BioCentury

REPRINT FROM JANUARY 23, 2017

TOOLS & TECHNIQUES

HOPES IN THE MACHINE

BY ERIN MCCALLISTER, SENIOR EDITOR

Applying machine learning tools to drug R&D could usher in a new era of effective therapies targeting causative disease drivers, higher success rates in clinical trials and real-world tools to manage adherence, prevent adverse events and reduce total healthcare costs.

Biopharma companies have been working at mining big data since the genomics era began, but traditional data mining parses only one type of data at a time to yield correlations. For example, mining sequencing data can reveal a correlation between a patient population and a disease target or mutation. But it doesn't reveal whether the mutation is a causative agent in pathogenesis, or its relationship to other factors that affect the safety and efficacy of targeted drugs in different patients.

In contrast, machine learning tools analyze multiple sets of data from disparate sources — often in wildly different formats — to sniff out connections among biology, genetics and phenotypic traits.

Using an iterative, learning approach, the technologies have the potential to identify previously unknown pathways, uncover which targets are causing rather than simply correlated with a disease, predict how patients will respond to therapies, and draw conclusions about whether a clinical trial will be successful or how a drug will perform in the real world.

These technologies are created from a series of neural networks — a computational technique that can be used to discover and predict patterns among a set of inputs and corresponding outcomes.

They employ natural language processing to analyze unstructured data from the literature and physician notes in electronic medical records (EMRs), and combine this information with millions of structured data points to develop hypotheses that can then be tested in preclinical animal models, clinical trials or the real world.

Adoption of these technologies remains modest but could increase over the next few years as early programs deliver successful clinical trials, approved drugs and better outcomes in the real world.

"WE CAN THEN IDENTIFY WHERE TO START INTERVENING AND HOW."

ALEX TURKELTAUB, ROAM ANALYTICS

A compound developed for a cancer population selected using GNS Healthcare Inc.'s reverse engineering forward simulation (REFs) technology could be on the market within a few years. The company also has shown its Efficacy to Effectiveness (E2E) tool can accurately predict real-world performance of a new drug prior to its launch.

Two companies using their platforms to discover targets and drug candidates, BenevolentAI and Berg LLC, plan to have compounds in the clinic in the next year.

Many, many more experiments are ongoing via partnerships as drug developers and providers seek to understand whether and how these tools can save time and money, and improve research and clinical outcomes.

"I think value-based arrangements with machine learning tracking outcomes will be the sharp ends that I think will pay off very quickly within the next year, for sure," said John Glasspool, head of corporate strategy and customer operations at Shire plc, which is evaluating a handful of undisclosed machine learning technologies.

As these tools start to deliver improvements in patient adherence and identification of specific patients who will benefit from a drug, "we will get a better understanding of the disease biology and targets" he added. That in turn will result in greater Phase II and Phase III successes, he said.

Realizing that potential, however, will require much more data sharing than is currently taking place, along with creation of new data standards (see "Driving Adoption").

Profiles of nine companies who spoke to BioCentury and are developing machine learning technologies for use in drug R&D, from discovery to real-world evidence development, follow.

BENEVOLENTAI: MACHINE PLUS MAN

The aim of using machine learning technologies in drug discovery is to dramatically reduce the time and expense of traditional wet lab work.

"Our goal is to be at least fourfold more efficient based on the combination of more successful candidates, and getting into and through the clinic quicker and for less money," said Jackie Hunter, CEO of BenevolentAI's BenevolentBio unit.

The company's Judgment Augmented Cognition platform uses private and public genomic databases, as well as the scientific literature and patent databases, to tease out new drug targets and candidates in rare diseases.

"We run our tool to help us generate interesting connections and associations that were previously unknown, but are there in the data," said Hunter.

The system's output is a ranking of targets along with recommendations for molecular structures of drug candidates.

An internal team of five life science researchers then reviews the output to ensure it makes "biological sense," Hunter said. Human review of the machine learning output helps avoid potential confounding caused by the published literature's bias toward successful studies.

The most advanced program identified using the tool is a repurposed program from Johnson & Johnson's Janssen Pharmaceutica N.V. unit that will enter Phase II trials this year.

"We're taking [the pharma's] chemical data, biologic data, genomic data and ingesting all of that for a more holistic picture, as well as the larger set of data we have," Hunter said.

Janssen granted BenevolentAI exclusive, worldwide rights in November to develop, manufacture and commercialize undisclosed compounds for undisclosed disease areas.

BenevolentAI also plans to begin preclinical studies this year of candidates against undisclosed targets that the technology discovered to be associated with an undisclosed disease.

Proof of concept came from using the tool to identify new targets that were previously unknown to be causative in amyotrophic lateral sclerosis (ALS). Studies in a cell model derived from stem cells of ALS patients showed that four out of five compounds against the top hits, which are in development for other diseases, had a beneficial effect.

"One did not work, two worked and two others worked extremely well," Hunter said.

BERG: INTERROGATING SAMPLES

Berg is developing an interrogative biology platform that allows it to query data from the company's *in vivo* systems and patient-level data to uncover cause and effect relationships in disease biology.

The system has generated a portfolio of undisclosed novel preclinical compounds for diabetes, cancer, and Parkinson's and Alzheimer's diseases, and plans to start Phase I testing of a diabetes compound this year or next.

Berg's *in vivo* systems consist of cell lines generated from patient tissue and serum samples. The company manipulates them to induce signaling that results in a phenotypic response or outcome. Berg uses mass spectrometry and other technologies to capture in real-time all the proteomic, metabolomic, DNA and RNA expression, phenotypic and other data generated from these perturbations.

"We don't analyze static molecular data, we create intelligent experimental designs of disease perturbations to induce signaling in disease-specific molecular pathways," said Chief Analytics Officer Slava Akmaev.

It then applies its Bayesian network machine learning technology to analyze the data to find correlations and assess whether or not there is a causal relationship.

“OUR GOAL IS TO BE AT LEAST FOURFOLD MORE EFFICIENT.”

JACKIE HUNTER, BENEVOLENTAI

“The Bayesian networks allow us to actually see the mechanics and reasoning behind certain inferences so we can assess the cause and effect relationships and answer the question of what drives the [phenotypic or signaling] outcome,” Akmaev said.

“We can identify pathways and targets, not through decades of research, but in a matter of months,” he added.

twoXAR: WORKING THE NETWORK

twoXAR Inc. is pursuing a service model using its DUMA technology to reduce the amount of time and money its customers spend synthesizing and screening thousands of potential drug candidates.

DUMA uses both public and private data, including physical structures of proteins and molecules, data on protein binding and protein interaction, clinical data and gene expression data.

It identifies disease signatures from the biological data sets and then builds networks of protein-protein and protein-compound interactions to generate a computational disease model.

DUMA uses the model to score the anticipated efficacy of compounds against the disease. Finally, it produces a prediction report that ranks the compounds according to how potent they are likely to be against the disease, and describes the factors involved in making those rankings.

DUMA also includes a patent search to determine whether the candidates are novel.

In rheumatoid arthritis, DUMA identified 10 novel candidates. Three produced an efficacy signal in a rat model. The targets are not disclosed, but CEO Andrew Radin said, “they were hitting targets that aren’t traditionally associated with RA.”

Radin thinks the tool could shave years off drug discovery, but he thinks more data will be required to replace traditional practices.

“I can see this as supplementing that. Maybe the computer comes up with 10 options and the researchers also come up with six of their own, and then they test them,” he said. If candidates selected by DUMA show efficacy in the clinic, he thinks it may start to edge out the traditional approach.

DUMA does not assess the compounds’ toxicity, but Radin said the company is in the early stages of developing a toxicity predictor tool.

twoXAR has at least nine collaborations with undisclosed companies and academic groups to identify new drug candidates and expects to publish data this year.

INSILICO: DISEASE SELECTION

InSilico Medicine Inc. is using its machine learning tool to repurpose drugs and identify the optimal therapeutic area for new drug candidates.

The company built its system based on gene expression data from the Broad Institute of MIT and Harvard and NIH’s Library of Integrated Cellular Signals (LINCS), which contains perturbation profiles across multiple cell lines and drug modalities. Perturbation profiles include changes in gene expression and signaling in cell lines after they’ve been exposed to drugs and/or drug candidates. The database includes more than 1 billion data points.

InSilico used the data to create neural networks that link gene expression profiles for an agent to a specific disease area such as cancer or cardiovascular disease.

InSilico is now adding on a module to analyze phenotypic data from animal models as well as the gene expression data to predict the disease area in which a therapeutic candidate will have the greatest efficacy.

The company is working with an undisclosed biopharma company to screen and prioritize molecules in its library.

“If they have all of these molecules in their library and come to us with their phenotypic and gene expression data, we can tell them to focus on these three for a particular type of cancer, to reduce the experimental time and cost,” said COO Qingsong Zhu.

WATSON: RUNNING THE R&D GAMUT

IBM Corp. is using its Watson Health platform to identify novel targets and optimal drug combinations and to improve patient outcomes in the real world.

Its Watson for Drug Discovery tool was built using data from full-text medical journal articles, patents and genomics information to uncover new disease pathways and relationships between new genes and proteins that could be targeted.

“We believe that we’ll be able to identify new pathways and relationships with new genes by looking at the relationship between ontology, amino acids and other data sources," Lauren O'Donnell, VP of life sciences, told BioCentury.

In December, IBM and Barrow Neurological Institute announced that Watson had identified new genes associated with ALS using its own databases and data provided by the institute. Watson analyzed 1,500 proteins and genes. Barrow then tested the top 10 in preclinical models and confirmed that eight were linked to ALS, including five never before associated with ALS.

IBM is now partnered with Pfizer Inc. to use the drug discovery tool to determine which patients would benefit most from cancer immunotherapies, as well as identify optimal combinations of immunotherapies that could produce the greatest efficacy.

IBM will gain access to Pfizer's data, which Watson can then use to augment its neural networks.

"It's tapping into the machine learning, the natural language and cognitive reasoning to support the discovery of new targets, combinations and patient selection strategies for immunotherapy," O'Donnell said.

IBM has 10 other pilot studies ongoing with biopharma companies and other clients.

Separately, Celgene Corp. and IBM partnered in November to co-develop Watson for Patient Safety, an outcome- and evidence-based tool to identify factors that influence adverse events and determine whether they are preventable, or linked to a drug.

Watson for Patient Safety will use machine learning to collect, collate and automate the analysis of high volumes of data from diverse sources, including data from EMRs, medical claims databases and other healthcare information sources to identify what factors are influencing the adverse events and whether or not they are preventable or linked to the drug.

IBM is also building a clinical trials module and a real-world evidence module that could be used to design and manage value-based payment models. IBM gained access to clinical data on over 50 million patient lives in 2015 via its acquisition of Explorys Inc. The 2016 acquisition of Truven Health Analytics Inc. in 2016 added hospital, payer and other postmarket data for 215 million patient lives.

DRIVING ADOPTION

Data sharing among academia, industry and other healthcare stakeholders — along with standards for collecting and managing the data — will be necessary to improve accuracy and increase adoption of machine learning technologies in drug development.

Machine learning tools rely on a mix of public and private data from disparate sources including the literature, clinical trials, medical records and insurance claims. Because these tools build upon each new piece of data in an iterative, learning fashion, the lack of data-sharing could stifle progress, and limit the tools' usefulness.

"What I would hate to see is that the data gets siloed and then we have multiple versions of what the data look like and it won't be as beneficial as it could be," said John Glasspool, head of corporate strategy and customer operations at Shire plc.

A few ongoing initiatives are laying the ground work for the kind of sharing needed to maximize knowledge. For example, Project Data Sphere LLC was formed in 2014 to share data from the control arms of cancer trials. The group now has data representing more than 29,000 lives.

Separately, Johnson & Johnson provides third-party researchers with access to full clinical study reports and patient-level data from trials conducted for the pharma's drugs approved in the U.S. and Europe, trials that were terminated and trials that have been accepted for publication. The initiative is being managed by the Yale School of Medicine Open Data Access (YODA) Project.

With so much data coming from so many different sources, standards to ensure consistency and comparability across studies is also necessary.

Machine learning companies currently use both manual and automated systems to get the disparate sets of data to communicate with one another.

First companies must annotate the data and build their own dictionaries that teach the algorithms the difference between a drug target or disease pathway and an unrelated object.

"We spend the time up front to really do the detailed annotation, curation and building the appropriate dictionaries on both the chemical and literature sides so when the algorithm comes across CBT2, it knows it's the cannabinoid receptor 2 and not a post office in Cambridge," said Jackie Hunter, CEO of BenevolentAI's BenevolentBio unit.

Similarly, Medidata Solutions Inc. uses machine learning to bring consistency to treatment of data in different datasets, such as when different physicians can use different descriptions to report the same adverse event. "We're turning it into a consistent set of coded results," said President Glen de Vries.

Developing and adopting standards for collecting and managing these and other data would eliminate some of that work, and improve the results of machine learning algorithms.

Here, too, initiatives have begun. The Clinical Data Interchange Standards Consortium (CDISC) is developing platform-independent data standards that allow for interoperability between different data types and data generated from different platforms. FDA adopted CDISC standards for all NDAs, ANDAs and BLAs. Japan's PMDA has adopted similar standards.

— BY ERIN MCCALLISTER

IBM plans to eventually combine the modules to build an integrated machine learning platform that could be used across R&D.

WUXI NEXTCODE: PICKING PATIENTS

New WuXi Life Science Ltd.'s WuXi NextCODE unit is augmenting its contract genomics service with deep learning capabilities to identify patient selection signatures for clinical trials and treatment decision making.

"We're focused on bringing together different data sources, data types like DNA and RNA array data and a variety of different phenotypic information to create sophisticated mathematical models," said COO Hannes Smarason.

This year the company expects to publish results of a study in which the technology identified a signature of patients who respond to a specific cancer therapy based not only on genomic biomarkers but also other biologic and demographic data.

NextCODE is working with researchers at Massachusetts Institute of Technology, Harvard University and Stanford University.

NextCODE is also enriching its machine learning technology through contract genomics collaborations such as its Jan. 9 partnership with AbbVie Inc. and Genomics Medicine Ireland Ltd. to collect and analyze genomic and phenotypic data for 45,000 individuals in Ireland.

AbbVie wants to identify new drug targets and biomarkers, while NextCODE hopes to enhance the power of its machine learning technologies. "By virtue of processing this information on a national scale and through these partnerships, we are able to train the network and get more insights into the data. We have no ownership of this data or the targets identified, but there will be an awful lot of learning for our algorithms," Smarason said.

GNS: ENABLING PRECISION MEDICINE

GNS is applying its machine learning platform to both R and D with the aim of enabling personalized medicine.

"The big focus at GNS from the start has been to match drugs to individual patients, and maybe simultaneously lower total costs of care," said CEO Colin Hill.

An undisclosed cancer drug candidate from Eisai Co. Ltd. is making its way through a clinical trial that enrolled patients based on a treatment response signature identified using GNS's REFS platform.

"This is making its way through drug development, along with other similar types of markers applied to drugs for major depressive disorder that are making their way to the market. So we certainty expect to see, in the next few years, the signature from GNS technology actually having an impact on real-world patients," said Hill.

REFS uses Bayesian network inference to build disease models using algorithms that can identify new linkages or draw conclusions from genomic, postmarket and clinical data.

It doesn't require the user to start with a hypothesis. Instead, the algorithms use the raw data to generate potentially trillions of hypotheses about which biological network or pathway is responsible for the treatment effect to create a learning disease model.

GNS has used REFS to create models in Huntington's disease, multiple myeloma (MM), PD and AD and multiple other diseases, Hill said.

"With the creation of these models straight from the data, we have the ability to simulate interventions into these models to see if this pathway causes a change in phenotypes," he told BioCentury.

In 2012, GNS partnered with Covance Inc. to use REFS to predict the likelihood of a therapy's success, with an initial focus on Type II diabetes.

GNS has not yet disclosed results from the Covance partnership, but "we demonstrated that it was possible to do," Hill said. Covance was acquired by Laboratory Corp. of America Holdings in 2015.

GNS also has a collaboration with the Multiple Myeloma Research Foundation (MMRF) to use REFS to analyze data from the CoMMpass sequencing study. CoMMpass is following over 1,000 patients for eight years to determine how molecular profiles change with response to treatment.

GNS has analyzed the sequence and treatment outcome data generated thus far to identify a series of new targets in MM as well as to identify a signature of high-risk patients unlikely to benefit from current standard of care. Data were presented in December at the American Society of Hematology meeting.

"WE CAN IDENTIFY PATHWAYS AND TARGETS, NOT THROUGH DECADES OF RESEARCH, BUT IN A MATTER OF MONTHS."

SLAVA AKMAEV, BERG

"THE BIG FOCUS AT GNS FROM THE START HAS BEEN TO MATCH DRUGS TO INDIVIDUAL PATIENTS, AND MAYBE SIMULTANEOUSLY LOWER TOTAL COSTS OF CARE."

COLIN HILL, GNS HEALTHCARE

"About 80% of patients will respond to the first-line therapy, but 10-15% do horribly and continue to do horribly. We've identified a combination of factors that define the molecular and clinical subtype of that patient so now you can target those patients in your Phase II or Phase III study because that is where the real unmet need is," said Chief Commercial Officer Iya Khalil.

GNS also has a second technology called E2E that could be used to negotiate value-based deals.

E2E creates a clinical model for a new drug using clinical data from the new entrant, along with clinical data for drugs that are marketed to treat the same indication, and a real-world model using claims data for the marketed drugs. The models require data on the same endpoints for all the drugs included, as well as other lab values, co-morbidities and demographics.

E2E then compares the clinical and real-world models of the marketed drugs, and applies the information to the clinical data for the new drug to predict how it will perform in the real world.

In October, GNS and partner Novartis AG presented data that showed E2E accurately predicted the real-world effectiveness of relapsing-remitting MS drug Gilenya fingolimod based on Phase III data.

Such information could be used by companies to set real-world performance goals for drugs as part of outcomes-based contracts with payers.

Khalil said GNS is working with Novartis and other pharmas in other disease areas.

ROAM: SAFETY IN THE REAL WORLD

Roam Analytics is deploying its Health Knowledge Graph in the postmarket setting to predict potential AEs based on baseline characteristics of patients and/or providers, and suggest an opportunity for intervention.

The technology incorporates genomics and clinical data, as well as data on physician practices and where the patient is being treated. The Knowledge Graph uses propensity scoring, a statistical technique that estimates the effect of a treatment or other intervention and is able to make causal inferences.

"It's not just about predicting that the patient will have a bad outcome, but what is it about the patient or the patient pathway that he or she followed that leads to the bad outcome," said CEO Alex Turkeltaub.

Roam has been testing the technology for about six months with undisclosed biopharma companies and healthcare providers.

"The thing we want to allow our customers to do is answer questions that previously were impossible to answer. So we can set up these *de facto* real-world evidence trials where they can take patient groups, split them into two groups that are identical in every aspect except one to get answers to their questions," Turkeltaub told BioCentury.

"We aren't just tracking the data, but we can then identify where to start intervening and how. So if we see this type of behavior or symptoms in a patient that looks like 'X,' we can provide them with a supplemental therapy to avoid any serious side effects," Turkeltaub said.

The tool can be used to set the parameters for and manage performance-based outcomes deals between companies and payers, where Roam can serve as a third-party to analyze the data.

Roam expects to have data in the next year showing how well the tool can help to control costs and/or improve patient outcomes, including adherence.

MEDIDATA: REAL-WORLD RESPONSES

Medidata Solutions Inc. is also using machine learning to build disease models that can predict how a particular candidate drug will perform in the real world and to continually refine which patients are most likely to respond.

The technology uses clinical data from client studies that are stored on Medidata's Clinical Cloud platform, along with real-world data. Once a drug is launched, the technology can incorporate data from patient registries, claims, wearable devices and other sources to refine patient selection criteria.

If, for example, real-world use reveals differential responses among patient subgroups that were not observed in the clinic, Medidata's system can pinpoint which patients are or aren't responding and why. Drug companies could use the data to work with providers to ensure the right patients get the drug.

For patients who aren't achieving a benefit, machine learning could also determine whether better side effect management or other interventions could improve response.

"We are giving our clients the agility where they can take this information, react quickly in the context of a particular patient or a particular drug to get better outcomes," said President Glen de Vries.

On Jan. 12, Celgene said it would use Medidata's Clinical Cloud, including its monitoring tools and machine learning technologies, to "optimize clinical research progress, improve patient outcomes and drive value-based, personalized healthcare."

Details of the partnership aren't disclosed.

COMPANIES AND INSTITUTIONS MENTIONED

AbbVie Inc. (NYSE:ABBV), Chicago, Ill.
American Society of Hematology (ASH), Washington, D.C.
BenevolentAI, London, U.K.
Berg LLC, Farmingham, Mass.
Broad Institute of MIT and Harvard, Cambridge, Mass.
Celgene Corp. (NASDAQ:CELG), Summit, N.J.
Clinical Data Interchange Standards Consortium, Austin, Texas
Eisai Co. Ltd. (Tokyo:4523), Tokyo, Japan
Genomics Medicine Ireland Ltd., Dublin, Ireland
GNS Healthcare Inc., Cambridge, Mass.
Harvard University, Cambridge, Mass.
IBM Corp. (NYSE:IBM), Armonk, N.Y.
InSilico Medicine Inc., Baltimore, Md.
Johnson & Johnson (NYSE:JNJ), New Brunswick, N.J.
Laboratory Corp. of America Holdings (NYSE:LH), Burlington, N.C.
Massachusetts Institute of Technology (MIT), Cambridge, Mass.
Medidata Solutions Inc. (NASDAQ:MDSO), New York, N.Y.
Multiple Myeloma Research Foundation (MMRF), Norwalk, Conn.
National Institutes of Health (NIH), Bethesda, Md.
Novartis AG (NYSE:NVS; SIX:NOVN), Basel, Switzerland
New WuXi Life Science Ltd., Shanghai, China
Pfizer Inc. (NYSE:PFE), New York, N.Y.
Pharmaceuticals and Medical Devices Agency (PMDA), Tokyo, Japan
Project Data Sphere LLC, Cary, N.C.
Roam Analytics, San Mateo, Calif.
Shire plc (LSE:SHP; NASDAQ:SHPG), Dublin, Ireland
Stanford University, Stanford, Calif.
twoXAR Inc., Palo Alto, Calif.
U.S. Food and Drug Administration (FDA), Silver Spring, Md.
Yale School of Medicine, New Haven, Conn.

REFERENCES

McCallister, E. "Predicting reality." *BioCentury* (2016)
McCallister, E. "Big data: Then and now." *BioCentury* (2012)
Zipkin, M. "A need for speed." *BioCentury Innovations* (2016)

BIOCENTURY INC.

NEWSROOM
pressreleases@biocentury.com

SAN CARLOS, CA
+1 650-595-5333; Fax: +1 650-595-5589

CHICAGO
+1 312-755-0798; Fax: +1 650-595-5589

WASHINGTON, DC
+1 202-462-9582; Fax: +1 202-667-2922

UNITED KINGDOM
+44 (0)1865-512184; Fax: +1 650-595-5589

Use of Images: Certain Images used in BioCentury Inc.'s Publications, Video Content, Websites, Services, Notices and/or Marketing Materials are licensed from Getty Images (US), Inc. Any such image of a person or object so displayed is being used for illustrative purposes only and any such person or object depicted, if any, is merely a model. For more information see "Use of Images" found under the "About Us" tab on the Homepage at www.biocentury.com.

All information provided through BioCentury Inc.'s Publications, Video and Audio Content, and Websites is gathered from sources that BioCentury believes are reliable; however, BioCentury does not guarantee the accuracy, completeness, or timeliness of such information, makes no warranties regarding such information, and is not responsible for any investment, business, tax or legal decision made or action taken in reliance upon such information.