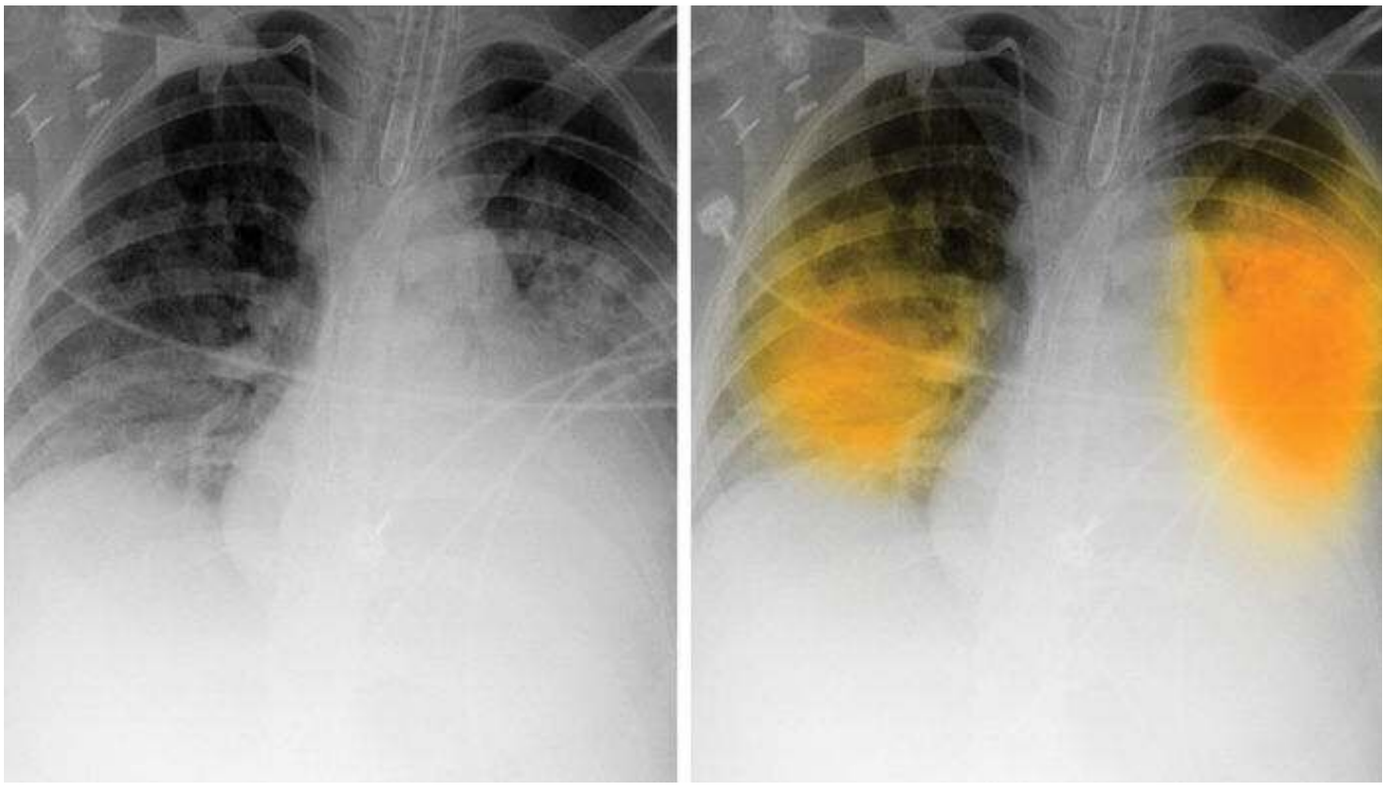
Artificial intelligence could revolutionize medical care. But don’t trust it to read your x-ray just yet

By Jennifer Couzin-FrankelJun. 17, 2019 , 12:45 PM



Scientists are developing a multitude of artificial intelligence algorithms to help radiologists, like this one that lights up likely pneumonia in the lungs.

Albert Hsiao and Brian Hurt/UC San Diego AiDA Laboratory

Artificial intelligence (AI) is poised to upend the practice of medicine, boosting the efficiency and accuracy of diagnosis in specialties that rely on images, such as radiology and pathology. But as the technology gallops ahead, experts are grappling with its potential downsides. “Just working with the technology, I see lots of ways it can fail,” says Albert Hsiao, a radiologist at the University of California, San Diego, who has developed algorithms for reading cardiac images and improving their quality. One major concern: Most AI software is designed and tested in one hospital, and it risks faltering when transferred to another.

Last month in the *Journal of the American College of Radiology*, U.S. government scientists, regulators, and doctors published a road map describing how to convert research-based AI into improved medical imaging on patients. Among other things, the authors urged more collaboration across disciplines in building and testing AI algorithms, and intensive validation of algorithms before they reach patients. For now, Hsiao says, “I would want a human physician no matter what,” even if a machine hums alongside.

Most AI in medicine is used in research, but regulators have already approved some algorithms for radiologists. Physicians are also developing their own—which they’re permitted to use without regulatory approval as long as companies aren’t marketing the new technology. Studies are testing algorithms to read x-rays, detect brain bleeds, pinpoint tumors, and more.

The algorithms learn as scientists feed them hundreds or thousands of images—of mammograms, for example—training the technology to recognize patterns faster and more accurately than a human could. "If I'm doing an MRI of a moving heart, I can have the computer predict where the heart's going to be in the next fraction of a second and get a better picture instead of a blurry" one, says Krishna Kandarpa, a cardiovascular and interventional radiologist at the National Institute of Biomedical Imaging and Bioengineering in Bethesda, Maryland. Or AI might analyze computerized tomography heads scans of suspected strokes, label those more likely to harbor a brain bleed, and put them on top of the pile for the radiologist to examine. An algorithm could help spot breast tumors in mammograms that a radiologist's eyes risk missing.

But Eric Oermann, a neurosurgeon at Mount Sinai Hospital in New York City, has explored one downside of the algorithms: The signals they recognize can have less to do with disease than with other patient characteristics, the brand of MRI machine, or even how a scanner is angled. With colleagues, Oermann developed a mathematical model for detecting patterns consistent with pneumonia and trained it with x-rays from patients at Mount Sinai. The hospital has a busy intensive care unit with many elderly people, who are often admitted with pneumonia; 34% of the Mount Sinai x-rays came from infected patients.

When the algorithm was tested on a different batch of Mount Sinai x-rays it performed admirably, accurately detecting pneumonia 93% of the time. But Oermann also tested it on tens of thousands of images from two other sites: the National Institutes of Health Clinical Center in Bethesda and the Indiana Network for Patient Care. With x-rays from those locations—where pneumonia rates just squeaked past 1%—the success rate fell, ranging from 73% to 80%, the team reported last year in *PLOS Medicine*. "It didn't work as well because the patients at the other hospitals were different," Oermann says.

At Mount Sinai, many of the infected patients were too sick to get out of bed, and so doctors used a portable chest x-ray machine. Portable x-ray images look very different from those created when a patient is standing up. Because of what it learned from Mount Sinai's x-rays, the algorithm began to associate a portable x-ray with illness. It also anticipated a high rate of pneumonia.

Few multisite studies like Oermann's have been published, and last month's road map deemed this worrying. This year, a South Korean team reported in the *Korean Journal of Radiology* an analysis of 516 studies of AI algorithms designed to interpret medical images. The authors found that just 6% of the studies tested their algorithm at more than one hospital. "It's very concerning," says Elaine Nsoesie, a computational epidemiologist at Boston University who wasn't involved in the work. Even the brand of scanner matters, as the pixel pattern can vary, disrupting how AI assesses the image.

One way to avoid this pitfall, Nsoesie says, is to test an algorithm using data from several hospitals. Researchers are starting to do this, she says, "but less than you would think." A rare example is an algorithm first trained and tested on data at Stanford's Lucile Packard Children's Hospital in Palo Alto, California, and Children's Hospital Colorado in Aurora. It's now undergoing testing in a clinical trial on scans from nine sites. The software measures skeletal maturity in hand x-rays, which orthopedists use to guide treatment for growth disorders in children and teenagers.

In another field that relies on images, pathology, Jeroen van der Laak, a computer scientist at Radboud University Medical Center in the Netherlands, tried a new way to encourage researchers to test their algorithms across hospitals: a competition. In 2015, van der Laak gathered and digitized 400 lymph node slides from breast cancer patients at two Dutch centers. Then, he invited all comers to train their algorithm on 270 of those slides and test it on the remaining 130, to see whether it could do better than pathologists hunting for tiny cancers. Twenty-three teams submitted 32 algorithms.

The results, published in 2017 in *JAMA*, showed that 10 matched or exceeded a panel of 11 pathologists. The top-performing algorithm, from a group at Harvard University and the Massachusetts Institute of Technology in Cambridge, matched a pathologist who took an entire weekend to go over 130 slides. "To actually see that you could be as good as a pathologist" was "a shock," van der Laak says.

He and others say that to achieve that kind of accuracy, AI algorithms should train on data that are diverse not only in their hospital of origin but also in racial and geographic diversity, because disease can manifest differently across populations.

The U.S. Food and Drug Administration (FDA) continues to weigh how to assess algorithms for patient care. The agency considers "locked" AI software, which is unchanging, as a medical device. But earlier this year, it announced it was developing a framework for regulating more cutting-edge AI software that continues to learn over time. Still, there "are major questions everyone is struggling with" around regulation, says Hugo Aerts, who directs the computational imaging and bioinformatics laboratory at the Dana-Farber Cancer Institute in Boston. What if developers update an algorithm that works in 96% of cases to achieve a 99% success rate; do they have to go through the regulatory process again? What if an approved algorithm is applied to a patient population it wasn't originally tested on?

FDA has already issued some approvals. One algorithm, created by Hsiao, measures heart size and blood flow in a cardiac MRI. Hsiao was frustrated that analyzing the data by hand took at least 6 hours, so he returned to his roots as a computer science major and wrote his own software. He subsequently formed a company, Arterys, based in San Francisco, California, and won FDA approval in about 6 months, he says. Hsiao is now working on algorithms to make it easier to pick up pneumonia by mapping its likely location in the lungs.

But, he says, the doctor, not the machine, is still the boss and entitled to override the technology. "If I think it's not pneumonia," Hsiao says, "it's not."