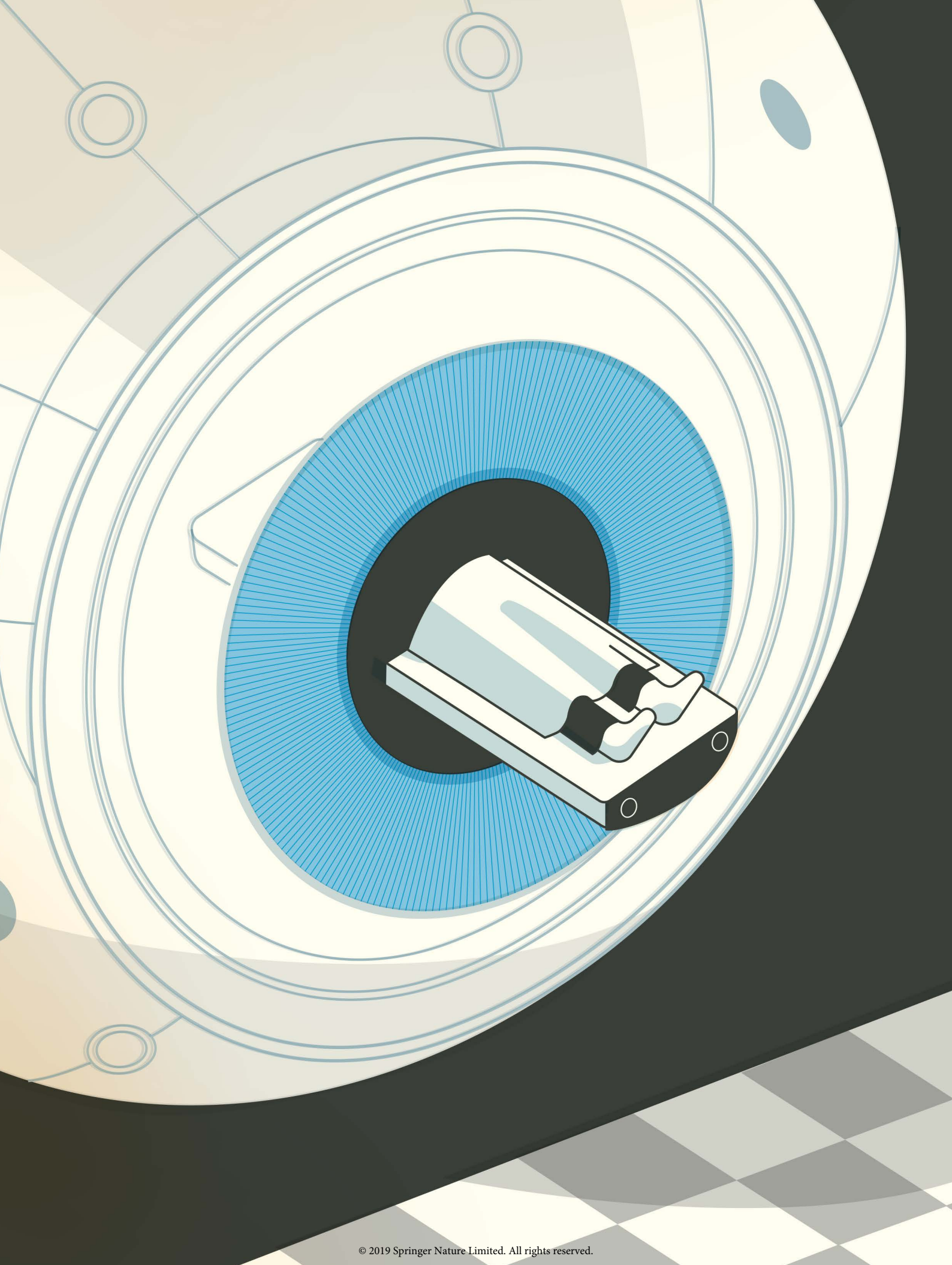

Rise of Robot Radiologists

Deep-learning algorithms are peering into MRIs and x-rays with unmatched vision, but who is to blame when they make a mistake?

By Sara Reardon

WHEN REGINA BARZILAY had a routine mammogram in her early 40s, the image showed a complex array of white splotches in her breast tissue. The marks could be normal, or they could be cancerous—even the best radiologists often struggle to tell the difference. Her doctors decided the spots were not immediately worrisome. In hindsight, she says, "I already had cancer, and they didn't see it."

Over the next two years Barzilay underwent a second mammogram, a breast MRI and a biopsy, all of which continued to yield ambiguous or conflicting findings. Ultimately she was diagnosed with breast cancer in 2014, but the path to that diagnosis had been unbelievably frustrating. "How do you do three tests and get three different results?" she wondered.

Barzilay was treated and made a good recovery. But she remained horrified that the uncertainties of reading a mammogram could delay treatment. "I realized to what extent we are unprotected with current approaches," she says, so she made a career-altering decision: "I absolutely have to change it."

A computer scientist at Massachusetts Institute of Technology, Barzilay had never studied health before. Her research used machine-learning techniques—a form of artificial intelligence—for natural-language processing. But she had been looking for a new line of research and decided to team up with radiologists to develop machine-learning algorithms that use computers' superior visual analysis to spot subtle patterns in mammograms that the human eye might miss.

Over the next four years the team taught a computer program to analyze mammograms from about 32,000 women of different ages and races and told it which women had been diagnosed with cancer within five years of the scan. They then tested the computer's matching abilities in 3,800 more patients. Their resulting algorithm, published last May in *Radiology,* was significantly more accurate at predicting cancer—or the absence of cancer—than practices generally used in clinics. When Barzilay's team ran the program on her own mammograms from 2012—ones her doctor had cleared—the algorithm correctly predicted she was at a higher risk of developing breast cancer within five years than 98 percent of patients.

AI algorithms not only spot details too subtle for the human eye to see. They can also develop entirely new ways of interpreting medical images, sometimes in ways humans do not understand. The numerous researchers, start-up companies and scanner manufacturers designing AI programs hope they can improve the accuracy and timeliness of diagnoses, provide better treatment in developing countries and remote regions that lack radiologists, reveal new links between biology and disease, and even help to predict how soon a person will die.

AI applications are entering clinics at a rapid rate, and physicians have met the technology with equal parts excitement about its potential to reduce their workload and fear about losing their jobs to machines. Algorithms also raise unprecedented questions about how to regulate a machine that is constantly learning and changing and who is to blame if an algorithm gets a diagnosis wrong. Still, many physicians are excited about the promise of AI programs. "If these models can be sufficiently vetted and we can raise our level of understanding of how they work, this can help raise the level of health care for everybody," says Matthew Lungren, a radiologist at Stanford University.

"A VERY, VERY HOT TOPIC"

THE IDEA OF using computers to read radiological scans is not new. In the 1990s radiologists started using a program called computer-assisted diagnosis (CAD) to detect breast cancer in mammograms. The technology was hailed as revolutionary, and clinics adopted it rapidly. But CAD proved to be more time-consuming and difficult to use than existing methods,

and according to some studies, clinics that used it made more errors than those that did not. The failure made many physicians dubious of computer-aided diagnostics, says Vijay Rao, a radiologist at Jefferson University in Philadelphia.

In the past decade, however, computer vision has improved by leaps and bounds—in everyday applications such as face recognition and in medicine. The advance has been largely driven by the development of deep-learning methods, in which a computer is given a set of images and then left to draw its own connections between them, ultimately developing a network of associations. In medical imaging, this might, for example, involve telling the computer which images contain cancer and setting it free to find features common to those images but absent in cancer-free images.

Development and adoption of AI technologies in radiology has spread rapidly. "Last year, at every large meeting I went to, the main theme was AI and imaging," says Rao, past president of the Radiological Society of North America. "Clearly, this is a very, very hot topic."

The U.S. Food and Drug Administration says that it does not keep a list of AI products that it has approved. But Eric Topol, a digital medicine researcher at the Scripps Research Institute in La Jolla, Calif., estimates that the agency is approving more than one medical imaging algorithm per month. A 2018 survey by marketing-intelligence firm Reaction Data found that 84 percent of U.S. radiology clinics had adopted or planned to adopt AI programs. The field is growing especially quickly in China, where more than 100 companies are designing AI applications for health care.

"It's a fascinating time to be in this market," says Elad Walach, CEO of the Tel Aviv–based start-up Aidoc. The company develops algorithms to analyze CT scans for abnormalities and move those patients to the top of a doctor's priority list. Aidoc also tracks how often doctors use the program and how long they spend second-guessing its conclusions. "Initially they're skeptical, but after two months they get used to it and are very trusting," Walach says.

Saving time can be crucial to saving a patient. One recent study of chest x-rays for collapsed lungs found that radiologists flag more than 60 percent of the scans they order as high priority, which suggests that they might spend hours wading through nonserious cases before getting to those that are actually urgent. "Every doctor I talk to has a story where they lost a patient because of a collapsed lung," says Karley Yoder, vice president and general manager of AI at Boston-based GE Healthcare, one of the leading manufacturers of medical imaging equipment. Last September the FDA approved a set of AI tools that will now come embedded in GE scanners, automatically flagging the most urgent cases.

Because they can process massive amounts of data, computers can perform analytical tasks that are beyond human capability. Google, for instance, is using its computing power to develop AI algorithms that construct two-dimensional CT images of lungs into a three-dimensional lung and look at the entire structure to determine whether cancer is present. Radiologists, in contrast, have to look at these images individually and attempt to reconstruct them in their heads. Another Google algorithm can do something radiologists cannot do at all: determine patients' risk of cardiovascular disease by looking at a scan of their retinas, picking up on subtle changes related to blood pressure, cholesterol, smoking history and aging. "There's potential signal there beyond what was known before," says Google product manager Daniel Tse.

84

Percentage of U.S. radiology clinics that have adopted or plan to adopt AI programs, according to a 2018 survey.

THE BLACK BOX PROBLEM

AI PROGRAMS COULD END UP revealing entirely new links between biological features and patient outcomes. A 2019 paper in *JAMA Network Open* described a deep-learning algorithm trained on more than 85,000 chest x-rays from people enrolled in two large clinical trials that had tracked them for more than 12 years. The algorithm scored each patient's risk of dying during this period. The researchers found that 53 percent of the people the AI put into a high-risk category died within 12 years, as opposed to 4 percent in the low-risk category. The algorithm did not have information on who died or on the cause of death. The lead investigator, radiologist Michael Lu of Massachusetts General Hospital, says that the algorithm could be a helpful tool for assessing patient health if combined with a physician's assessment and other data such as genetics.

To understand how the algorithm worked, the researchers identified the parts of images that it used to make its calculations. Some, such as waist circumference and the structure of women's breasts, made sense because these areas can hint at known risk factors for certain diseases. But the algorithm also looked at the region under patients' shoulder blades, which has no known medical significance. Lu guesses that flexibility might be one predictor of a shorter life span. Taking a chest x-ray often requires patients to hug the machine, and less healthy people who cannot put their arms all the way around it might position their shoulders in a different way. "They're not things I would have thought of de novo and might not understand," Lu says.

The disconnect between the way computers and humans think is known as the black box problem: the idea that a computer brain operates in an obscured space that is inaccessible to humans. Experts differ on whether this presents a problem in medical imaging. On the one hand, if an algorithm consistently improves doctors' performance and patients' health, doctors do not need to know how it works. After all, researchers still do not fully understand the mechanisms of many drugs such as lithium, which has been used to treat depression since the 1950s. "Maybe we shouldn't be so fixated, because the way humans work in medicine is about as black box as you can get," Topol says. "Do we hold machines to a higher standard?"

"AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't."
—Curtis Langlotz, Stanford University

Still, there is no denying that the black box presents ample opportunity for human-AI misunderstanding. For instance, researchers at the Icahn School of Medicine at Mount Sinai were deeply puzzled by a discrepancy in the performance of a deep-learning algorithm they had developed to identify pneumonia in lung x-rays. It performed with greater than 90 percent accuracy on x-rays produced at Mount Sinai but was far less accurate with scans from other institutions. They eventually figured out that instead of just analyzing the images, the algorithm was also factoring in the odds of a positive finding based on how common pneumonia was at each institution—not something they expected or wanted the program to do.

Confounding factors like these worry Samuel Finlayson, who studies biomedical applications of machine learning at Harvard Medical School. He notes that data sets on which AI is trained can be biased in ways that developers fail to consider. An image taken in an emergency room or one taken in the middle of the night may be more likely to show a sick person than one taken during a routine examination, for instance. An algorithm could also learn to look at scars or medical device implants that indicate a previous health problem and decide that people without these marks did not have the condition. Even the way that institutions label their images can confuse an AI algorithm and prevent the model from functioning well in another institution with a different labeling system. "If you naively train [an algorithm] at a hospital from one location, one time, and one population group, you're unaware of all the thousands of little factors that models are taking into account. If any of those change, you can be in for a world of hurt," Finlayson warns.

The solution, Finlayson says, is to train an algorithm with data from many locations and in diverse patient populations, then test it prospectively—without any modifications—in a new patient population. But very few algorithms have been tested this way. According to Topol's recent *Nature Medicine* review, among dozens of studies claiming an AI performs better than radiologists, only a handful were tested in populations that were different from the population where they were developed. "Algorithms are very, very delicate," says Cynthia Rudin, a computer scientist at Duke University. "If you try to use one outside the training set [of images], it doesn't always work."

As researchers become aware of this problem, more prospective studies in novel settings could be on the horizon. Barzilay's team recently finished testing its mammogram AI on 10,000 scans from the Karolinska Institute in Sweden and found that it performed just as well there as it did in Massachusetts. The group is now working with hospitals in Taiwan and Detroit to test it in more diverse patient groups. The team found that current standards for assessing breast cancer risk are much less accurate in African-American women, Barzilay says, because those standards were developed mostly using scans from white women: "I think we really are in a position to revamp this sad state of affairs."

LEGAL TERRA INCOGNITA

EVEN IF THE AI'S conclusions are medically relevant, the black box still presents a number of concerns from a legal perspective. If an AI gets a diagnosis wrong, it can be hard to determine whether the doctor or the program is at fault. "Lots of bad things happen in health care, and you don't necessarily know why the bad things happened," says Nicholson Price, a health law expert at the University of Michigan. If an AI system leads a physician to make an incorrect diagnosis, the physician may not be able to explain why and the company's data on the test's methodology are likely to be a closely guarded trade secret.

Medical AI systems are still too new to have been challenged in medical malpractice lawsuits, so it is unclear how courts will determine responsibility and what kind of transparency should be required.

The tendency to build black box algorithms frustrates Rudin. The problem comes from the fact that most medical algorithms are built by adapting deep-learning tools developed for other types of image analysis. "There's no reason you can't build a robot that can explain itself," she insists. But it is exponentially harder to build a transparent algorithm from scratch than to repurpose an existing black box algorithm to look at medical data. That is why Rudin suspects most researchers let an algorithm run and then try to understand later how it came to its conclusion.

Rudin is developing transparent AI algorithms that analyze mammograms for suspected tumors and constantly inform researchers what they are doing. But her research has been stymied by the lack of available images on which to train the algorithm. The images that are publicly available tend to be poorly labeled or taken with old machines that are no longer in use, Rudin says, and without enormous, diverse data sets, algorithms tend to pick up confounding factors.

Black boxes, along with an AI algorithm's ability to learn from experience, also present challenges to regulators. Unlike a drug, which will always work in the same way, machine-learning algorithms change and improve over time as they gain access to more patient data. Because the algorithm draws meaning from so many kinds of input, seemingly innocuous changes such as a new IT system at a hospital could suddenly ruin the AI program. "Machines can get sick just like humans get sick, and they can be in-

"You can't trust an algorithm when someone's life is on the line." —*Eric Topol, Scripps Research*

fected with malware," Topol says. "You can't trust an algorithm when you have someone's life on the line."

Last April the FDA proposed a set of guidelines to manage algorithms that evolve over time. Among them is an expectation that producers keep an eye on how their algorithms are changing to ensure they continue to work as designed and asking them to notify the agency if they see unexpected changes that might prompt reevaluation. The agency is also developing best manufacturing practices and may require companies to spell out their expectations for how algorithms might change and a protocol for how to manage those changes. "We need to understand that not one size fits all," says Bakul Patel, director of digital health at the FDA.

WILL MACHINES REPLACE DOCTORS?

THE LIMITATIONS OF AI should reassure radiologists who worry about machines taking their jobs. In 2012 technology venture capitalist and Sun Microsystems co-founder Vinod Khosla horrified a medical audience by predicting that algorithms would replace 80 percent of doctors, and more recently he claimed that radiologists still practicing in 10 years will be "killing patients." Such remarks caused panic and backlash in the radiology field, Rao says. "I think the hype is creating a lot of expectations."

But that concern has also had real impacts. In 2015 only 86 percent of radiology resident positions in the U.S. were filled, compared with 94 percent the previous year, although those numbers have improved over the past several years. And according to a 2018 survey of 322 Canadian medical students, 68 percent believed AI would reduce the demand for radiologists.

Still, most experts and AI manufacturers doubt AI will be replacing doctors any time soon. "AI solutions are becoming very good at doing one thing very well," Walach says. But because human biology is complex, he says, "you typically have to have humans who do more than one thing really well." In other words, even if an algorithm is better at diagnosing a particular problem, combining it with a physician's experience and knowledge of the patient's individual story will lead to a better outcome.

An AI that can do a single task well could free radiologists from drudgework, allowing them more time to interact with patients. "They could come out of the basement, which is where they live in the dark," Topol says. "What we need in medicine is more interhuman contact and bonding."

Still, Rao and others believe that the tools and training that radiologists receive, including their day-to-day work, will change drastically over the coming years as a result of artificial-intelligence algorithms. "AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't," says Curtis Langlotz, a radiologist at Stanford.

There are some exceptions, however. In 2018 the FDA approved the first algorithm that can make a medical decision without the need for a physician to look at the image. The program, developed by IDx Technology in Coralville, Iowa, looks at retinal images to detect diabetic retinopathy and is 87 percent accurate, according to the company's data. IDx chief executive officer Michael Abramoff says that because no doctor is involved, the company has assumed legal liability for any medical errors.

In the short term, AI algorithms are more likely to assist doctors than replace them. For instance, physicians working in developing countries might not have access to the same kinds of scanners as a major medical institution in the U.S. or Europe or trained radiologists who can interpret scans. As medicine becomes more specialized and dependent on image analysis, the gap between the standard of care provided in wealthier and poorer areas is growing, Lungren says. Running an algorithm can be a cheap way to close that gap and may even be done on a mobile phone.

Lungren's group is developing a tool that allows doctors to take cell-phone pictures of an x-ray film—not the digital scans that are standard in wealthy nations—and run an algorithm on the photographs that detects problems such as tuberculosis. "It's not replacing anybody," he says—many developing countries have no radiologists in the first place. "We're augmenting nonradiologists to bring expertise to their fingertips."

Another short-term goal of AI could be to examine medical records to determine whether a patient needs a scan in the first place, Rao says. Many medical economists believe that imaging is overused—more than 80 million CT scans are performed every year in the U.S. alone. Although this abundance of data is helpful to researchers using it to train algorithms, scans are extraordinarily costly and can expose patients to unnecessary amounts of radiation. Similarly, Langlotz adds that algorithms could one day analyze images while a patient is still in the scanner and predict the final outcome, thus reducing the amount of time and radiation exposure required to get a good image.

Ultimately, Barzilay says, AI will be most useful when it serves as a sharp-eyed partner in tackling problems that doctors cannot detect and solve alone. "If there were a convenient and describable pattern," she notes, "humans would already be able to do it." She knows firsthand that, too often, this is not the case.

Sara Reardon is a freelance journalist based in Bozeman, Mont. She is a former staff reporter at *Nature*, *New Scientist* and *Science* and has a master's degree in molecular biology.