STAT E-BOOK

The advance of artificial intelligence into health care

SPONSORED BY NUANCE

STAT

TABLE OF CONTENTS

In medical parlance, "stat" means important and urgent, and that's what we're all about — quickly and smartly delivering good stories. We take you inside science labs and hospitals, biotech boardrooms, and political backrooms. We dissect crucial discoveries. We examine controversies and puncture hype. We hold individuals and institutions accountable. We introduce you to the power brokers and personalities who are driving a revolution in human health. These are the stories that matter to us all.

BOSTON • WASHINGTON • NEW YORK

SAN FRANCISCO • LOS ANGELES • CLEVELAND

Our team includes talented writers, editors, and producers capable of the kind of explanatory journalism that complicated science issues sometimes demand. And even if you don't work in science, have never stepped foot in a hospital, or hated high school biology, we've got something for you. The world of health, science, and medicine is booming and yielding fascinating stories. We explore how they affect us all. And, with our eBook series, we regularly do deep dives into timely topics to get you the inside scoop you need.

The advance of artificial intelligence into health care

Artificial intelligence is taking hold in medicine.

Every sector of the health care system is testing the technology to improve patient outcomes, speed up research, and cut costs. Biopharma companies are turning to artificial intelligence to speed screening and find new uses for old drugs, while electronic health records giants and startups alike are building out new models to guide care.

And as the role of artificial intelligence grows, so, too, do the questions surrounding that shift. How can investors and health system buyers separate the truly innovative technologies from the ones that won't prove useful? What will it take for the health care system and drug companies to seamlessly integrate this technology into long-standing workflows? And how will developers, health systems, and regulators ensure that everyone benefits from new models — and that they don't exacerbate existing problems in medicine?

In these stories, STAT examines how the biopharma and health care industries are navigating those questions.

NUANCE

Restore joy to the practice of medicine with AI

Nuance® Dragon® Ambient eXperience™ (DAX) captures conversations as they happen in the exam room or during virtual interactions with the patient's consent. It then utilizes AI to convert what was said into clinical notes containing all pertinent details—lessening feelings of burnout and allowing physicians to give their undivided attention to patients

- 79% of physicians state that DAX has improved documentation quality AND 50% reduction in documentation time
- 70% of physicians have reported reductions in feelings of burnout and fatigue because of DAX
- 83% of patients say their physician is more personable and conversational

Better patient-physician experience, all around.
Nuance.com/DAX

STAT

HEALTH TECH

‘We need to be much more diverse’: More than half of data used in health care AI comes from the U.S. and China

By Katie Palmer | AUGUST 6, 2022

As medicine continues to test automated machine learning tools, many hope that low-cost support tools will help narrow care gaps in countries with constrained resources. But new research suggests it’s those countries that are least represented in the data being used to design and test most clinical AI — potentially making those gaps even wider.

Researchers have shown that AI tools often fail to perform when used in real-world hospitals. It’s the problem of transferability: An algorithm trained on one patient population with a particular set of characteristics won’t necessarily work well on another. Those failures have motivated a growing call for clinical AI to be both trained and validated on diverse patient data, with representation across spectrums of sex, age, race, ethnicity, and more.

But the patterns of global research investment mean that even if individual scientists make an effort to represent a range of patients, the field as a whole skews significantly toward just a few nationalities.

In a review of more than 7,000 clinical AI papers, all published in 2019, researchers revealed more than half of the databases used in the work came from the U.S. and China, and high-income countries represented the majority of the remaining patient datasets.

"Look, we need to be much more diverse in terms of the datasets we use to create and validate these algorithms," said Leo Anthony Celi, first author of the paper in PLoS Digital Health (he is also the journal's editor). "The biggest concern now is that the algorithms that we're building are only going to benefit the population that's contributing to the dataset. And none of that will have any value to those who carry the biggest burden of disease in this country, or in the world."

The skew in patient data isn't unexpected, given Chinese and American dominance in machine learning infrastructure and research. "To create a dataset you need electronic health records, you need cloud storage, you need computer speed, computer power," said co-author William Mitchell, a clinical researcher and ophthalmology resident in Australia. "So it makes sense that the U.S. and China are the ones that are in effect storing the most data." The survey also found Chinese and American researchers accounted for more than 40% of the clinical AI papers, as measured by the inferred nationality of first and last authors; it's no surprise that researchers gravitate toward the patient data that's closest — and easiest — to access.

But the risk posed by the global bias in patient representation makes it worth calling out and addressing those ingrained tendencies, the authors argue. Clinicians know that algorithms can perform differently in neighboring hospitals that serve different patient populations. They can even lose power over time within the same hospital, as subtle shifts in practice alter the data that flows into a tool.

"Between an institution from São Paulo and an institution in Boston, I think the differences are going to be much, much bigger," said Celi, who leads the Laboratory of Computational Physiology at MIT. "Potentially, the scale and the magnitude of errors would be greater."

Clinician guidelines are already tailored to well-resourced countries, and a lack of diverse patient data only stands to widen global health care inequality. "Most of the research that informs how we practice medicine is performed in a few rich countries, and then there's an assumption that whatever we learn from these studies and trials performed in a few rich countries will generalize to the rest of the world," said Celi. "This is also going to be an issue if we don't change the trajectory with respect to the creation of artificial intelligence for health care."

The answer isn't straightforward, because nations that are resource-poor are also more likely to be data-poor. One popular research target for clinical AI in low-resourced settings is automated screening for eye disease. Using a portable fundus camera to image the eye, or even a smartphone camera, an algorithm could identify the signs of problems like diabetic retinopathy early enough to intervene. But as the authors note, 172 countries accounting for 3.5 billion people have no public ophthalmic data repository for researchers to draw from — data deserts that frequently also affect other fields of medicine.

That's why Celi and others are investing in programs to encourage data collection and pooling of machine learning resources in poorly-represented countries. One consortium is assembling multidisciplinary experts from Mexico, Chile, Argentina, and Brazil to "identify best practices in data diplomacy," said Celi. "It turns out the biggest challenge here is really the politics and economics of data," encouraging those with access to clinical data to open it up for local and international research rather than hoarding it for commercial purposes.

That work can also help double down on efforts to test existing models in areas with data disparities. If local data collection and curation isn't possible yet, validation can help ensure that algorithms trained in data-rich countries can, at least, be safely deployed in other settings. And along the way, those efforts can start to lay the groundwork for long-term data collection, and the ultimate growth of international data repositories.

By quantifying the international bias in AI research, Celi says, "we just don't end up with 'things are pretty bad.'" The group hopes to use this as a baseline against which to measure improvement. Another recent paper led by Joe Zhang at Imperial College London detailed the creation of a dashboard that tracks the publication of clinical AI research, including the nationality of the first author on each paper. The first step to solving the problem is measuring it.

STAT

HEALTH TECH

When AI takes a human touch: How a team effort to improve patient care in hospitals paid off

By Casey Ross | MARCH 21, 2022

The project began with a vexing problem. Imaging tests that turned up unexpected issues — such as suspicious lung nodules — were being overlooked by busy caregivers, and patients who needed prompt follow-up weren't getting it.

After months of discussion, the leaders of Northwestern Medicine coalesced around a heady solution: Artificial intelligence could be used to identify these cases and quickly ping providers.

If only it were that easy.

It took three years to embed AI models to flag lung and adrenal nodules into clinical practice, requiring thousands of work hours by employees who spanned the organization — from radiologists, to human resources specialists, to nurses, primary care doctors, and IT experts. Developing accurate models was the least of their problems. The real challenge was building trust in their conclusions and designing a system to ensure the tool's warnings didn't just lead providers to click past a pop-up, and instead translated to effective, real-world care.

"There were so many surprises. This was a learning experience every day," said Jane Domingo, a project manager in Northwestern's office of clinical improvement. "It's amazing to think of the sheer number of different people and expertise that we pulled together to make this work."

Ultimately, the adrenal model failed to produce the necessary level of accuracy in live testing. But the lung model, by far the most common source of suspicious lesions, proved highly adept at notifying caregivers, paving the way for thousands of follow-up tests for patients, according to a paper published last week in NEJM Catalyst. Additional study is needed to determine whether those tests are reducing the number of missed cancers.

STAT interviewed employees across Northwestern who were involved in building the algorithm, incorporating it into IT systems, and pairing it with protocols to ensure that patients received the rapid follow-up that had been recommended. The challenges they faced, and what it took to overcome them, underscores that AI's success in medicine hinges as much on human effort and understanding as it does on the statistical accuracy of the algorithm itself.

Here's a closer look at the players involved in the project and the obstacles they faced along the way.

The annotators

To get the AI to flag the right information, it needed to be trained on labeled examples from the health system. Radiology reports had to be marked up to note incidental findings and recommendations for follow-up. But who had the time to mark up tens of thousands of clinical documents to help the AI spot the telltale language?

The human resources department had an idea: Nurses who had been put on light duty due to work injuries could be trained to scan the reports and pluck out key excerpts. That would eliminate the need to hire a high-priced third party with unknown expertise.

However, highlighting discreet passages in lengthy radiology reports is not as easy as it sounds, said Stacey Caron, who oversaw the team of nurses doing the annotation. "Radiologists write their reports differently, and some of them will be more specific in their recommendations, and others will be more vague," she said. "We had to make sure the education on how [to mark relevant excerpts] was clear."

Caron met with nurses individually to orient them to the project and created a training video and written instructions to guide their work. Each report had to be annotated by multiple nurses to ensure accurate labeling. In the end, the nurses logged about 8,000 work hours annotating more than 53,000 distinct reports, creating a high-quality data stream to help train the AI.

The model builders

Developing the AI models may not have been the hardest task in the project, but it was crucial to its success. There are several different approaches to analyzing text with AI — a task known as natural language processing. Picking the wrong one means certain failure.

The team started with a model known as regular expression, or regex, which searches for manually defined word sequences within text, like "non-contrast chest CT." But because of the variability in wording used by radiologists in their reports, the AI became too error-prone. It missed an unacceptable number of suspicious nodules in need of follow-up, and flagged too many reports where they didn't exist.

Next, the AI specialists, led by Mozziyar Etemadi, a professor of biomedical engineering at Northwestern, tried a machine learning approach called bag-of words, which counts the number of times a word is used from a pre-selected list of vocabulary, creating a numeric representation that can be fed into the model. This, too, failed to achieve the desired level of accuracy.

The shortcomings of those relatively simple models pointed to the need for a more complex architecture known as deep learning, where data are passed through multiple processing layers in which the model learns key features and relationships. This method allowed the AI to understand dependencies between words in the text.

Early testing showed the model almost never missed a report that flagged a suspicious nodule.

"It's really a testament to these deep learning tools," said Etemadi. "When you throw more and more data at it, it gets it. These tools really do learn the underlying structure of the English language."

But technical proficiency, though an important milestone, was not enough for the AI to make a difference in the clinic. Its conclusions would only matter if people knew what to do with them.

"AI cannot show up and give the clinicians more work," said Northwestern Medicine's chief medical officer, James Adams, who championed the project in the health system's executive ranks. "It needs to be an agent of the frontline people, and that's different from how health care technology of this past generation has been implemented."

The alert architects

A commonly used vehicle for delivering timely information to clinicians is known as a best practice alert, or BPA — a message that pops up in health records software.

Clinicians are already bombarded with such alerts, and adding to the list is a touchy subject. "We kind of have to have our ducks in a row, because if it's interruptive, it's going to face some resistance from physicians," said Pat Creamer, a program manager for information services.

The solution in this case was to embed the alert in clinicians' inboxes, where two red exclamation marks signify a message requiring immediate attention. To reinforce trust in the validity of the AI's alert, the relevant text from the original report was embedded within the message, along with a hyperlink that allows physicians to easily order the recommended follow-up test.

Creamer said the message also allows clinicians to reject the recommendation if other information indicates follow-up is not needed, such as ongoing management of the patient by someone else. The message can also be transferred to that other caregiver.

The most important part of the alert, Creamer said, was building it into the record-keeping system so that the team could keep tabs on each part of the process. "It's not a normal BPA," he said, "because it's got programming behind it that's helping us track the findings and recommendations throughout the whole lifecycle."

And in cases where patients didn't receive follow-up, they were ready with plan B.

The loop closers

The alert system needed a backstop to ensure that patients didn't fall through the cracks. That challenge fell into the lap of Domingo, the project manager who had to figure out how to ensure patients would show up for their next test.

The first line of defense was a dedicated team of nurses tasked with following up with patients if the ordered test was not completed within a certain number of days. Given the difficulty of reaching patients by phone, however, they needed another option. The idea was floated of sending a letter to patients by mail, but some physicians worried that a notification of a suspicious lesion would induce panic, triggering a flood of nervous phone calls.

"The letter became one of my passions," Domingo said. "It was something I really pushed for."

The wording of the letter was especially tricky. She reached out to Northwestern's patient advisory councils for input. "There was overwhelming feedback that we should alert them that there was a finding that may need follow-up," she said. But a suggestion was made to add another clause noting that such findings are not always serious and may just require additional consultation. The letter is now sent to patient's within seven days of the initial AI alert to physicians.

"From the limited number of complaints we've gotten," Domingo said, "this was an important piece to help improve patient safety."

Since the onset of the project, the AI has prompted more than 5,000 physician interactions with patients, and more than 2,400 additional tests have been completed.

It remains a work in progress, with additional tweaks to ensure the AI remains accurate and that the alerts are finely-tuned. Some doctors remain skeptical, but others said they see a value in AI that wasn't so clear when the project started.

"The bottom line is the burden is no longer on me to track everything," said Cheryl Wilkes, an internal medicine physician. "It makes me sleep better at night. That's the best way I can explain it."

SPONSOR

NUANCE

Supporting clinicians and unlocking the future of patient care with ambient clinical intelligence

By Kenneth Harper

Ambient clinical intelligence (ACI) solutions already support clinicians by capturing patient-provider conversations and contextualizing them to create accurate clinical documentation. But in the future, ACI will go much further, analyzing thousands of voice characteristics to detect disease, injury, and mental illness—delivering powerful patient insights and decision support to clinicians at the point-of-care.

The healthcare industry is under enormous pressure, perhaps more than ever before. Research from the AHRQ found that clinicians report feelings of stress due to time pressures at work – they have limited time with patients, especially due to administrative tasks. One-third of clinicians said they needed more than 50% of additional time for patient care functions. Further, A recent Elsevier report showed that one in three clinicians are considering leaving their current role by 2024, with as much as half in some countries leaving healthcare for good.

Amid a worsening labor shortage, clinician burnout is rising, while patient expectations have never been higher.

One of the biggest contributory factors to burnout is the huge documentation burden placed on clinicians, but voice AI and ACI solutions are helping reduce that burden significantly.

Ambient clinical intelligence alleviates burnout and improves experiences
AI-powered speech recognition solutions have been widely used by clinicians for some time, enabling them to capture the patient story using only their voice. But more recently, ACI technology has taken those automatic documentation capabilities a step further.

ACI solutions securely record patient-provider conversations and automatically create accurate clinical notes directly in the EHR. This dramatically reduces the time clinicians must spend documenting care, helping to alleviate burnout.

Freed from the demands of manually documenting patient encounters, clinicians can focus instead on the patient in front of them, delivering a much better patient experience and rediscovering the joy of practicing medicine. Plus, the time saved allows clinicians to see more patients—without increasing their documentation burden.

But this is just the beginning for ACI. In the future, we'll use AI-driven ambient clinical intelligence to become clinically aware and do so much more on behalf of the care team and the patient.

What if we could do more with clinician-patient conversations to further support care delivery?

Speaking is the byproduct of a complex system. Every time we speak, we use our lungs, vocal cords, tongue, lips, nasal passages, and brain. And a disease, injury, or medical event involving any of these systems may leave diagnostic clues—biomarkers—that ACI solutions can detect in a patient's voice.

There are more than 2,500 biomarkers in the sub-language elements of human speech, and they can offer all kinds of insights into a patient's health and wellbeing.

For example, patients with Parkinson's have weak, soft voices, including characteristics such as breathiness. Patients with Alzheimer's use shorter words and more sentence fragments. And children with ADD speak louder and faster than their peers.

Right now, data scientists, researchers, and AI developers are working on ways to use sensory and signal data in patient voice samples to detect disease, injury, mental illness, and even environmental conditions. Their goal is to use ACI technology to give clinicians real-time patient insights and decision support that will fundamentally transform the way we deliver healthcare—and have a huge positive impact on patient outcomes.

The potential clinical applications for these AI-driven solutions are almost limitless, but I'm going to zoom in on two use cases that already show significant promise – detecting depression and anxiety and identifying social determinants of health.

Detecting depression and anxiety from a patient's voice

Several companies are now validating vocal biomarkers related to symptoms of disease. For example, one of our potential partners, Ellipsis Health, is using vocal biomarkers to give clinicians insights into patients' emotional state. The company's clinical support tool uses machine learning algorithms to measure and monitor the severity of depression and anxiety at scale by analyzing the words people say and how they say them—clinically validated vital signs for deep anxiety.

It's a great example of using voice-based AI to provide real-time, evidence-based clinical decision support. Just imagine the societal impact we can have if a clinician is given a proactive notification of depression symptoms and is then able to successfully intervene and get the patient the help they need, which would have gone untreated otherwise.

Identifying social determinants of health from a conversation

Another high-value use case for ACI in the future is identifying social determinants of health (SDoH)—factors like socioeconomic status, employment, food security, education, and community cohesion that can have a profound impact on healthcare outcomes.

Future ACI solutions may be able to capture SDoH insights in conversations and, when necessary, help mitigate the effects of SDoH on patient populations. By being aware of SDoH, clinicians and other stakeholders may be able to make better-informed decisions about patient treatment and support. And if healthcare organizations can identify and incorporate SDoH into patient care plans to treat the whole person and not just a disease, patient healthcare outcomes should greatly improve.

ACI opens an interconnected AI ecosystem and a world of limitless opportunities

Since the launch of Nuance's ambient clinical intelligence solution, the Dragon Ambient eXperience (DAX), clinicians have captured millions of patient voiceprints that are a goldmine for teams researching powerful new applications for ACI technologies. Over time, Nuance DAX will open a vast clinical intelligence ecosystem, where healthcare organizations can select from AI-powered solutions to help care teams improve care delivery and patient outcomes.

As we look to the road ahead, technological advancements like this will remove extraneous, manual tasks from clinicians' workflows, which will in turn remove the discrete functions that can contribute to burnout. While patient documentation can play a role in exacerbating burnout among providers – despite its critical importance to care – ambient sensing technology that securely "listens" to provider-patient encounters effectively lets the clinical documentation write itself. Instead of spending two hours documenting every one hour of patient care, providers are free to focus on what matters most: caring for people.

Experience DAX, innovated by Nuance and Microsoft, in action. Learn more about how AI-driven ambient clinical intelligence helps improve care delivery and outcomes.

STAT

HEALTH TECH

Can open datasets help machine learning solve medical mysteries?

By Katie Palmer | JANUARY 25, 2022

The medical data housed in patient records are a clinical researcher's dream: the key, potentially, to better tools to treat disease and screen with precision. They're also a computer scientist's nightmare: locked away in hospital systems, subject to restrictive data-sharing agreements, and often too messy to make use of.

A new open science project wants to accelerate ethical AI in medicine by doing the hard work of collecting and cleaning that data. Nightingale Open Science launched in December with $6 million in funding, led by Schmidt Futures, the philanthropy of ex-Google CEO Eric Schmidt and his wife, Wendy Schmidt. (It has no affiliation with Google's controversial health record-mining partnership with Ascension, which went by the code name Project Nightingale). It will freely share de-identified clinical datasets with researchers, linking medical images like X-rays, ECG results, and biopsy slides — 40 terabytes worth, to start — to outcomes from partnered health systems. Hundreds of researchers have signed up for access in its first month.

The nonprofit, co-founded by UC Berkeley ER physician and machine learning researcher Ziad Obermeyer and University of Chicago computation and behavioral science researcher Sendhil Mullainathan, also aims to shift the way researchers think about applying artificial intelligence to medicine.

Instead of aiming to reproduce clinical judgement with algorithms, each of the initial datasets are structured like a problem set, including data about a medical mystery that has stymied doctors — but may well be answerable by machine learning algorithms that see medical imaging differently. "It's really like there's this alien intelligence looking at this picture in totally different ways from the way a human looks at it," Obermeyer said.

STAT spoke with Obermeyer about how the project assembled and curated its datasets while maintaining privacy, and how it hopes to accelerate computational medicine.

What's the problem that you're aiming to solve by making these datasets open to researchers?

I think that like many researchers in health, I just spend an enormous amount of time negotiating for access to data. That's for a variety of reasons that are good — the health systems are very careful about privacy and you know, making sure that the data is being used for good and not for evil. But also there's just a lot of these frictions that mean it's really hard to get access to data.

I think that's a huge problem, because there's enormous value in the data sitting on these servers not being used except for billing and administrative purposes. We've had a lot of interest from foundations, and I think it's disproportionately foundations that are tech-adjacent, who kind of look around at other sectors of the economy and how technology and data are being applied very fruitfully and then they look at health care and they're like: What the hell?

How do Nightingale's datasets differ from what's currently being used in clinical machine learning research, other than being more accessible?

Let's say somebody waved their magic wand and gave you access to the entirety of the health dataset of a huge health system. What you do next is actually not entirely obvious. We have some models for this, because there's one really fantastic data set called MIMIC: It's like everything you could want from one hospital's ICU. It's generated a lot of papers, but it's actually not generated a lot of tools that people are even close to using in the hospital. Because it's just hard to structure the right question when you haven't spent a lot of time thinking about health.

I don't think you need to be a doctor to do this, but you need to make it easy for people to get pointed in the right direction. So one of the other things we do, besides making the data open, is try to curate the data sets around these really interesting and unsolved medical questions like sudden cardiac death, or silent heart attacks that don't get diagnosed, or cancer metastasis. All those problems are really great use cases for machine learning.

What does that curation process look like?

One of the first things we did, once we found systems that were willing to consider the arrangement, was spend a lot of time talking to them and asking what problems were important to them. In April or May of 2020, one of the systems we had an agreement with was Providence St. Joseph, and they were seeing some of the highest volumes of Covid patients. One of the issues that surfaced was how hard it was to figure out which patients with Covid were going to do OK, and which ones were going to deteriorate rapidly.

This is a great machine learning problem, because we have the X-ray, we have what happened to the patient — although that takes some stitching together of a lot of complex data sources.

That co-creation process was what led to the dataset on Covid: It's X-rays from a bunch of patients who tested positive for Covid, linked to what happened to the patients in the ER. Did they basically do fine and recover, or did they really go south in the few days afterward?

One thing that's de-emphasized is the clinical judgment of the X-ray at the time it was taken — you're focusing on the outcome.

There's certainly a bunch of people doing work that links to patient experience or outcomes, but it's not as common as I think it should be. That's in part because it's a lot harder to get those outcome data than it is to get doctor's interpretations. It's a lot easier to go to a hospital system's radiology information system and extract the images and what the radiologist said about the images because they live in the same place. So a lot of the datasets that are floating around are like that: ECG plus what the cardiologist said about the ECG, digital pathology plus what the pathologist said.

Curating those other outcomes requires a ton of extra work and knowing where to look. What does a missed heart attack look like if a doctor misses a heart attack? What happens to that patient, what kind of outcomes did they have? You need to kind of triangulate across laboratory studies and test results and procedure codes and all these other things. It's not impossible; it just requires a lot more of that detective work.

What have been the most complicated elements of that detective work?

There's one really big part which actually is now invisible. It's the part where we had to talk to a lot of health systems and get rejected by a lot of health systems before we found ones that wanted to partner with us.

We were very lucky that our early funders, and specifically Schmidt Futures, Eric Schmidt's foundation, went out on a limb, because they supported us before we had signed agreements. That was about a year of a lot of work by a lot of people trying to find health systems that shared our vision around unlocking the value of this data in a safe and ethical way.

Then there's a bunch of work that goes into not just the curating and the abstract process of coming up with a dataset, but the reality of deciding on the data elements, cleaning it, extracting it. This is part of the philosophy: We're not just dumping the data on you, we've done a ton of quality control and made sure that the data that we're giving you are both answering the set of questions we set out to answer, but also including enough other interesting stuff that you can branch out and ask a lot of other interesting questions.

Why focus these datasets on medical images, as opposed to other sources of clinical data?

One of the reasons why we focus on imaging data is because under HIPAA safe harbor, imaging is not one of the identified patient data types. So the imaging choice was partially driven by the fact that images are easy to de-identify while still preserving so much richness that keep them interesting from an algorithmic point of view. It's very unlikely that human doctors are getting all of the information that there is out of an image, but at the same time it's, under HIPAA, not identified patient information. So as long as we link it to a handful of variables and not like, a million variables that constitute the whole of the EHR, then we're on the right side of the law.

They're also the most commonly collected things in our healthcare system, and they're the kinds of things, especially ECG, that are increasingly going to be put on wearable devices and accessible via smartphones and that can be pushed out to really new and exciting settings.

For example, one data set that we're working on in India right now is basically collecting mobile ECG from smartphones and linking those to health outcome data in really under-resourced areas where it's very hard to get to a doctor, but where you could imagine helping triage people in and out of the hospital.

What do you see as the major limitations of what you've collected and made accessible so far?

We're still skewed towards better-resourced institutions. We thought a lot about the diversity of the institutions that we were working with, and one thing that we're particularly proud of is that, even though we do have partners like Stanford and Providence who are at the frontier of tech plus health, we also have a partner that is a county health system. But even with a lot of our efforts to be very inclusive about the kinds of systems that we talked to, there's just going to be a bias based on where we were able to get agreements signed. That's a problem that we're hoping to make progress on over the next year or so.

Another limitation is that this isn't a data dump of 100% of the data that anyone could ever want. This is a pretty highly curated dataset around a specific question. We tried to build so that we have those continuing relationships with the health systems where, if there are really high priority issues that a lot of researchers are interested in, we can add those data elements. We want to be as responsive to what that community wants as we can.

What's the shortest possible timeline you can imagine for an impact to come out of one of these first curated datasets?

One dataset that comes to mind, because it's both really exciting from a scientific point of view, but also in some ways a well-trodden ground, is around breast cancer screening. We worked with Providence oncology, the cancer center in Oregon, and they have all of these breast biopsy specimens that were literally collecting dust on a shelf.

We were able to partner with both them and Hamamatsu, which is a company that makes slide scanners, to start physically scanning these specimens at very high resolution and then linking them to cancer registry data, to mortality data, to electronic health record data on treatments. I think we will learn a ton from just applying algorithms to ask the question: Which cancers are going to spread and which ones are not?

There are some really interesting papers that suggest that machine learning algorithms can see things in those slides that pathologists don't see, and that they even get signal from parts of the images that pathologists don't systematically look at.

This is really the dilemma of screening: We see a mammogram or we see a biopsy specimen, and we need to make the call: Is this woman someone for whom watchful waiting would be OK, or is this someone who needs surgical management and chemotherapy? I think that there's a lot more work that needs to be done, but I don't think it's so far-fetched to think about in the next five to 10 years, AI-augmented screening strategies that incorporate the output of an algorithm into a pathologist or a radiologist decision being deployed in a clinical trial.

STAT

HEALTH TECH

Medicine's first autonomous AI could prevent blindness due to diabetes — if it can reach those most in need

By Katie Palmer | **DECEMBER 16, 2021**

Once feared as a job-erasing technology devoid of clinical nuance, autonomous artificial intelligence in medicine is growing up. The Food and Drug Administration in 2018 gave its stamp of approval to the first automated AI screening system, an algorithm that can analyze retinal images to detect diabetic retinopathy. And starting Jan. 1, primary care doctors across the country can get paid more reliably for automated screenings of the vision-threatening condition, which ultimately impacts more than half of people with diabetes.

"That, in my opinion, is going to be a sort of game-changing moment," said Aaron Lee, an associate professor of ophthalmology at the University of Washington, as a national Medicare reimbursement rate helps more primary care practices decide whether the technology is worth investing in.

As the technology matures and adoption widens, clinical AI developers can look to diabetic retinopathy screening as a case study in what it takes to implement automated technology in the real world.

It has cleared many of the hurdles that have tripped up AI: It has demonstrated good results in prospective trials, and appears to perform accurately across different races, ages, sexes, and ethnicities. But the technology still requires a massive amount of buy-in: from health systems that need to shell out for the tech, from providers that need to add yet another task to their workflow, and from patients who must be convinced to take extra time during a routine primary care appointment to have their eyes scanned by an AI.

The first system from IDx, which recently changed its name to Digital Diagnostics, will soon be in more than 1,000 clinics in the U.S. Eyenuk, which received clearance for its own automated screen in 2020, has 70 clients so far. "We've really seen the revenue inflection point here in the last six months, where we've gone from, 'Hey, this is a cool tool, let me try it out a little bit,' to system-wide deployments," said John Bertrand, CEO of Digital Diagnostics.

The hope is that AI-driven systems can improve low rates of screening for diabetic retinopathy, the leading cause of blindness in the U.S. working age population. Well over half of diabetes patients miss their recommended annual screenings, which can help identify the eye disease before high blood glucose levels irrevocably damage blood vessels in the retina. The tools can enable primary care offices to screen patients during routine checkups: No need for a separate eye specialist appointment, no waiting for the results to be transferred, and fewer opportunities to put off follow-ups if the screen comes back positive.

To have a meaningful impact, though, the tools need to reach the patients who are missing their screenings. Among the most vulnerable populations, including those with geographic and financial barriers to screening, only 10% to 30% of patients may receive annual diabetic eye exams. "Many health systems that have the resources to deploy these technologies are not necessarily in those settings," said David Ramsey, director of ophthalmic research at Boston's Lahey Hospital & Medical Center.

The new reimbursement rates are one important step to getting screens to patients who don't receive them regularly, as are incentive payments from insurers and government programs to providers that meet certain screening thresholds among patients.

But the value proposition — and in turn, the likelihood of adoption — can vary dramatically from practice to practice, in part because the system relies on dedicated equipment, called a color fundus camera, to take photos of the back of the eye. "You have to have enough diabetics with reimbursement to justify the expense of the camera," said Lee Herman, whose practice in suburban Atlanta, one of the first to implement the IDx system, serves more than 600 patients with diabetes. In the U.S., the automated systems are only approved for use with one or two specific cameras that cost in the range of $15,000 to $25,000.

While both companies have systems in use at federally qualified health centers, many of the earliest adopters of these technologies have been academic centers, or provider groups that have both primary care and ophthalmology under the same umbrella, allowing eye specialists to focus on referrals for patients they can actually treat. Still others are testing the system alongside typical eye exams, or using it as part of a hybrid telescreening approach for other eye diseases.

"I think that [these technologies] can increase access," said Ramsey. "But our system is not necessarily designed to put cameras intrinsically where they will do the most good." Digital Diagnostics, recognizing this barrier, has built the cost of a camera into its subscription, allowing customers to pay off the cost of the device over the course of a two- or three-year contract, and making it easier for cash-constrained operations to get the program off the ground.

In the future, new validated camera models, or even hardware-independent algorithms for grading retinal images, could create more affordable options.

But for now, "the FDA's approach has been to ask companies to prospectively collect data" on other cameras, said Eyenuk CEO Kaushal Solanki — making it an expensive, time-consuming endeavor to validate use in each additional model.

If a primary care office has a clear financial case for AI diabetic retinopathy screening, its next hurdle is one shared by many algorithmic point solutions. While the tools stand to improve efficiency and accuracy, most practices don't have the human or technological bandwidth to implement the tools, as highlighted by a recent study of an uncleared deep learning-based screening system developed by Google Health.

"Originally the questions always were about, 'Is this safe? Can I trust this?'" said Michael Abramoff, the University of Iowa ophthalmologist who developed the Digital Diagnostics screening algorithm. Now, "it's more, 'How can I fit this best into the workflow? How can I make this efficient?'"

Some of the barriers are technical. "There's so, so much that goes into starting one of these," said Risa Wolf, assistant professor of pediatric endocrinology at Johns Hopkins Hospital who led the integration in pediatrics. That included vetting the system's safety and privacy — a special consideration with eye images, which are considered biometric identifiers. (The systems are not FDA-cleared for use in children, but Wolf is conducting approved research into their application in pediatric populations.)

And while taking usable retinal images can take as little as five minutes if the patient's eyes don't need to be dilated, that isn't always the case — as Baltimore-based Johns Hopkins Community Physicians discovered in a pilot program of the Digital Diagnostics system. Older patients are more likely to need dilation, which requires waiting for about 15 minutes while eye drops take effect.

Older patients are more likely to need dilation, which requires waiting for about 15 minutes while eye drops take effect. People with darker skin are also more likely to have darker irises, which sometimes require dilation to get good retinal images, explained Alvin Liu, a Johns Hopkins ophthalmologist.

"The barrier has been partially patient acceptance," said Scott Feeser, medical director for the system's 30-some primary care offices, "especially when they weren't expecting to get their eyes dilated when they came into the office — they're coming in just for the regular checkup."

If patients or practices aren't prepared to take the extra time for dilation, some at-risk patients may be more likely to walk out the door without getting screened. So Liu is working on a system to predict a patient's likelihood of requiring dilation. "That should make the workload better," he said. But custom-developing that kind of system may not be possible outside a deep-benched network like Johns Hopkins.

To help other providers make the call, Liu is planning a prospective study to measure the impact of an AI install on screening rates in the Johns Hopkins primary care network as the pilot project expands; Wolf is enrolling for a randomized controlled trial to answer the same question in her pediatric population. "Because we know that our minority, low-income youth are less likely to get screened, we want to see if autonomous AI can level the playing field," said Wolf.

At the first four primary care sites in the Johns Hopkins pilot, Feeser said they've seen screening rates go up by about 20%, compared to 10% at practices that simply doubled down on patient referrals to eye doctors for annual screenings and retrieving records from those appointments. "It's an important piece of getting these scans done," he said. "But it's not a miracle."

As more data comes in, it may become clearer how diverse primary care systems can successfully implement automated screening programs – whether or not it makes sense to buy a camera outright. "Maybe we'll put one on a truck, and say, 'Hey, on Tuesday the camera's going to be here,'" said Feeser.

In the meantime, the University of Washington's Lee, who conducted a head-to-head study of automated algorithms in the VA system, thinks it's OK to wait for the dust to settle. "I actually go around conferences and I try to tap the brakes on people adopting these systems," he said. He worries that patients could end up missing other eye diagnoses that a specialist could pick up on in a typical exam, and recommends that any practice considering the model test it in their patient population before deploying without clinician oversight.

"How do we use these new innovative technologies to really mitigate disparities in diabetes outcomes?" asked Wolf. "I think this is just one piece of the big puzzle of how it can hopefully improve screening access, and really address the entire population, not just the ones who actually make it to the eye doctor."

STAT

HEALTH TECH

Artificial intelligence is making a pitch to transform radiology. Will it pay off for hospitals?

By Katie Palmer | DECEMBER 2, 2021

For radiologists working today, the specter of artificial intelligence is inescapable.

In the past year, venture capitalists have continued to invest significantly in startups developing AI for medical image analysis and support, with some groups projecting a $20 billion market by 2031. But despite the proliferation of research and investment, AI products are still a hard sell for many radiology practices — even those at academic centers leading clinical research into deep learning. The tools are still in the middle of the hype cycle for any new technology: Inflated expectations are giving way to skepticism and barriers to adoption.

"We are right now smack dab in the trough of disillusionment," where applications are expanding but revenue is still hard to come by, said Paul Chang, vice chair of radiology at the University of Chicago School of Medicine, in a discussion about the business case for AI in radiology this week at a meeting of the Radiological Society of North America in Chicago.

The popularity of AI tools in radiology — and the pitch some radiologists and AI developers alike are making for them — was on full display this week at the conference, where the exhibit hall hosted an AI showcase and half of the scientific sessions invoked AI or deep learning.

Chang and other radiologists at the session said that for AI to climb out of its current trough, business leaders at academic centers and private practices need convincing that they'll get a significant return on their investments in new software and the infrastructure to run it. And the footholds they find are unlikely to look like the dramatic image-crunching skills that have drawn the greatest attention.

"When everyone started hearing about AI, all the vendors were doing image interpretation type tasks — detection, maybe diagnosis," said Nina Kottler, associate chief medical officer for clinical AI at Radiology Partners. Radiologists, like any highly skilled clinicians, are expensive, and anything that could cut down on their workload seemed like a decent target for automated technology. An AI that could count the number of lung nodules in a chest CT, instead of forcing a radiologist to manually count them, could potentially improve accuracy and also allow a radiologist to spend more time on tasks that take more specialized knowledge.

That bias still exists among researchers, too. Chang noted that even in the exhibit hall and at the papers presented at the conference, there remains a significant emphasis on image-based AI tools.

But the incremental improvements provided by those types of programs — so far — don't seem to be appealing as much to the decision-makers in charge of choosing which AI systems to invest in, said Hari Trivedi, co-director of the Health Innovation and Translational Informatics lab at Emory University. Trivedi's perspective pulls from his interviews with two dozen peers in radiology about the drivers of AI adoption.

"If I had to pitch an AI model I would want to adopt, I would focus less on things that improve incremental efficiency," he said.

At Radiology Partners, a national practice of onsite radiology clinics that conducts about 10% of imaging in the U.S., early tests of AI for visual detection in images didn't pay off, said Kottler. "Maybe you get a 10% — not even, maybe a 5% — ROI on that."

That's not to say the practice abandoned AI. It just implemented it in different places. The first successful AI implementations at Radiology Partners, which employs 3,000 radiologists across the country, used natural language processing to help clinicians dictate their reports, followed by other NLP systems that streamline radiologist workflows. With those applications, Kottler said, "it's like a 90% ROI. It's a total game-changer."

Any AI implementation with an obvious impact on a hospital system's bottom line will be an easier sell — especially those that keep more patients coming in for imaging, which has high reimbursement rates. "If we can catch additional patients, retain them, and simultaneously improve quality, that's what people are getting excited about," said Trivedi. "Of the hundred patients a month that are supposed to have a follow-up scan right now, my guess is less than half of them are coming back. If you implement a model that closes the loop on follow-ups … you are now generating tens of thousands of dollars of additional revenue for that hospital system."

Another practical challenge for providers looking to invest in new AI systems is algorithmic overload. Emory's radiology department already has 140 different pieces of software in place, said Trivedi — so "it's almost a nonstarter if I say I want this single model for lung nodule detection, or intracranial hemorrhage detection." Integrated systems for multiple AI tools will be necessary for many providers to buy in. "Nobody wants single solutions anymore. Nobody wants to hunt and peck," said Trivedi. "It's got to be a unified platform."

Building that kind of infrastructure takes time, both to get disparate clinical systems to work together and to navigate the bureaucracy of provider systems. "Depending on how funding and governance is laid out at your institution, there can be a thirst for AI, but the mechanism or the means to actually do that implementation and get it across can be a multiyear process," said Trivedi. Despite his academic team's leadership in the development of algorithmic tools, he said Emory has yet to use AI in its radiology department. "About 12 months from now, we will have something in the works. We're very early."

As more radiology practices decide to give AI tools a go, the panelists anticipated a winnowing of the field — both in terms of individual solutions, and the companies that provide them. When people talk about AI, they often believe it will make everything easier and faster, said Luciano Prevedello, associate chief clinical information officer at The Ohio State University Wexner Medical Center. But some of the algorithms he has implemented have done the opposite. "We had to remove one of the algorithms that we've implemented," he said, "because it was really not adding very much value for us. We were spending more time reviewing the results of the tool than if we interpreted by ourselves."

The number of companies selling AI tools is also heading for a significant consolidation, the panelists all agreed. "If you sort of lift up the covers and you talk to some vendors more deeply, you'll hear that some vendors are going to other vendors and saying, 'Will you buy me?'" said Kottler. "They're just not making money, there's not enough dollars to go around to support all of them."

A spate of mergers among health tech companies could create the unified platforms hospitals want. "I think what the future is going to be is consolidation of the platform — a way to connect multiple players into one single place," said Prevedello.

But consolidation has risks, too: "We will have consolidation, we'll have buyouts of small companies by big companies, but I don't think that we necessarily have to expect, and I hope it doesn't happen, where everything is provided by you know, two or three vendors. I don't think that will be a good thing," Trivedi said.

When consolidation starts happening at a rapid clip, though, it will be a sign that the business of artificial intelligence in radiology is crawling out of its trough, said Chang. "We're going to get there, one way or another."

STAT

FIRST OPINION

Health-related artificial intelligence needs rigorous evaluation and guardrails

By John D. Halamka, Suchi Saria & Nigam H. Shah | MARCH 17, 2022

Algorithms can augment human decision-making by integrating and analyzing more data, and more kinds of data, than a human can comprehend. But to realize the full potential of artificial intelligence (AI) and machine learning (ML) for patients, researchers must foster greater confidence in the accuracy, fairness, and usefulness of clinical AI algorithms.

Getting there will require guardrails — along with a commitment from AI developers to use them — that ensure consistency and adherence to the highest standards when creating and using clinical AI tools. Such guardrails would not only improve the quality of clinical AI but would also instill confidence among patients and clinicians that all tools deployed are reliable and trustworthy.

STAT, along with researchers from MIT, recently demonstrated that even "subtle shifts in data fed into popular health care algorithms — used to warn caregivers of impending medical crises — can cause their accuracy to plummet over time."

Experts have been aware that data shifts — which happen when an algorithm must process data that differ from those used to create and train it — adversely affect algorithmic performance.

State-of-the-art tools and best practices exist to tackle it in practical settings. But awareness and implementation of these practices vary among AI developers.

Also variable is adherence to existing guidelines for development and testing of clinical algorithms. In a recent examination of AI algorithms provided by a commercial electronic health record system vendor, most of the recommendations from such guidelines were not reported. Just as concerning is the fact that about half of AI development and testing guidelines suggest reporting technical performance (how well the model's output matches truth on one dataset) but do not address fairness, reliability, or bottom-line usefulness of the algorithms.

Without rigorous evaluation for accuracy, safety, and the presence of bias, AI developers are likely to repeat mistakes similar to those documented in a classic study by Ziad Obermeyer and colleagues, in which a poorly chosen outcome — using health costs as a proxy for health needs — during algorithm development led to major racial bias.

For nearly a year, we and many other colleagues from academia, industry, and government have convened to discuss ways to overcome these challenges. Among the many perceptive observations offered by the group, a number of them stand out as actionable suggestions:

Create a label for every algorithm — analogous to a nutrition label, or a drug label — describing the data used to develop an algorithm, its usefulness and limitations, its measured performance, and its suitability for a given population. When you buy a can of soup, you decide if the calories, fat, and sodium align with your needs and preferences. When health systems decide on a drug to use, a medical review board assesses its utility. The same should be true of AI in health care.

Test and monitor the performance of algorithm-guided care within the settings in which it is deployed in an ongoing way. Testing should include screening for potential demographic-specific losses in accuracy with tools that find error hotspots that can be hidden by average performance metrics.

Create best practices for establishing the usefulness, reliability, and fairness of AI algorithms that bring together different organizations to develop and test AI on data sets drawn from diverse and representative groups of patients.

Create a standard way for government, academia, and industry to monitor the behavior of AI algorithms over time.

Understand clinical context and goals of each algorithm and know what attributes — quality, safety, outcomes, cost, speed, and the like — are being optimized.

Learn how local variations in lifestyle, physiology, socioeconomic factors, and access to health care affect both the construction and fielding of AI systems and the risk of bias.

Assess the risk that AI might be used, intentionally or not, to maintain the status quo and reinforce, rather than eliminate, discriminatory policies.

Develop approaches for appropriate clinical use of AI in combination with human expertise, experience, and judgment, and discourage overreliance on, or unreflective trust of, algorithmic recommendations.

The informal dialogues that yielded these observations and recommendations have continued to evolve. More recently, they have been formalized into a new Coalition for Health AI to ensure progress toward these goals.

The steering committee for this project includes the three of us and Brian Anderson from MITRE Health; Atul Butte from the University of California, San Francisco; Eric Horvitz from Microsoft; Andrew Moore from Google; Ziad Obermeyer from the University of California, Berkeley; Michael Pencina from Duke University; and Tim Suther from Change Healthcare. Representatives from the Food and Drug Administration and the Department of Health and Human Services serve as observers in our meetings.

We are hosting a series of virtual conferences to advance the work over the next few months followed by an in-person conference to finalize the material for publication.

The coalition has identified three key steps needed to pave the path toward addressing these concerns:

- Describe consistent methods and practices to assess the usefulness, reliability, and fairness of algorithms. Tech companies have developed toolkits for assessing the fairness and bias of algorithmic output. But everyone in the field must remain mindful of the fact that automated libraries are no substitute for careful thinking about what an algorithm should be doing and how to define bias.

- Facilitate the development of broadly accessible evaluation platforms that bring together diverse data sources and standard tools for algorithm testing. Currently, there are no publicly accessible evaluation platforms that have both data and evaluation libraries in one place.

- Ensure that robust and validated measures of reliability, fairness, and usefulness of AI interventions are incorporated into clinical algorithms.

By working together as a multi-stakeholder group and engaging policy makers, this coalition can develop the standards, guardrails, and guidance needed to enhance the reliability of clinical AI tools. By earning the public's confidence in the underlying methods and principles, they will be assured that the humanistic values of medicine remain paramount and protected.

STAT

OPINIONS+

Forget the hype and embrace the reality: How AI can improve health outcomes right now

By Prasad Dindigal | DECEMBER 21, 2021

The story of artificial intelligence (AI) driving better health care outcomes has been a convoluted one, with hype-laden chapters on algorithmic cures for cancer and the future of robots replacing doctors. As those promises eventually proved overly ambitious, many people have lost the plot.

When the world fixed its collective gaze on the Covid-19 pandemic and AI innovation fell off the hype cycle, it did not die. Instead, it has quietly emerged as a critical link between patients, providers, and payers by helping identify gaps in care, guide strategic decision-making, and improve patient engagement with care managers and primary care providers. Bots may not have replaced clinicians, but they have emerged as an important link in the care-management process.

Take, for example, the widespread deferral of routine care during the pandemic. An estimated 41% of U.S. adults delayed or avoided medical care, including urgent or emergency care, at some point between March and June 2020. In many areas of health care, these rates have still not bounced back.

This isn't a new problem. Even before the pandemic, only half of people who were prescribed medications for chronic conditions like diabetes or heart disease ended up taking them regularly.

When disrupted by Covid-19, things got worse. Forced into makeshift work-from-home settings and unable to rely on many of their traditional outreach touchpoints, many workplace care management initiatives and disease management programs struggled to stay connected with patients and keep them following prescribed care regimens.

Recent advances in real-world patient data, AI-enabled analytics, and digital engagement strategies based on behavioral science have made it possible to execute highly personalized strategies that keep care managers in regular contact with their patients to make sure they stick with their medications and recommended care. Used properly, technology is helping providers, payers, and self-insured employers recommend treatment for at-risk patients earlier, engage deeper, and nudge patients to the best possible outcomes.

My company, EXL, recently worked on a project with a large, self-insured employer to reduce inefficiencies by driving increased use of preventive care for patients at risk of cancer. The example provides a look at how AI can play a key role in identifying the right patients and help them stick to a care plan by delivering a steady flow of follow-ups, prompts, and guidance.

Stratify the population: We used AI-powered analytics to dig into medical and pharmacy claims, provider data, member enrollment data, and sociodemographic data to get a complete picture of the health care experience of each individual within the employee population.

Analyze behavior: This stratification process allowed us to identify the patients most at risk of being diagnosed with cancer. Once they were identified, the patterns of clinical engagement and health care utilization they experienced over time could be tracked. Through this process, we were able to prioritize employees who were most at risk of being diagnosed with cancer based on a combination of their previous medical histories, patterns of under-utilization of care or deferred care, and behavioral risk factors.

Optimize outreach: Once employees were assigned cancer risk scores, they could be flagged for further intervention based on their specific needs. In addition to being used in the data-analysis phase, AI was also used in the patient outreach process to power automated, programmatic patient communication via e-mail, text message, and speech-enabled applications that respond to the human voice and text prompts in real-time. This omni-channel approach helps care managers efficiently reach their patients outside of the confines of traditional corporate channels. It also logs activity, so care managers know when further, one-on-one human outreach is required for people who are either ignoring or missing the automated outreach.

With 100,00 employees working from homes around the world during the pandemic, this employer needed a systematic way for its care-management team to connect with patients that was far more personalized than a generic e-mail or form letter, but also more scalable than having nurse care managers phone each person.

The Goldilocks solution came in the form of a series of outreach escalations starting with AI-powered chat, e-mail, and programmatic outbound calling, building toward nurse care manager dialogue when at-risk patients were not following care guidelines or responding to automated prompts.

As health care providers and payers continue to shift to value-based models, which incentivize the use of the most effective therapies, it is essential that care management protocols are designed to help patients follow the path to optimal outcomes. AI can help that happen by improving the accuracy of patient risk stratification, facilitating more frequent contact with patients, and reducing costs associated with manual, labor-intensive processes.

For a behavioral approach like this to work in the real world, it must marry AI and nurses or other care managers to build patient engagement plans that actually prompt patients to follow through with recommended care and consistently check in to make sure they are staying on course.

From conversational AI that can recognize patterns of speech to text messages to old-fashioned, one-on-one contact with human nurse care managers, the gradual integration of AI-powered technologies into the care management workflow is making it possible to deliver more personalized, impactful patient engagement at scale.

While that may not sound like the amped-up vison of AI prognosticators had in mind five years ago, it represents a tremendous step forward in the development of practical, real-world solutions. It's time to embrace not what we hope AI might do down the road but what it can do now. Because the capabilities are powerful, they can have a substantial impact, and, if leveraged properly, they will help save lives.

STAT

EXCLUSIVEV

Alphabet has a new drug discovery company building on DeepMind's AI chops

By Katie Palmer | NOVEMBER 4, 2021

Alphabet is plunging into the drug discovery business.

Over the last year, the tech behemoth and Google parent company has made a large and unlikely splash in the world of biology. DeepMind, its AI research outfit, wowed structural biologists last November by cracking the longstanding problem of predicting protein structure with its deep learning model, AlphaFold2. Eight months later, it amplified the impact of those discoveries by releasing both the model's code and a database of more than 350,000 predicted protein structures to the public.

At the time, DeepMind CEO Demis Hassabis said he hoped the protein structures would be used to accelerate work by drug discovery companies. But the wheels were already in motion for him to head up his own.

Thursday, DeepMind is announcing the launch of Isomorphic Laboratories, a new Alphabet company that aims to leapfrog off the success of the protein-folding work to apply deep learning methods to drug discovery. At least for now, Hassabis will serve as CEO of both DeepMind and the new company, which will be based in London and plans to hire in the range of 30 people in its startup phase.

"You can think of it as a sort of sister company to DeepMind," Hassabis told STAT. DeepMind was acquired by Google in 2014 and turned a profit for the first time in 2020. "The idea is to really forge ahead with the potential for computational AI methods to reimagine the whole drug discovery process," he said.

Companies have been deploying artificial intelligence against different challenges in drug discovery for years, and the last year has seen significant investment in both new and longstanding efforts from companies including Exscientia, Relay Therapeutics, and Valo Health. Pharma giants like Genentech, Pfizer, and Merck are also investing in their own modeling capabilities. But Hassabis said Isomorphic Labs will tackle the problem of drug discovery differently than many existing companies.

"A lot of pharma companies did come to us to discuss AlphaFold and how they could use it, and through that we understood a lot more about what other things they would need," said Hassabis.

Isomorphic will focus on building predictive or generative models of biological phenomena to fill those gaps, using computers to both anticipate how drugs will perform and potentially design novel molecules. "I think a lot of AI companies use AI for analytics," said Hassabis. "You have a lot of data of some form, genomics or something like that, and you're trying to find a correlation or an insight in that data. I think that's very valuable, and we'll be doing that too. But we're going to go beyond that."

Rather than developing its own pipeline of drug candidates, Hassabis said, the company may aim to sell its platform of models as a service.

"I think that building models of fundamental biological processes is a strategy that makes sense," said Amit Deshwar, senior director of predictive systems at Deep Genomics, a company that uses deep learning to identify RNA therapeutic candidates and targets.

For DeepMind, the idea of launching a drug discovery company has been in the background for a while, with discussions ramping up in the last couple of years as the company's protein folding project gained steam. "We could see AlphaFold was working, and for me that was always a lighthouse project that I was waiting for," said Hassabis. "If that works, a lot of these other ideas that are much more early stage have the potential to work as well."

Hassabis declined to go into the details of those early-stage projects, which will initially be focused on internal research more than explicit product development. But it's clear that while Isomorphic Labs may have been kickstarted by protein folding, it won't be limited to that small piece of the puzzle.

"That's a huge step forward, that AlphaFold is able to come up with a valid structure for a protein that people have been trying to get for ages," said Ben Perry, discovery open innovation leader at the Drugs for Neglected Diseases Initiative, which collaborated with DeepMind on protein structure prediction before the AlphaFold database was made publicly available. "And then we suddenly get stuck back in the sludge of normal drug discovery." Challenges such as designing and determining how to manufacture small molecules, and predicting the physical and metabolic effects of binding to protein sites "are a lot harder than protein folding," said Perry.

What parts of the drug discovery process might be particularly ripe for DeepMind's deep learning approach? Hassabis identified protein-protein interactions, small molecule design, binding affinity, and toxicity analysis as potential targets for predictive models. But the success of any of Isomorphic's research lines will depend on a number of external factors.

The first is the availability of training data. "One of the key learnings of deep learning has been that you need a ton of data," said Deshwar. AlphaFold was trained on an open source protein data bank, filled with hundreds of thousands of experimental structures generated in labs around the world.

But applied data on drug candidates and their performance is often siloed in proprietary databases.

"There are pharma companies now who are just spending enormous amounts of money generating that data internally so they can perfect their own algorithms," said Perry. "It'll be interesting to see how they [Isomorphic] go about accessing that data," possibly through outright purchase or partnership with biotechs.

"The second aspect is are you building a model that will actually solve a bottleneck in drug discovery," said Deshwar. "And there can be a bit of tension there. If data is widespread and easy to generate, you can imagine a situation where it's easier to do things experimentally rather than predict it."

One bottleneck to target could be predicting how, and how well, a protein binds to a potential drug. "For me, the next step, which is arguably more complex, is once you've got a structure, now you've got to get good at actually predicting how molecules bind to it," said Perry. "And the world's rubbish at that."

If Isomorphic chooses to dive into the docking problem, it may encounter more competition. Now that AlphaFold has taken the urgency out of the longstanding protein-folding contest, which is called CASP, structural biologists are looking for the next global benchmark. Perry is part of a group that proposed an initiative called CACHE, which would pit small molecule hit-finding algorithms against each other.

Isomorphic's most immediate task will be in staffing up a multidisciplinary group of deep learning experts, computational biologists, medicinal chemists, biophysicists, and engineers. It won't be alone under the Alphabet umbrella: The company's aging-focused subsidiary Calico Labs, which has its own drug development arm, has more than 30 of its own job openings.

"I don't think they're going to have a problem recruiting, when you think about the company sitting behind them," said Perry.

STAT

HEALTH TECH

'It's going to set the bar pretty high': A path forward for trusted AI in breast cancer risk prediction

By Katie Palmer | NOVEMBER 16, 2021

As the role of artificial intelligence grows in medicine, one of the leading concerns is that algorithmic tools will perpetuate disparities in care. Because AIs are trained on health records reflecting current standards of care, they could end up parroting bias baked into the medical system if not carefully designed. And if algorithms aren't trained and tested on data from diverse populations, they could be less effective when used to guide care for poorly-represented subsets of patients.

So some AI development groups are tackling that problem head on, training and testing their algorithms on diverse patient data to ensure they can apply to a wide range of patients — long before they're deployed in the wild.

One such group is at MIT, where researchers have previously shown that a model called Mirai — developed to predict a patient's five-year breast cancer risk by interpreting mammograms — could outperform an existing tool. In a paper published this month in Journal of Clinical Oncology, they have now expanded test sites for the algorithm to seven hospitals around the globe — in Taiwan, Sweden, Brazil, Israel, North Carolina, Georgia, and the model's originating site at Massachusetts General Hospital — to see how it performs in patient populations with a wide range of clinical and demographic backgrounds.

The paper's authors claim it is the "broadest validation to date of an AI-based breast cancer model." External validation isn't uncommon in AI research, but the standard number of test sites in the field is closer to two, said Olivier Elemento, director of the Englander Institute for Precision Medicine at Weill Cornell Medicine. "I personally think that there's never enough validation," he said. "It's the only way to understand the real world performance of these models."

Despite its importance, this kind of cross-site testing is rare, because it is painstaking, time-consuming work. "Validation was a really intense effort for us," said Adam Yala, a Ph.D. student at MIT in Regina Barzilay's lab and lead author on the study. "We reached out to well over 40 groups, many many dozens of meetings, invited talks and cold emails, to eventually find our partners."

And cleaning, accessing, and analyzing patient data from those diverse operations is no easy feat. Some facilities will share their data with collaborators after careful work to anonymize patients; in other cases, medical data can't leave the hospital's physical site at all.

To comply with patient data protections at Taiwan's Chang Gung Memorial Hospital, Yala had to fly to Taipei. "The only way to validate Mirai on their data was to physically show up, directly login into a machine and do the work on premise," he said. The trip almost didn't happen: As he was preparing for his trip in February 2020, Covid-19 border restrictions were beginning to kick in, and he had to reroute his flight through San Francisco instead of Hong Kong, which had just changed its quarantine policy. After scrounging a few masks from a paint store in Cambridge, Yala sat in a room near the hospital's MRI machine for three days while Mirai worked away at the numbers.

Yala had assembled that dataset for a study of Mirai published in January, along with patient information from another in-person trip to Karolinska University Hospital in Sweden in December 2019.

To complete the global tour for the latest analysis, he added data from sites in the U.S. at Novant Health and Emory University, along with Maccabi-Assuta in Israel and Barretos Cancer Hospital in Brazil.

"This validation is not for the faint-hearted," said Judy Gichoya, who worked with Hari Trivedi at Emory to incorporate the system's carefully annotated breast cancer data. "It's going to set the bar pretty high for AI studies in the future," said Elemento.

As in the last study, Mirai performed well compared to a widely-used risk assessment model called Tyrer-Cuzick, which inputs patient demographics, family history, and breast density to calculate its predictions. Mirai's C-index — a measure that generalizes the model's area under the curve over time — was similar at all seven test sites, ranging from 0.75 to 0.84. In practice, the model could be used to identify high-risk patients for further MRI screening. "We specifically looked at existing high risk guidelines that we have to decide who gets MRI," said Yala, "and we consistently beat that by a large margin."

Importantly, Mirai performed similarly across white and African-American patients at Emory, an especially critical competency when Black women are significantly more likely to die from breast cancer at any age. "If you look at the history of Tyrer-Cuzick, it was mainly validated for white patients," said Gichoya. Previous work had showed Mirai outperforming Tyrer-Cuzick in assessing the breast cancer risk of Black patients at MGH — and that performance held up at Emory, whose unique dataset has a relatively high representation of people who identify as Black.

The results offer a model in responsible development of AI-based predictive tools, which are not always tested on diverse datasets before deployment. But Mirai is still far from implementation beyond the walls of MGH, where it was used during the pandemic to encourage specific high-risk individuals to come in for screening after the pandemic led many patients to put off their scheduled mammograms.

"I think it's a really fantastic finding that we see this performing O.K. across multiple data sets — way better than existing clinical models that we use, and performing actually way better for special populations like Black women," said Gichoya, an assistant professor of interventional radiology and informatics at Emory. "But the problem is, imagine you're a radiologist and you're using this. Now that I'm using this type of model in my clinical work, what does that mean?"

To start answering that question, Mirai will need to be tested in prospective clinical trials. All of Mirai's analysis so far has looked back in time to see if it fits patterns of existing patient data that were not used to train the original model. "To make this tool useful for good, we have to do quite a lot of prospective validation," said Gichoya, "and even get different types of datasets to allow us to look at subpopulations — not just based on race, but other breast cancer types."

Yala said there are plans for a number of studies of Mirai's clinical performance — including a trial with a large hospital system in Mexico — and the free release of Mirai's code so hospitals can run their own studies and validate the system locally before they deploy it. As those years-long projects progress, Mirai may also be adapted to predict risk using mammograms from different machine manufacturers (the original model was only validated on images from one maker, Hologic) and three-dimensional tomosynthesis images.

If the model proves to accurately predict breast cancer risk in the real world, then its developers will have to tackle the thorny problem of how to translate an algorithmic risk calculation into clinical practice.

"There's an even bigger part of prospective validation in figuring out, how do you implement these models?" said Gichoya. "You have to think about how you communicate this risk to patients, because these models are black boxes still."

To build trust with radiologists and clinicians, she said, it will help to understand how Mirai makes its risk predictions — for the model's outcomes to be explainable. That's a particularly tall order for image-based predictive models, said Elemento. "These techniques right now give you pixels, and we need to translate them into words, semantic explanations," he said.

Explainability may end up having a significant impact on the acceptance of AI-based risk predictors by both providers and patients, especially those who have historically been mistreated because of systemic and institutional racism.

"All of our health care systems already capture the biases in some way or another, so how we effectively build systems off generally biased data is a really open challenge," said Yala. "How do you build technologies while making it so that what it's predicting isn't regurgitating the past? Finding the thing that's actually going to drive an equitable improvement is an open technical problem. But it's one we can study and we can solve."