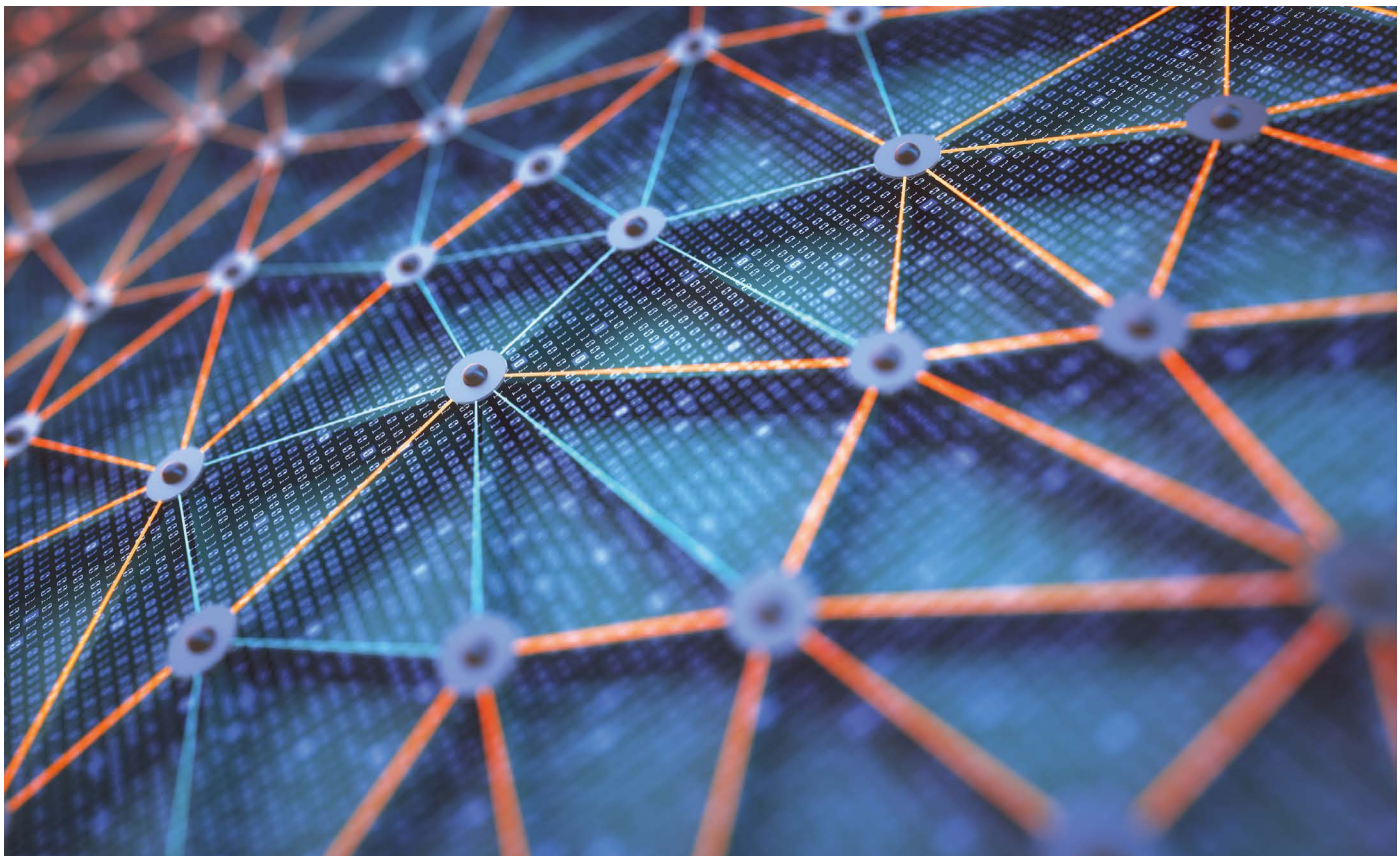

KTSIMAGE/GETTY

The use of artificial intelligence in medicine is growing rapidly.

THE REPRODUCIBILITY ISSUES THAT HAUNT HEALTH-CARE AI

Health-care systems are rolling out artificial-intelligence tools for diagnosis and monitoring. But how reliable are the models? **By Emily Sohn**

Each day, around 350 people in the United States die from lung cancer. Many of those deaths could be prevented by screening with low-dose computed tomography (CT) scans. But scanning millions of people would produce millions of images, and there aren't enough radiologists to do the work. Even if there were, specialists regularly disagree about whether images show cancer or not. The 2017 Kaggle Data Science Bowl set out to test whether machine-learning algorithms could fill the gap.

An online competition for automated lung cancer diagnosis, the Data Science Bowl provided chest CT scans from 1,397 patients to hundreds of teams, for the teams to develop and test their algorithms. At least five of the winning models demonstrated accuracy exceeding 90% at detecting lung nodules. But to be clinically useful, those algorithms would have to perform equally well on multiple data sets.

To test that, Kun-Hsing Yu, a data scientist at Harvard Medical School in Boston, Massachusetts, acquired the ten best-performing algorithms and challenged them on a subset of the data used in the original competition. On these data, the algorithms topped out at 60–70% accuracy, Yu says. In some cases, they were effectively coin tosses[1]. "Almost all of these award-winning models failed miserably," he says. "That was kind of surprising to us."

But maybe it shouldn't have been. The artificial-intelligence (AI) community faces a reproducibility crisis, says Sayash Kapoor, a PhD candidate in computer science at Princeton University in New Jersey. As part of his work on the limits of computational prediction, Kapoor discovered that reproducibility failures and pitfalls had been reported in 329 studies across 17 fields, including medicine. He and a colleague organized a one-day online workshop last July to discuss the subject, which attracted about 600 participants from 30 countries. The resulting videos have been viewed more than 5,000 times.

It's all part of a broader move towards increased reproducibility in health-care AI, including strategies such as greater algorithmic transparency and promoting checklists to avoid common errors.

These improvements cannot come soon enough, says Casey Greene, a computational biologist at the University of Colorado School of Medicine in Aurora. "Given the exploding nature and how widely these things are being used," he says, "I think we need to get better more quickly than we are."

Big potential, high stakes

Algorithmic improvements, a surge in digital data and advances in computing power and performance have quickly boosted the potential of machine learning to accelerate

diagnosis, guide treatment strategies, conduct pandemic surveillance and address other health topics, researchers say.

To be broadly applicable, an AI model needs to be reproducible, which means the code and data should be available and error-free, Kapoor says. But privacy issues, ethical concerns and regulatory hurdles have made reproducibility elusive in health-care AI, says Michael Roberts, who studies machine learning at the University of Cambridge, UK.

In a review[2] of 62 studies that used AI to diagnose COVID-19 from medical scans, Roberts and his colleagues found that none of the models was ready to be deployed clinically for use in diagnosing or predicting the prognosis of COVID-19, because of flaws such as biases in the data, methodology problems and reproducibility failures.

Health-related machine-learning models perform particularly poorly on reproducibility measures relative to other machine-learning disciplines, researchers reported in a 2021 review[3] of more than 500 papers presented at machine-learning conferences between 2017 and 2019. Marzyeh Ghassemi, a computational-medicine researcher at the Massachusetts Institute of Technology (MIT) in Cambridge who led the review, found that a major issue is the relative scarcity of publicly available data sets in medicine. As a result, biases and inequities can become entrenched.

For example, if researchers train a drug-prescription model on data from physicians who prescribe medications more to one racial group than another, results could be skewed on the basis of what physicians do rather than what works, Greene says.

Another issue is data 'leakage': overlap between the data used to train a model and the data used to test it. These data sets should be completely independent, Kapoor says. But medical databases can include entries for the same patient, duplications that scientists who use the data might not be aware of. The result could be an overly optimistic impression of performance, Kapoor says.

Septic shock

Despite these concerns, AI systems are already being used in the clinic. For instance, hundreds of US hospitals have implemented a model in their electronic health-record systems to flag early signs of sepsis, a systemic infection that accounts for more than 250,000 deaths in the United States each year. The tool, called the Epic Sepsis Model, was trained on 405,000 patient encounters at 3 health-care systems over a 3-year period, according to its creator Epic Systems, based in Verona, Wisconsin.

To evaluate it independently, researchers at the University of Michigan Medical School in Ann Arbor analysed 38,455 hospitalizations involving 27,697 people. The tool, they reported in 2021, produced a lot of false alarms, generating alerts on more than twice the number of people who actually had sepsis. And it failed to identify 67% of people who actually had sepsis[4]. (The company has since overhauled the models.)

Proprietary models make it hard to spot faulty algorithms, Greene says, and greater transparency could help to prevent them from becoming so widely deployed. "At the end of the day," Greene says, "we have to ask, 'Are we deploying a bunch of algorithms in practice that we can't understand, for which we don't know their biases, and that might create real harm for people?'"

Making models and data publicly available helps everyone, says Emma Lundberg, a bioengineer at Stanford University in California, who has applied machine learning to protein imaging. "Then someone could use it on their own data set and find, 'Oh, it's not working perfectly, so we're going to tweak it', and then that tweak is going to make it applicable elsewhere," she says.

Positive moves

Scientists are increasingly moving in the right direction, Kapoor says, producing large data sets covering institutions, countries and populations, and that are open to all. Examples include the national biobanks of the United Kingdom and Japan, as well as the eICU Collaborative Research Database, which includes data associated with around 200,000 critical-care-unit admissions, made available by Amsterdam-based Philips Healthcare and the MIT Laboratory for Computational Physiology.

"Almost all of these award-winning models failed miserably. That was kind of surprising to us."

Ghassemi and her colleagues say that having even more options would add value. They have called for[3] the creation of standards for collecting data and reporting machine-learning studies, allowing participants to give consent to the use of their data, and adopting approaches that ensure rigorous and privacy-preserving analyses. For example, an effort called the Observational Medical Outcomes Partnership Common Data Model allows patient and treatment information to be collected in the same way across institutions. Something similar, the researchers wrote, could enhance machine-learning research in health care, too.

Eliminating data redundancy would also help, says Søren Brunak, a translational-disease systems biologist at the University of Copenhagen. In machine-learning studies that predict protein structures, he says, scientists have had success in removing proteins from test sets that are too similar to proteins used in training sets. But in health-care studies, a database might include many similar individuals, which doesn't challenge the algorithm to develop insight beyond the most typical patients. "We need to work on the pedagogical side – what data are we actually showing to the algorithms – and be better at balancing that and making the data sets representative," Brunak says.

Widely used in health care, checklists provide a simple way to reduce technical issues and improve reproducibility, Kapoor suggests. In machine learning, checklists could help to ensure that researchers attend to the many small steps that need to be done correctly and in order, so that results are valid and reproducible, Kapoor says.

Multiple machine-learning checklists are already available, many spearheaded by the Equator Network, an international initiative to improve the reliability of health research. The TRIPOD checklist (see www.tripod-statement.org), for instance, includes 22 items to guide the reporting of studies of predictive health models. The Checklist for AI in Medical Imaging, or CLAIM, lists 42 items[5], including whether a study is retrospective or prospective, and how well the data match the intended use of the model.

In July 2022, Kapoor and colleagues published a list of 21 questions to help reduce data leakage (see go.nature.com/3veyw3j). For example, if a model is being used to predict an outcome, the checklist advises researchers to confirm whether data in the training set pre-dates the test set, a sign that they are independent.

Although there is still much to do, growing dialogue around reproducibility in machine learning is encouraging and helps to counteract what has been a siloed state of research, researchers say. After the July online workshop, nearly 300 people joined a group on the online collaboration platform Slack to continue the discussion, Kapoor says. And at scientific conferences, reproducibility has become a frequent focus, Greene adds. "It used to be a small esoteric group of people who cared about reproducibility. Now it feels like people are asking questions, and conversations are moving forward. I would love for it to move forward faster, but at least it feels less like shouting into the void."

Emily Sohn is a freelance journalist based in Minneapolis, Minnesota.

1. Yu, K. H. *et al. J. Med. Internet Res.* **22**, e16709 (2020).
2. Roberts, M. *et al. Nature Mach. Intell.* **3**, 199–217 (2021).
3. McDermott, M. B. A. *et al. Sci. Transl. Med.* **13**, eabb1655 (2021).
4. Wong, A. *et al. JAMA Intern. Med.* **181**, 1065–1070 (2021).
5. Mongan, J., Moy, L. & Kahn, C. E. *Radiol. Artif. Intell.* **2**, e200029 (2020).

Correction

This Technology feature erroneously stated that Sayash Kapoor discovered reproducibility failures and pitfalls in 329 studies across 17 fields. In fact, those studies had themselves reported reproducibility failures and pitfalls.

Corrected 12 January 2023