THE WALL STREET JOURNAL.

TUESDAY, APRIL 6, 2021

JOURNAL REPORTS: TECHNOLOGY

The People in This Medical Research Are Fake. The Innovations Are Real.

'Synthetic-data technology,' by creating artificial patient populations, has the potential to speed up innovations without compromising privacy

By Dov Lieber

Researchers in Israel were happy to get their hands on data about thousands of Covid-19 patients, including a 63-year-old father of two who was admitted to the emergency room with Covid-19 and soon recovered. It was the early days of the coronavirus pandemic and the treatments used for this patient could provide invaluable insight into the then little-understood virus.

Normally, it would have been unthinkable to share sensitive medical details, such as the patient's use of Lipitor for high cholesterol, so quickly, without taking measures to safeguard his privacy. But this man wasn't real. He was a fake patient created by algorithms that take details from real-life data sets such as electronic medical records, scramble them and piece them back together to create artificial patient populations that largely mirror the real thing but don't include any real patients.

Medical researchers and data scientists say such "synthetic" healthcare data has the potential to speed up medical innovation. The rapid digitization of health records has created troves of patient information that can be analyzed by algorithms and harnessed to improve disease-treatment models and develop new products and services. But patient information isn't easy to get because privacy laws require medical data to be stripped of names, addresses and other identifying details before it can be shared, a time-consuming process that can take months. Even those measures don't satisfy some privacy advocates, who point to studies showing that it is possible to re-identify patients even after data sets have been anonymized.

Enter synthetic-data technology.

"The key advantage that synthetic data offers for healthcare is a large reduction in privacy risks that have bugged numerous projects [and] to open up healthcare data for the research and development of new technologies," says Allan Tucker, a professor at Brunel University London and author of a study published in Nature in November showing the validity of using synthetic data as a substitute for real healthcare data.

Testing ground

Synthetic-data technology has been around for several years, and there are many synthetic-data companies working in industries such as finance and insurance. A new wave of techniques designed to better preserve statistical patterns while protecting privacy, however, cleared the way for its use in healthcare more recently, industry insiders say. While synthetic data isn't widely used in healthcare yet, the Covid-19 pandemic accelerated demand for it as providers and researchers rushed to understand the virus and find treatments.

Israel, which has an advanced digital national healthcare system and electronic medical records going back 20 years, has become a strong testing ground for the technology. All four of the country's HMOs and healthcare providers such as Sheba Medical Center, Israel's largest hospital, are clients of MDClone, an Israeli startup with a platform for creating synthetic data from medical records.

Early in the coronavirus pandemic, Sheba used MDClone's platform to synthesize data from its Covid-19 patients. The hospital then invited data scientists from around the country to glean whatever knowledge they could about Covid from the data set. More than 40 groups that specialize in data science signed up to help, the hospital says, including startups, academics and big multinational companies.

The initiative ultimately led the hospital to implement an algorithm, developed by Israeli company Data Science Group, that helps clinicians decide when patients should be given a drug treatment such as remdesivir or be sent to the ICU.

Sheba also was able to integrate its electronic health records with those of Israel's second-largest HMO, Maccabi Health Services, to give clinicians a far broader view into the medical histories of Sheba's Covid-19 patients, helping them find links between pre-existing conditions and the virus's toll on its victims.

"People were completely astounded by how quickly we did this," says Eyal Zimlichman, the hospital's chief medical officer and chief innovation officer. If Sheba or Maccabi had been using real patient data, it would have taken many months for committees to approve its use and make it privacy-friendly by stripping out personal details, he says.

Scientists aren't yet relying solely on synthetic data for their research, Dr. Zimlichman says. They are testing their hypotheses on synthetic data first, then retesting them on real patient data before submitting studies to medical journals for publication, he says. The use of synthetic data is sometimes described in the industry as a sandbox, allowing researchers to easily test as many hypotheses as they would like, without having to go through the long process of requesting real data.

Since the start of the pandemic, researchers at Sheba have published six peer-reviewed studies about Covid-19 that first relied on synthetic data.

In North America, a handful of health systems and medical schools, as well as government agencies such as the U.S. Department of Veterans Affairs and National Institutes of Health, are using synthetic healthcare data to study and develop new treatment options.

At Intermountain Healthcare, a health system based in Utah, researchers have used synthetic data produced from MDClone's

platform to advance close to 80 projects aimed at improving patient care, says Mike Phillips, a radiologist by training and a partner of Intermountain's venture fund, which invests in startups that spur healthcare innovation.

In one example, Intermountain used the technology to help create a preventative-care program for patients with kidney disease that dramatically reduced hospital admission for dialysis, saving tens of millions of dollars, he says. In that case, access to synthetic data allowed a wider circle of Intermountain employees, such as those looking at the financial and logistical impact, to advance the initiative.

"By democratizing the data, you can get to these changes much faster," Mr. Phillips says.

The NIH and the Bill and Melinda Gates Foundation, meanwhile, are working with California startup Syntegra to synthesize a data set of millions of Covid-19 patients that will be made available to researchers seeking insights into the disease.

Michael Lesh, the founder and chief executive of Syntegra and a professor of medicine at the University of California San Francisco, also sees a role for synthetic data in clinical drug trials. In cases where drug companies are struggling to find study subjects, especially those willing to take the placebo, synthetic patient populations offer a potential solution, he says.

"It's truly radical. It could shave years off the drug-approval process," says Dr. Lesh, adding that his company has discussed the potential use of synthetic data in clinical trials with the U.S. Food and Drug Administration.

The FDA said in statement that it's "following developments" in the synthetic-data space and "would encourage anyone seeking to use synthetic data approaches in a regulatory submission to FDA to seek advice from the relevant center at the FDA." The agency said it doesn't comment on talks with private companies.

Close enough

Not all synthetic healthcare data is built from the medical histories of real patients.

Mitre Corp., a U.S. nonprofit that oversees federally funded research-and-development projects across a variety of fields, has created Synthea, an open-source tool that can generate populations of fictional patients that aren't based on actual medical records.

Instead, Synthea enables researchers to create realistic patients from scratch, using health and disease statistics, demographic data, academic research and other publicly available data sources. The program can simulate what usually occurs to people with various health conditions from birth to death. Since the technology is open source, it is continually being tweaked by researchers to create more-accurate disease models.

While data created by Synthea isn't yet as realistic as synthetic data that originates from real patient records, the U.S. government has taken an interest in the technology. In January, the U.S. Department of Health and Human Services offered $100,000 in cash prizes to teams or individuals that could either improve on Synthea's technology or develop novel uses for data generated by the tool.

The marriage of healthcare and synthetic-data technology comes at a time when faith in the current methods for protecting privacy of healthcare data is eroding.

Deven McGraw, a former senior privacy regulator at the U.S. Department of Health and Human Services, says that with privacy concerns growing, more focus should be put on synthetic healthcare data in the U.S.

"I think the potential of synthetic data for enabling data analytics in a more privacy-protective way is enormous, and it could help ease public concerns around sharing health data with tech companies," says Ms. McGraw, who is now the chief regulatory officer at Ciitizen, a company that helps clients better manage their own health data. (She says she doesn't work with synthetic data in her current role.)

Still, some privacy experts worry that synthetic healthcare data in general hasn't been sufficiently vetted to ensure that the anonymity it promises is irreversible.

"My impression is that the work being done in this space has a lot of potential but it isn't being specifically scrutinized," says Katrina Ligett, associate professor of computer science at Hebrew University of Jerusalem, and head of the program on internet and society.

Prof. Ligett argues that mathematically speaking, you cannot have full privacy without introducing randomness into the data that would impact accuracy.

But practitioners of synthetic-data science say they aren't trying to create an exact mirror image, just something that approximates the real thing.

"Synthetic data has to tell the same story as the original data. It's never 100% identical. The key is it has to be close enough," says Ziv Ofek, CEO of MDClone.

Mr. Lieber is a reporter for The Wall Street Journal in Tel Aviv.