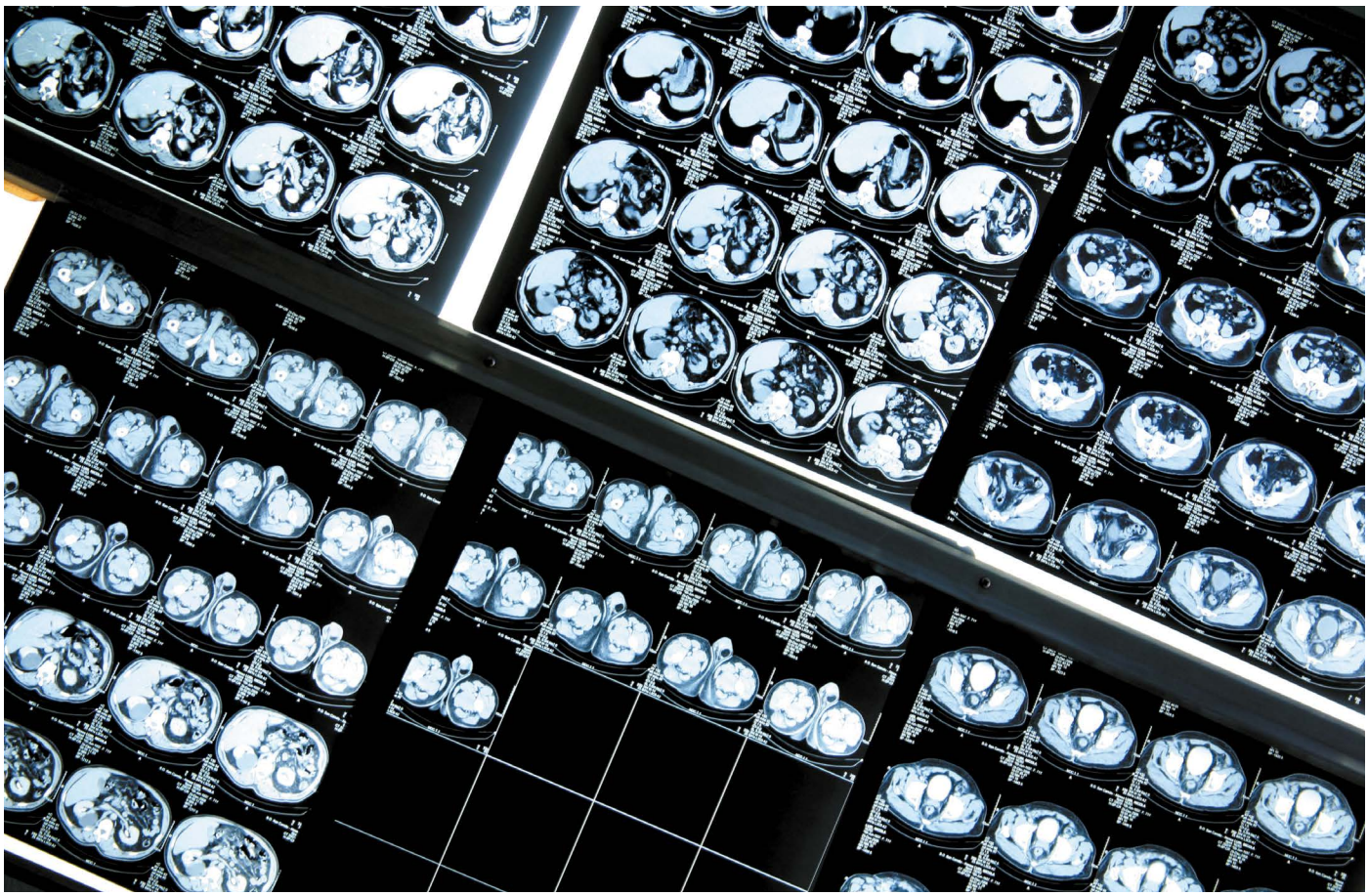


MASSIMO BREGA/SPL

Researchers are feeding machine-learning tools millions of medical scans to give them general diagnostic capabilities.

AN AI REVOLUTION IS BREWING IN MEDICINE. WHAT WILL IT LOOK LIKE?

Emerging generalist models could overcome some limitations of first-generation machine-learning tools for clinical use. **By Mariana Lenharo**

Jordan Perchik started his radiology residency at the University of Alabama at Birmingham near the peak of what he calls the field's "AI scare". It was 2018, just two years after computer scientist Geoffrey Hinton had proclaimed that people should stop training to be radiologists because machine-learning tools would soon displace them. Hinton, sometimes referred to as the godfather of artificial intelligence (AI), predicted that these systems would soon be able to read and interpret medical scans and X-rays better than people could. A substantial drop in applications for radiology programmes followed. "People were worried that they were going to finish residency and just wouldn't have a job," Perchik says.

Hinton had a point. AI-based tools are increasingly part of medical care; more than 500 have been authorized by the US Food and Drug Administration (FDA) for use in medicine. Most are related to medical imaging — used for enhancing images, measuring abnormalities or flagging test results for follow-up.

But even seven years after Hinton's prediction, radiologists are still very much in demand. And clinicians, for the most part, seem underwhelmed by the performance of these technologies.

Surveys show that although many

physicians are aware of clinical AI tools, only a small proportion – between 10% and 30% – has actually used them[1]. Attitudes range from cautious optimism to an outright lack of trust. "Some radiologists doubt the quality and safety of AI applications," says Charisma Hehakaya, a specialist in the implementation of medical innovations at University Medical Center Utrecht in the Netherlands. She was part of a team that interviewed two dozen clinicians and hospital managers in the Netherlands for their views on AI tools in 2019 (ref. 2). Because of that doubt, she says, the latest approaches sometimes get abandoned.

And even when AI tools accomplish what they're designed to do, it's still not clear whether this translates into better care for patients. "That would require a more robust analysis," Perchik says.

But excitement does seem to be growing about an approach sometimes called generalist medical AI. These are models trained on massive data sets, much like the models that power ChatGPT and other AI chatbots. After ingesting large quantities of medical images and text, the models can be adapted for many tasks. Whereas currently approved tools serve specific functions, such as detecting lung nodules in a computed tomography (CT) chest scan, these generalist models would act more like a physician, assessing every anomaly in the scan and assimilating it into something like a diagnosis.

Although AI enthusiasts now tend to steer clear of bold claims about machines replacing doctors, many say that these models could overcome some of the current limitations of medical AI, and they could one day surpass physicians in certain scenarios. "The real goal to me is for AI to help us do the things that humans aren't very good at," says radiologist Bibb Allen, chief medical officer at the American College of Radiology Data Science Institute, who is based in Birmingham, Alabama.

But there's a long journey ahead before these latest tools can be used for clinical care in the real world.

SOURCE: REF. 4

Current limitations

AI tools for medicine serve a support role for practitioners, for example by going through scans rapidly and flagging potential issues that a physician might want to look at right away. Such tools sometimes work beautifully. Perchik remembers the time an AI triage flagged a chest CT scan for someone who was experiencing shortness of breath. It was 3 a.m. – the middle of an overnight shift. He prioritized the scan and agreed with the AI assessment that it showed a pulmonary embolism, a potentially fatal condition that requires immediate treatment. Had it not been flagged, the scan might not have been evaluated until later that day.

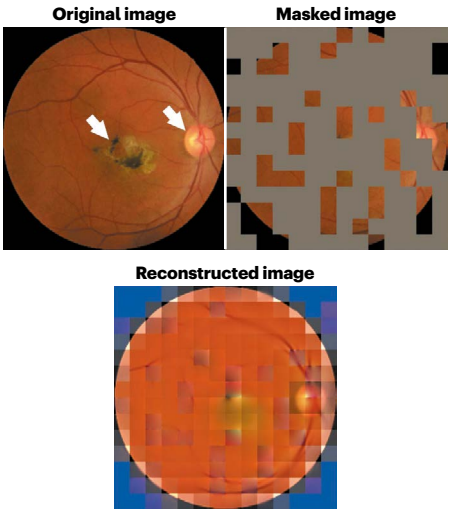
But if the AI makes a mistake, it can have the opposite effect. Perchik says he recently spotted a case of pulmonary embolism that the AI had failed to flag. He decided to take extra review steps, which confirmed his assessment but slowed down his work. "If I had decided to trust the AI and just move forward, that could have gone undiagnosed."

Many devices that have been approved

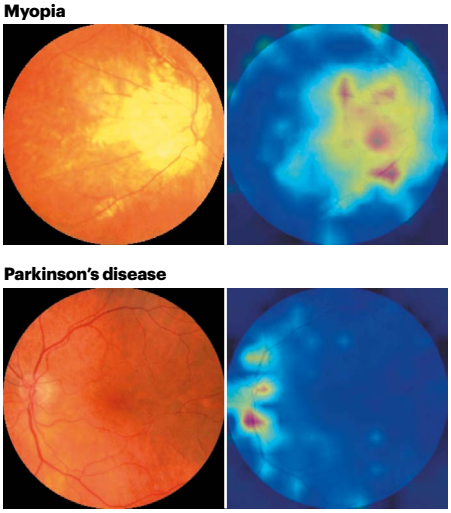
"If I had decided to trust the AI and just move forward, that could have gone undiagnosed."

EYE DIAGNOSTICS

A foundation model was trained on more than one million images of human retinas. When given masked data in which 75% of the image was covered, the model was able to reconstruct the image, accurately filling in details of anatomical structures such as the optic nerve.



After training the model to classify conditions such as myopia and Parkinson's disease, researchers were able to visualize the areas of the images that were most important for making a diagnosis. These matched the pathologies that physicians look for.



Heatmaps show the areas of the image that contribute most to diagnosis.

don't necessarily line up with the needs of physicians, says radiologist Curtis Langlotz, director of Stanford University's Center for Artificial Intelligence in Medicine and Imaging in Palo Alto, California. Early AI medical tools were developed according to the availability of imaging data, so some applications have been built for things that are common and easily spotted. "I don't need help detecting pneumonia" or a bone fracture, Langlotz says. Even so, multiple tools are available for assisting physicians with these diagnoses.

Another issue is that the tools tend to focus on specific tasks rather than interpreting a medical examination comprehensively – observing everything that might be relevant in an image, taking into account previous results and the person's clinical history. "Although focusing on detecting a few diseases has some value, it doesn't reflect the true cognitive work of the radiologist," says Pranav Rajpurkar, a computer scientist who works on biomedical AI at Harvard Medical School in Boston, Massachusetts.

The solution has often been to add more AI-powered tools, but that creates challenges for medical care, too, says Alan Karthikesalingam, a clinical research scientist at Google Health in London. Consider a person having a routine mammography. The technicians might be assisted by an AI tool for breast cancer screening. If an abnormality is found, the same person might require a magnetic resonance imaging (MRI) scan to confirm the diagnosis, for which there could be a separate AI device. If the diagnosis is confirmed, the lesion would be removed surgically, and there might be yet another AI system to assist with the pathology.

"If you scale that to the level of a health system, you can start to see how there's a plethora of choices to make about the devices themselves and a plethora of decisions on how to integrate them, purchase them, monitor them, deploy them," he says. "It can quickly become a kind of IT soup."

Many hospitals are unaware of the challenges involved in monitoring AI performance and safety, says Xiaoxuan Liu, a clinical researcher who studies responsible innovation in health AI at the University of Birmingham, UK. She and her colleagues identified thousands of medical-imaging studies that compared the diagnostic performance of deep-learning models with that of healthcare professionals[3]. For the 69 studies the team assessed for diagnostic accuracy, a main finding was that a majority of models weren't tested using a data set that was truly independent of the information used to train the model. This means that these studies might have overestimated the models' performance.

"It's becoming now better known in the field that you have to do an external validation," Liu says. But, she adds, "there's only a handful of institutions in the world that are very aware of

this". Without testing the performance of the model, particularly in the setting in which it will be used, it is not possible to know whether these tools are actually helping.

Solid foundations

Aiming to address some of the limitations of AI tools in medicine, researchers have been exploring medical AI with broader capabilities. They've been inspired by revolutionary large language models such as the ones that underlie ChatGPT.

These are examples of what some scientists call a foundation model. The term, coined in 2021 by scientists at Stanford University, describes models trained on broad data sets — which can include images, text and other data — using a method called self-supervised learning. Also called base models or pre-trained models, they form a basis that can later be adapted to perform different tasks.

Most medical AI devices already in use by hospitals were developed using supervised learning. Training a model with this method to identify pneumonia, for example, requires specialists to analyse numerous chest X-rays and label them as 'pneumonia' or 'not pneumonia', to teach the system to recognize patterns associated with the disease.

The annotation of large numbers of images, an expensive and time-consuming process, is not required in foundation models. For ChatGPT, for example, vast collections of text were used to train a language model that learns by predicting the next word in a sentence. Similarly, a medical foundation model developed by Pearse Keane, an ophthalmologist at Moorfields Eye Hospital in London, and his colleagues used 1.6 million retinal photos and scans to learn how to predict what missing portions of the images should look like[4] (see 'Eye diagnostics'). After the model had learnt all the features of a retina during this pre-training, the researchers introduced a few hundred labelled images that allowed it to learn about specific sight-related conditions, such as diabetic retinopathy and glaucoma. The system was better than previous models at detecting these ocular diseases, and at predicting systemic diseases that can be detected through tiny changes in the blood vessels of the eye, such as heart disease and Parkinson's. The model hasn't yet been tested in a clinical setting.

Keane says that foundation models could be especially suitable for ophthalmology, because almost every part of the eye can be imaged at high resolution. And huge data sets of these images are available to train such models. "AI is going to transform health care," he says. "And ophthalmology can be an example for other medical specialities."

Foundation models are "a very flexible framework", says Karthikesalingam, adding that their characteristics seem to be well suited to addressing some of the limitations of first-generation medical AI tools.

Big tech companies are already investing in medical-imaging foundation models that use multiple image types — including skin photographs, retinal scans, X-rays and pathology slides — and incorporate electronic health records and genomics data.

In June, scientists at Google Research in Mountain View, California, published a paper describing an approach they call REMEDIS ('robust and efficient medical imaging with self-supervision'), which was able to improve diagnostic accuracies by up to 11.5% compared with AI tools trained using supervised learning[5]. The study found that, after pre-training a model on large data sets of unlabelled images, only a small number of labelled images were needed to achieve those results. "Our key insight was that REMEDIS was able to, in a really efficient way, with very few examples, learn how to classify lots of different things in lots of different medical images," including chest X-rays, digital pathology scans and mammograms, says Karthikesalingam, who is a co-author of the paper.

"You can basically start to have conversations with images just like you are talking with ChatGPT."

The following month, Google researchers described in a preprint[6] how they had brought that approach together with the firm's medical large language model Med-PaLM, which can answer some open-ended medical queries almost as well as a physician. The result is Med-PaLM Multimodal, a single AI system that demonstrated that it could not only interpret chest X-ray images, for example, but also draft a medical report in natural language[6].

Microsoft is also working to integrate language and vision into a single medical AI tool. In June, scientists at the company introduced LLaVA-Med (Large Language and Vision Assistant for biomedicine), which was trained on images paired with text extracted from PubMed Central, a database of publicly accessible biomedical articles[7]. "Once you do that, then you can basically start to have conversations with images just like you are talking with ChatGPT," says computer scientist Hoifung Poon, who leads biomedical AI research at Microsoft Health Futures and is based in Redmond, Washington. One of the challenges of this approach is that it requires huge numbers of text–image pairs. Poon says he and his colleagues have now collected more than 46 million pairs from PubMed Central.

As these models are trained on ever more data, some scientists are optimistic that they might be able to identify patterns that humans cannot. Keane mentions a 2018 study by Google researchers that described AI models capable of identifying a person's characteristics — such as age and gender — from retinal images[8]. That is something that even experienced ophthalmologists can't do, Keane says. "So, there's a real hope that there's a lot of scientific information embedded within these high-dimensional images."

One example of where AI tools could surpass human abilities, according to Poon, is the use of digital pathology to predict tumoral responses to immunotherapy. It is thought that the tumour microenvironment — the milieu of cancerous, non-cancerous and immune cells that can be sampled using a biopsy — influences whether an individual will respond well to various anti-cancer drugs. "If you can see millions and millions of patients that have already taken a checkpoint inhibitor or other immunotherapy, and you look at the exceptional responders and the non-responders, you could start to actually discern a lot of these patterns that an expert may not be able to see," says Poon.

He cautions that, although there's a lot of excitement around the diagnostic potential of AI devices, these tools also have a high bar for success. Other medical uses for AI, such as matching participants to clinical trials, are likely to have a more immediate impact.

Karthikesalingam also notes that even the best results achieved by Google's medical imaging AI are still no match for humans. "An X-ray report by a human radiologist is still considered significantly superior to a state-of-the-art multimodal generalist medical system," he says. Although foundation models seem to be particularly well poised to broaden the applications of medical AI tools, there is a long way to go to demonstrate that they can safely be used in clinical care, Karthikesalingam adds. "While we want to be bold, we also think it's very important to be responsible."

Perchik has no doubt that the role of AI will continue to grow in his field of radiology, but rather than replacing radiologists, he thinks people will need to be trained to use AI. In 2020, he organized a free AI literacy course for radiologists that has since expanded to 25 programmes across the United States. "A lot of the work that we do is demystifying AI and managing the hype versus what the reality of AI is," he says.

Mariana Lenharo is a reporter for *Nature* in New York City.

1. Chen, M. *et al. Front. Med.* **9**, 990604 (2022).
2. Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P. C. & Moors, E. H. M. *Eur. Radiol.* **30**, 5525–5532 (2020).
3. Liu, X. *et al. Lancet Digit. Health* **1**, E271–E297 (2019).
4. Zhou, Y. *et al. Nature* **622**, 156–163 (2023).
5. Azizi, S. *et al. Nature Biomed. Eng.* **7**, 756–779 (2023).
6. Tu, T. *et al.* Preprint at https://arxiv.org/abs/2307.14334 (2023).
7. Li, C. *et al.* Preprint at https://arxiv.org/abs/2306.00890 (2023).
8. Poplin, R. *et al. Nature Biomed. Eng.* **2**, 158–164 (2018).