ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

AP

When covid-19 struck Europe in March 2020, hospitals were plunged into a health crisis that was still badly understood. “Doctors really didn’t have a clue how to manage these patients,” says Laure Wynants, an epidemiologist at Maastricht University in the Netherlands, who studies predictive tools.

But there was data coming out of China, which had a four-month head start in the race to beat the pandemic. If machine-learning algorithms could be trained on that data to help doctors understand what they were seeing and make decisions, it just might save lives. “I thought, ‘If there’s any time that AI could prove its usefulness, it’s now,’” says Wynants. “I had my hopes up.”

It never happened—but not for lack of effort. Research teams around the world stepped up to help. The AI community, in particular, rushed to develop software that many believed would allow hospitals to diagnose

or triage patients faster, bringing much-needed support to the front lines—in theory.

In the end, many hundreds of predictive tools were developed. None of them made a real difference, and some were potentially harmful.

That's the damning conclusion of multiple studies published in the last few months. In June, the Turing Institute, the UK's national center for data science and AI, put out a report summing up discussions at a series of workshops it held in late 2020. The clear consensus was that AI tools had made little, if any, impact in the fight against covid.

Not fit for clinical use

This echoes the results of two major studies that assessed hundreds of predictive tools developed last year. Wynants is lead author of one of them, a review in the British Medical Journal that is still being updated as new tools are released and existing ones tested. She and her colleagues have looked at 232 algorithms for diagnosing patients or predicting how sick those with the disease might get. They found that none of them were fit for clinical use. Just two have been singled out as being promising enough for future testing.

"It's shocking," says Wynants. "I went into it with some worries, but this exceeded my fears."

Wynants's study is backed up by another large review carried out by Derek Driggs, a machine-learning researcher at the University of Cambridge, and his colleagues, and published in Nature Machine Intelligence. This team zoomed in on deep-learning models for diagnosing covid and predicting patient risk from medical images, such as chest x-rays and chest computer tomography (CT) scans. They looked at 415 published tools and, like Wynants and her colleagues, concluded that none were fit for clinical use.

"This pandemic was a big test for AI and medicine," says Driggs, who is himself working on a machine-learning tool to help doctors during the pandemic. "It would have gone a long way to getting the public on our side," he says. "But I don't think we passed that test."

Both teams found that researchers repeated the same basic errors in the way they trained or tested their tools. Incorrect assumptions about the data often meant that the trained models did not work as claimed.

Wynants and Driggs still believe AI has the potential to help. But they are concerned that it could be harmful if built in the wrong way because they could miss diagnoses or underestimate risk for vulnerable patients. "There is a lot of hype about machine-learning models and what they can do today," says Driggs.

Unrealistic expectations encourage the use of these tools before they are ready. Wynants and Driggs both say that a few of the algorithms they looked at have already been used in hospitals, and some are being marketed by private developers. "I fear that they may have harmed patients," says Wynants.

So what went wrong? And how do we bridge that gap? If there's an upside, it is that the pandemic has made it clear to many researchers that the way AI tools are built needs to change. "The pandemic has put problems in the spotlight that we've been dragging along for some time," says Wynants.

What went wrong

Many of the problems that were uncovered are linked to the poor quality of the data that researchers used to develop their tools. Information about covid patients, including medical scans, was collected and shared in the middle of a global pandemic, often by the doctors struggling to treat those patients. Researchers wanted to help quickly, and these were the only public data sets available. But this meant that many tools were built using mislabeled data or data from unknown sources.

Driggs highlights the problem of what he calls Frankenstein data sets, which are spliced together from multiple sources and can contain duplicates. This means that some tools end up being tested on the same data they were trained on, making them appear more accurate than they are.

It also muddies the origin of certain data sets. This can mean that researchers miss important features that skew the training of their models. Many unwittingly used a data set that contained chest scans of children who did not have covid as their examples of what non-covid cases looked like. But as a result, the AIs learned to identify kids, not covid.

Driggs's group trained its own model using a data set that contained a mix of scans taken when patients were lying down and standing up. Because patients scanned while lying down were more likely to be seriously ill, the AI learned wrongly to predict serious covid risk from a person's position.

In yet other cases, some AIs were found to be picking up on the text font that certain hospitals used to label the scans. As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.

Errors like these seem obvious in hindsight. They can also be fixed by adjusting the models, if researchers are aware of them. It is possible to acknowledge the shortcomings and release a less accurate, but less misleading model. But many tools were developed either by AI researchers who lacked the medical expertise to spot flaws in the data or by medical researchers who lacked the mathematical skills to compensate for those flaws.

A more subtle problem Driggs highlights is incorporation bias, or bias introduced at the point a data set is labeled. For example, many medical scans were labeled according to whether the radiologists who created them said they showed covid. But that embeds, or incorporates, any biases of that particular doctor into the ground truth of a data set. It would be much better to label a medical scan with the result of a PCR test rather than one doctor's opinion, says Driggs. But there isn't always time for statistical niceties in busy hospitals.

That hasn't stopped some of these tools from being rushed into clinical practice. Wynants says it isn't clear which ones are being used or how. Hospitals will sometimes say that they are using a tool only for research purposes, which makes it hard to assess how much doctors are relying on them. "There's a lot of secrecy," she says.

Wynants asked one company that was marketing deep-learning algorithms to share information about its approach but did not hear back. She later found several published models from researchers tied to this company, all of them with a high risk of bias. “We don’t actually know what the company implemented,” she says.

According to Wynants, some hospitals are even signing nondisclosure agreements with medical AI vendors. When she asked doctors what algorithms or software they were using, they sometimes told her they weren’t allowed to say.

How to fix it

What’s the fix? Better data would help, but in times of crisis that’s a big ask. It’s more important to make the most of the data sets we have. The simplest move would be for AI teams to collaborate more with clinicians, says Driggs. Researchers also need to share their models and disclose how they were trained so that others can test them and build on them. “Those are two things we could do today,” he says. “And they would solve maybe 50% of the issues that we identified.”

Getting hold of data would also be easier if formats were standardized, says Bilal Mateen, a doctor who leads the clinical technology team at the Wellcome Trust, a global health research charity based in London.

Another problem Wynants, Driggs, and Mateen all identify is that most researchers rushed to develop their own models, rather than working together or improving existing ones. The result was that the collective effort of researchers around the world produced hundreds of mediocre tools, rather than a handful of properly trained and tested ones.

“The models are so similar—they almost all use the same techniques with minor tweaks, the same inputs—and they all make the same mistakes,” says Wynants. “If all these people making new models instead tested models that were already available, maybe we’d have something that could really help in the clinic by now.”

In a sense, this is an old problem with research. Academic researchers have few career incentives to share work or validate existing results. There’s no reward for pushing through the last mile that takes tech from “lab bench to bedside,” says Mateen.

To address this issue, the World Health Organization is considering an emergency data-sharing contract that would kick in during international health crises. It would let researchers move data across borders more easily, says Mateen. Before the G7 summit in the UK in June, leading scientific groups from participating nations also called for “data readiness” in preparation for future health emergencies.

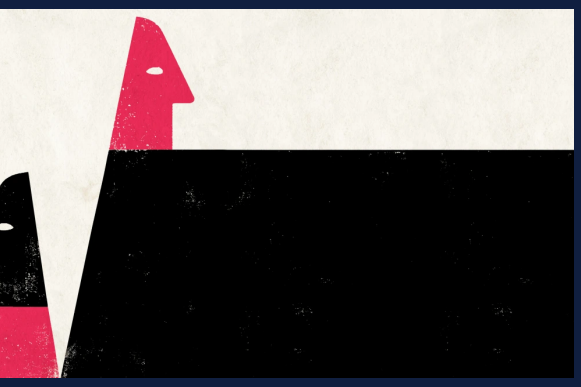
Such initiatives sound a little vague, and calls for change always have a whiff of wishful thinking about them. But Mateen has what he calls a “naïvely optimistic” view. Before the pandemic, momentum for such initiatives had stalled. “It felt like it was too high of a mountain to hike and the view wasn’t worth it,” he says. “Covid has put a lot of this back on the agenda.”

“Until we buy into the idea that we need to sort out the unsexy problems before the sexy ones, we’re doomed to repeat the same mistakes,” says Mateen. “It’s unacceptable if it doesn’t happen. To forget the lessons of this pandemic is disrespectful to those who passed away.”

by Will Douglas Heaven

DEEP DIVE

ARTIFICIAL INTELLIGENCE



Artificial intelligence is creating a new colonial world order

An MIT Technology Review series investigates how AI is enriching a powerful few by dispossessing communities that have been dispossessed before.

By Karen Hao

This horse-riding astronaut is a milestone in AI's journey to make sense of the world

OpenAI's latest picture-making AI is amazing—but raises questions about what we mean by intelligence.

By Will Douglas Heaven

What do psychedelic drugs do to our brains? AI could help us find out

The words people used to describe their trip experiences could lead to better drugs to treat mental illness.

By Jessica Hamzelou

This is the reason Demis Hassabis started DeepMind

A year after it took biologists by surprise,

AlphaFold has changed how researchers work and set DeepMind on a new course.

By Will Douglas Heaven

STAY CONNECTED

Illustration by Rose Wong

Get the latest updates from MIT Technology Review

Discover special offers, top stories, upcoming events, and more.

Enter your email

Privacy Policy

MIT Technology Review

Our in-depth reporting reveals what's going on now to prepare you for what's coming next.

Subscribe to support our journalism.

Cover Art by Michael Byers

Back to top ↑