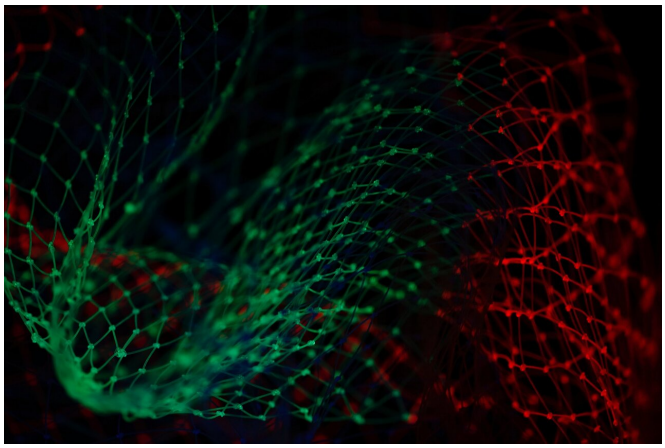
Synthetic data mimics real health-care data without patient-privacy concerns

4 June 2021, by Julia Evangelou Strait



Credit: Unsplash/CC0 Public Domain

The COVID-19 pandemic has accelerated the need to quickly understand how best to fight the virus, but it also presents challenges to initiating studies involving actual patients, such as obtaining consent when patients are critically ill or recruiting patients who may be reluctant to leave their homes.

But what if some research could be conducted using synthetic datasets that mimic real patient populations but don't carry the risk of disclosing protected health information? That's the aim behind an initiative at the Institute for Informatics at Washington University School of Medicine in St. Louis. The institute is making synthetic datasets more widely available to university researchers, with the goal of speeding up research that could save lives.

The institute has shown that software, called MDClone, can accurately produce synthetic data based on real patient data in electronic health records.

In a study published recently in the *Journal of the American Medical Informatics Association*: Open, researchers at the Institute for Informatics showed that synthetic data accurately mimicked the results of clinical studies that had been performed using the real patient datasets.

Rather than take traditional steps to conceal the identities of real patients in the dataset, the software instead produces a new set of simulated patients that, in aggregate, recreate the characteristics of the real patients, such as measures of body mass index, blood pressure and kidney function. These simulated patients have no direct counterparts in the real data, so the real patients' identities and privacy are protected.

"We've realized the power of synthetic data to accelerate the process of asking and answering questions involving real patient data," said senior author Philip R.O. Payne, the Janet and Bernard Becker Professor and director of Washington University's Institute for Informatics. "Instead of taking weeks and months, we're able to interact with data in real time, while also maintaining the highest levels of privacy and data security.

"We want to ensure that every investigator at Washington University has access to these same capabilities, in order to advance research and discovery across a range of diseases, conditions and populations," he said. "We are working hard to reach out to our research community and help them to access this new capability, and look forward to a future in which the use of this software becomes the standard for assessing hypotheses involving clinical data."

The university is collaborating with MDClone, the company that provides this software for research use. The approach used by the company's software to generate synthetic data, as well as the computational and network environments where the software is used, have been designed to comply with the strictest patient privacy and confidentiality

requirements. As a result, there is no way to tie any synthetic data back to real people and their identities. However, investigators do complete a training curriculum and sign a data use agreement that ensure such synthetic data is used responsibly and for scientific research purposes only.

Researchers could run queries asking, for example, which hospitalized patients with COVID-19 are at highest risk of death, or which drugs correlate with better outcomes for patients with COVID-19.

"Through this system, researchers can build their own queries and download synthetic datasets within minutes or hours," said first author Randi E. Foraker, associate professor of medicine and director of the Center for Population Health Informatics. "It really accelerates the research process. What might normally take months can be done same day, sometimes in a matter of minutes, with synthetic data."

The recent study compared the results of analyses on three different datasets. The first dataset was used to analyze the risk of death among pediatric trauma patients. The second dataset was harnessed to predict which hospitalized patients were most likely to develop sepsis, a life-threatening systemic response to infection. And the third was used to produce a map of rates of chlamydia infections by ZIP code in the St. Louis region over a single year.

The researchers found that the results of the synthetic data analyses were statistically similar to the analyses of the real data, drawing the same conclusions using either type of data. In more than one situation, the results were identical, and in only rare cases was there a statistical difference found between the real and synthetic datasets.

"Our three analyses demonstrated that the synthetic data performed well relative to the original data, but we're still testing the outer limits of what synthetic data can do," Foraker said. "It's not a guarantee that in every scenario the synthetic data will fully mimic the original data. We encourage researchers to run their own validation studies. If researchers want to run queries on synthetic data, get some preliminary results or generate some hypotheses before requesting access to real data, that would be a good use of this platform. It's also an excellent resource for students to get the opportunity to work with real-world patient data."

More information: Randi E Foraker et al, Spot the difference: comparing results of analyses from real patient data and synthetic derivatives, *JAMIA Open* (2020). DOI: 10.1093/jamiaopen/ooaa060

Provided by Washington University School of Medicine in St. Louis