ANDRZEJ WOJCICKI/SCIENCE PHOTO LIBRARY/GETTY IMAGES

BIG DATA TAKES ON CANCER

Data science and machine learning technologies are helping CANCER RESEARCHERS EXTRACT NEW MEANING from large clinical and molecular datasets.

Regina Barzilay wasn't looking to study cancer. She had made a name for herself as an artificial intelligence (AI) researcher at the Massachusetts Institute of Technology (MIT) in Cambridge, developing machine learning models to process and understand human language and unstructured text.

Then came her diagnosis of breast cancer, in 2014, at the age of 43. She was shocked by the paltry amount of data upon which her doctors based their clinical decisions. Maybe, Barzilay thought, algorithmic models like hers could extract more from the clinical records. If so, perhaps machine learning could detect tumours like hers at an earlier stage and offer personalized treatment recommendations.

Back in the lab, Barzilay put her ideas to the test. She adjusted her protocols to parse patient medical reports, and developed new deep learning methods to interpret diagnostic images. The adapted models have since proven their worth on retrospective datasets, and at least one of her tools has been implemented[1] in clinical practice as a diagnostic aid for radiologists.

The algorithms work so well that, had they been available, Barzilay suspects they may have helped doctors spot signs of her cancer a year or two earlier, possibly before the disease had spread to her lymph nodes. "By reducing uncertainty and truly personalizing patient care," Barzilay says, "machine learning can totally transform this area."

Delivering on that promise remains a challenge. In medicine today, there's so much raw clinical data generated — from the pathology lab and the imaging suite to the surgical ward and the oncologist's office — that it's rarely obvious how best to design and train algorithms to connect all the disparate threads of information for patients.

But if AI can teach a car to drive on its own or a social media platform to recognize faces, machine learning should be able to radically improve the diagnosis and treatment of cancer. It just needs to be applied in the right way, which is why *Nature* and the cBio Center in the Department of Data Sciences at the Dana-Farber Cancer Institute (DFCI) in Boston, Massachusetts, convened a two-day conference in October — spearheaded by DFCI Chief Scientific Officer Barrett Rollins — devoted to the interface of big data and cancer precision medicine.

SAM OGDEN (DFCI)

"BY REDUCING UNCERTAINTY AND TRULY PERSONALIZING PATIENT CARE, MACHINE LEARNING CAN TOTALLY TRANSFORM THIS AREA."
REGINA BARZILAY

In addition to machine learning experts like Barzilay, the meeting organizers invited pioneering thinkers in the fields of cancer medicine, tumour genetics and data science with the aim of forging connections between AI and oncology. "Each of these two fields has made major advances in the recent past, such as the ability to generate genomic and molecular profiles of tumours, resulting in massive data, and deep machine learning for sophisticated prediction methods," says Chris Sander, a computational and systems biologist who directs the DFCI cBio Center and helped to organize the meeting. "We are building more bridges between clinicians, cancer researchers and machine learning experts to create major collaborative opportunities."

Meeting of minds

Cancer research didn't always command such a multidisciplinary approach. Once the provenance of mostly cell biologists and mouse geneticists, successive developments in fundamental science and biomedical research have made the field data-rich and data-driven — and that's brought with it new challenges for researchers hoping to make sense of the complexity.

Many experts, recognizing the scale of the challenge, have openly embraced data sharing and collaboration. Case in point: Project GENIE. Short for Genomics Evidence Neoplasia Information Exchange, GENIE launched[2] in 2015 as a vehicle for sharing tumour genetic profiles from patients in active clinical treatment for use in cancer research by a broad community.

Project architects built this massive database to identify novel therapeutic targets, design biomarker-driven clinical trials and find genomic determinants of response to therapy. But now with genomic data from around 50,000 patients and counting — and 19 participating institutions from around the world — GENIE leaders such as Charles Sawyers, a cancer biologist from the Memorial Sloan Kettering Cancer Center (MSK) in New York City, are struggling to draw actionable insights from the glut of data.

At the Big Data and Cancer Precision Medicine meeting, Dana-Farber's Deborah Schrag (top right) hosts a breakout discussion on strategies for clinical data curation; Matthew Meyerson of Dana-Farber gives a talk on cancer genomes and therapeutic responses (middle left); and attendees participate in the Q&A and poster sessions.

TOP AND MIDDLE LEFT IMAGES ©SAM OGDEN (DFCI); MIDDLE RIGHT AND BOTTOM ROW ©ERICA FEICK

Sawyers is thus appealing to computer scientists to join GENIE and help make important translational discoveries: "I really want you to help us learn how to extract important insights more efficiently," he says, "particularly in extracting clinical data from electronic health records."

Computer scientist David Blei, from Columbia University in New York City, welcomes that invitation. "I don't know anything about biology or medicine, but I know about machine learning," he says — and he's keen to apply his skills in the life sciences, as evidenced by a project recently started with postdoc Wesley Tansey to model dose responses in high-throughput drug screens.

Blei offers a few cautionary notes for cancer biologists hoping to jump into AI. When it comes to making causal inferences from large observational datasets, he says, a basic AI model has certain flaws. Give an algorithm a task such as predicting which actors boost movie revenues, and it is likely to overestimate[3] the importance of action stars since action movie franchises tend to earn more money than art-house dramas, regardless of the name on the billboard. Transpose this to more consequential applications in medicine, and it is clear that there is need for careful consideration of assumptions and probabilistic modelling when developing AI methods.

David Sontag, head of the MIT Clinical Machine Learning Group, echoes this point. "There's a lot of subtlety in correctly applying machine learning in biomedical research," he says.

Sontag's own research has lately focused on analysing gene expression patterns in the bone marrow of patients with multiple myeloma. His team has identified putative disease subtypes defined by gene activity patterns and responses to treatment. "We want to build probabilistic models of disease progression that can help provide accurate prognoses and personalized treatment plans for patients," says Rebecca Peyser, a data scientist in Sontag's lab.

But while Peyser and Sontag have taken great care to ensure the integrity of their models, they emphasize that it is easy to make mistakes. Sontag warns that as cancer research makes increasing use of real-world evidence from electronic medical records and other sources, additional care needs to be taken to guarantee that any retrospective evaluation mimics how the algorithms would actually be used prospectively.

Still, even the best algorithms can draw spurious connections if given spurious kinds of training. "You have to care about the data-generating process — including the sources of noise and bias in recording and generating the data," says Suchi Saria, an expert in machine learning and healthcare at Johns Hopkins University in Baltimore, Maryland. "If you don't understand the data-generating process and reason about how it impacts your model, you're likely in trouble."

Opportunities abound

Methodological pitfalls notwithstanding, AI algorithms are already beginning to affect cancer research and clinical care, such as early diagnosis and prevention, drug discovery, matching patients to clinical trials and treatment decisions.

Thomas Fuchs, a computational pathologist at MSK, predicts that one of the first clinical settings in which AI will see widespread adoption is in pathology. Scanned images of fixed and stained tissue samples represent “an enormously dense data modality”, he says. And as good as humans are at recognizing patterns in the cellular masses, he argues that expert clinicians and computers, together, can do even better.

To prove his point, Fuchs and his team trained and tested[4] a deep learning model on more than 12,000 slides of prostate biopsies. “The model actually learns what prostate cancer looks like, without manual annotations,” Fuchs says. These results are presented in a slide viewer, offering a tool that's akin to “Google Maps for histological slides”.

Already, many drug developers are routinely incorporating these kinds of computational pathology models into their clinical research programmes, relying on machine learning to, for example, quantify levels of biomarkers that may help explain why only some patients respond to immunotherapy.

Advanced computational techniques are also helping to stratify patients undergoing immunotherapy. Two years ago, a team led by Gunnar Rätsch, a data scientist at the Swiss Federal Institute of Technology in Zurich, created[5] a tool that measures alternative splicing events in gene transcripts within tumours. This tool, he explains, can flag putative splice-associated neoantigens that may predict success rates to checkpoint inhibitors or, as Rätsch and his colleagues showed through a systematic analysis of tumours from 8,705 patients reported[6] in August 2018, provide the basis for designing personalized cancer vaccines.

In a similar vein, computational immuno-oncologist John-William Sidhom, working with Drew Pardoll and Alexander Baras at Johns Hopkins, is applying deep learning to profile the so-called immune synapse — the interface between the immune cells that process signals from the tumour and those that respond by taking action against the rogue tissue — to better predict drug responses and guide therapeutic decision-making.

Sidhom's analytical framework, known as a convolutional neural network, is more often used among cancer researchers to study diagnostic image data from radiology or pathology. But the same technologies can be applied to the immune synapse — which, Sidhom notes, “is a very important part of having a very potent and specific anti-cancer response”.

AI also has application in more traditional areas of cancer therapy, such as in predicting who is likely to suffer severe side effects from radiation treatment. In what she's calling “Big-RT”, computational biologist Bissan Al-Lazikani, from the Institute of Cancer Research in London, and her colleagues recently fed data on outcomes, clinical metrics and genetic profiles from nearly 1,000 prostate cancer patients into a machine learning model. In as-yet unpublished work, they found dozens of different parameters that all seemed to affect an individual's likelihood of severe toxicity.

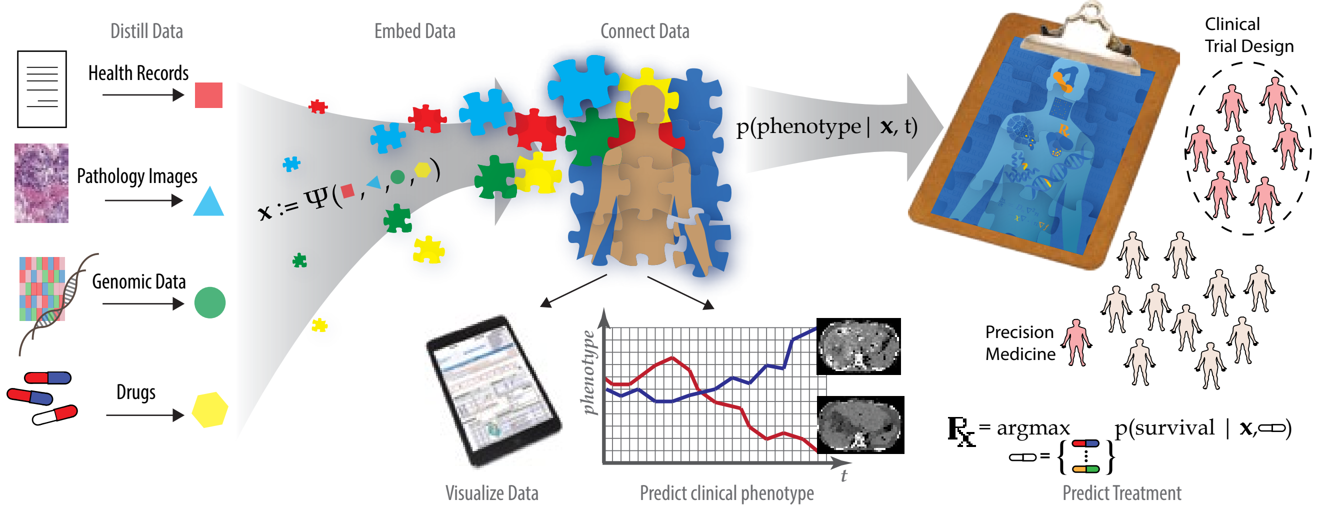
DATA TO KNOWLEDGE TO ACTION

Precision medicine integrates many strands of data using machine learning algorithms to help doctors predict what will happen to the patient and decide on the best treatment.

DATA
Make clinical data computable

KNOWLEDGE
Predict clinical phenotypes

ACTION
Select optimal therapy



ANDRE KAHLES, GUNNAR RÄTSCH, CHRIS SANDER

ERIKA FEICK

“I REALLY WANT YOU TO HELP US LEARN HOW TO EXTRACT IMPORTANT INSIGHTS MORE EFFICIENTLY.”
CHARLES SAWYERS

A collect call

Determining which potential risk factors matter most will require collecting much more data. But unfortunately, says Al-Lazikani, even though around 60% of all people diagnosed with cancer undergo radiation therapy at some point, comprehensive data of the kind she needs are rarely pulled together in a systematic way. “We have to be collecting all the information about all our patients all of the time,” she says.

AI has great potential as an aid in drug discovery, too. In a recent paper[7], members of the International Cancer Genome Consortium used a resource that Al-Lazikani helped create called canSAR (and the publicly available machine learning software it comes with) to identify more than 60 new potential drug targets for treating prostate cancer.

Olga Troyanskaya, a computer scientist at Princeton University in New Jersey, is also helping to identify new druggable targets — and illuminate new aspects of the basic biology of cancer — through analyses of non-coding gene mutations. Her deep learning techniques have revealed causal roles for regulatory elements in triggering autism, but can be used to implicate sequences in tumour formation as well.

One issue cited repeatedly by cancer researchers hoping to bring machine learning into their field is the dearth of suitable big datasets. Sawyers and his colleagues spent years and millions of dollars amassing, harmonizing and curating the Project GENIE database. Nikhil Wagle, an oncologist at DFCI and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, has found an alternative approach through direct engagement of patients.

Using a patient-driven research initiative called Count Me In, he and his colleagues have rapidly amassed medical records and tissue samples donated by thousands of patients with various tumour types. “It's a very quick way to get huge amounts of data,” says Wagle. “And unlike many other studies, we aim to combine genomic, molecular, medical record, and patient reported data in the same database, making it quite comprehensive.”

Machine learning curve

As AI penetrates all aspects of cancer research, clinicians have begun to think about what the future holds for the practice of medicine (see ‘Data to knowledge to action’). And while they acknowledge the potential for advances in precision oncology, some are also ringing alarm bells about potential risks.

For example, humans might be overburdened by their new computational aids, says Mia Levy, a cancer informatics researcher who directs the Rush University Medical Center in Chicago, Illinois. “When you think about these algorithms,” she says, “you've got to think about how you're going to implement these into the workflow of the clinician.”

These considerations are important. But what matters ultimately to patients is not the process but the clinical outcome — and for that, says Barzilay, machine learning algorithms are a vast improvement on the status quo.

Barzilay often gives talks about her work to public audiences. After one such talk, an attendee emailed about an abnormal growth she had developed in her milk-producing glands. The woman said she was also taking hormone replacement therapy (HRT) to deal with the symptoms of menopause, and wondered if these drugs might increase the odds of her breast mass turning malignant.

She asked Barzilay what to do — which sent the MIT professor on a search looking into reams of clinical data. Barzilay crunched the numbers with her AI algorithms, and found that HRT, on average, does not seem to significantly affect outcomes among women with this type of pre-cancerous breast disease. Although she would never want to offer medical advice, Barzilay says she does think that “the patient needs to know this kind of information, and we need to provide it”.

Fortunately, new transformative technologies, both laboratory-based and computational, are making it possible to collect high-resolution clinical and biological data, extract meaningful insights, and then offer patients just this sort of personalized medical advice. And that means cancer patients can hope for earlier diagnosis and therefore better outcomes than Barzilay herself received just four years ago — an eternity in today's rapidly advancing field of precision oncology — thanks to advances in genomic and molecular profiling and in computational modelling and AI. ■

REFERENCES

1. Lehman, C. D., *et al. Radiology* **00**:1–7 (2018) https://doi.org/10.1148/radiol.2018180694.

2. The AACR Project GENIE Consortium. *Cancer Discov*; **7**(8); 818–831 (2017).

3. Wang, Y. & Blei, D. M. Preprint at https://arxiv.org/pdf/1805.06826.pdf (2018).

4. Campanella, G., Krauss Silva, V. W. & Fuchs, T. J. Preprint at https://arxiv.org/abs/1805.06983 (2018).

5. Kahles, A., Ong, C. S., Zhong, Y. & Rätsch, G. *Bioinformatics* **32**(12); 1840–1847 (2016).

6. Kahles, A., *et al. Cancer cell.* **34**(2); 211-224 E6 (2018).

7. Wedge, D. C., *et al. Nature Genetics* **50**; 682–692 (2018).