectal lesions: an international web-based study. Endoscopy 2016;48:1110-1118. DOI: 10.1055/s-0042-113609.

9. Lai EJ, Calderwood AH, Doros G, Fix OK, Jacobson BC. The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. Gastrointest Endosc 2009;69:620-625. DOI: 10.1016/J.GIE.2008.05.057.

10. Kudo SE, Misawa M, Mori Y, et al. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. Clin Gastroenterol Hepatol 2020;18:1874-1881. DOI: 10.1016/j.cgh.2019.09.009.

11. Participants in the Paris Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointest Endosc 2003;58(Suppl):S3-S4. DOI: 10.1016/s0016-5107(03)02159-x.

12. Hewett DG, Kaltenbach T, Sano Y, et al. Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. Gastroenterology 2012;143:599-607.e1. DOI: 10.1053/J.GASTRO.2012.05.006.

13. IJspeert JEG, Bastiaansen BAJ, van Leerdam ME, et al. Development and validation of the WASP classification system for optical diagnosis of adenomas, hyperplastic polyps and sessile serrated adenomas/polyps. Gut 2016;65:963-970. DOI: 10.1136/GUTJNL-2014-308411.

14. Sumimoto K, Tanaka S, Shigita K, et al. Diagnostic performance of Japan NBI Expert Team classification for differentiation among noninvasive, superficially invasive, and deeply invasive colorectal neoplasia. Gastrointest Endosc 2017;86:700-709. DOI: 10.1016/j.gie.2017.02.018.

15. Barua I, Vinsard DG, Jodal HC, et al. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. Endoscopy 2021;53:277-284. DOI: 10.1055/a-1201-7165.

16. Hassan C, Pickhardt PJ, Rex DK. A resect and discard strategy would improve cost-effectiveness of colorectal cancer screening. Clin Gastroenterol Hepatol 2010;8:865-869, 869.e1-869.e3. DOI: 10.1016/J.CGH.2010.05.018.

17. Wieszczy P, Kaminski MF, Løberg M, et al. Estimation of overdiagnosis in colorectal cancer screening with sigmoidoscopy and faecal occult blood testing: comparison of simulation models. BMJ Open 2021;11:e042158. DOI: 10.1136/bmjopen-2020-042158.

18. Rees CJ, Rajasekhar PT, Wilson A, et al. Narrow band imaging optical diagnosis of small colorectal polyps in routine clinical practice: the Detect Inspect Characterise Resect and Discard 2 (DISCARD 2) study. Gut 2017;66:887-895. DOI: 10.1136/gutjnl-2015-310584.

19. Ladabaum U, Fioritto A, Mitani A, et al. Real-time optical biopsy of colon polyps with narrow band imaging in community practice does not yet meet key thresholds for clinical decisions. Gastroenterology 2013;144:81-91. DOI: 10.1053/j.gastro.2012.09.054.