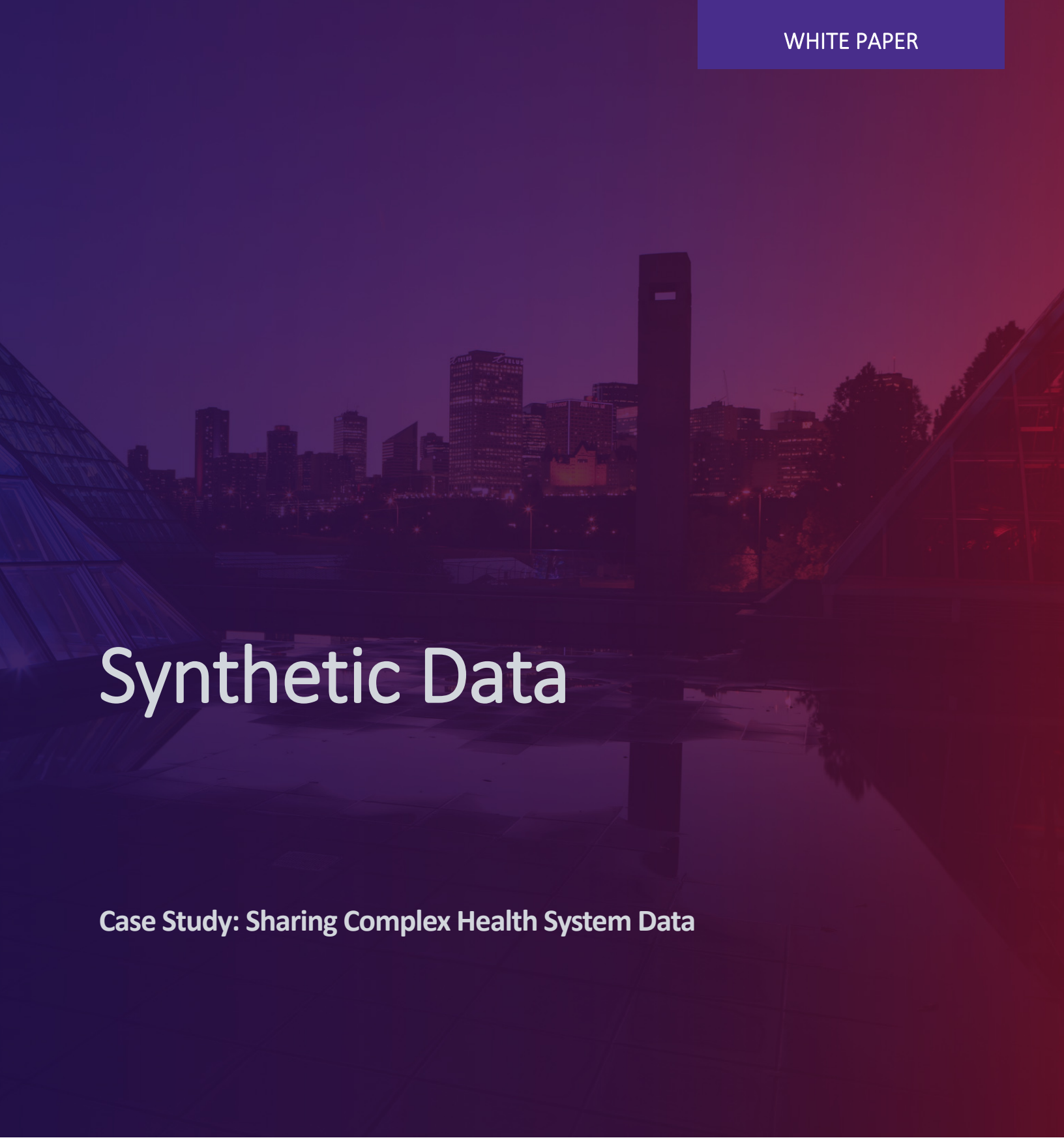
WHITE PAPER



Synthetic Data

Case Study: Sharing Complex Health System Data

Health City

Synthetic Data

Case Study: Sharing Complex Health System Data

Background

Accessing individual level-health data is a process that can be convoluted and time consuming and can act as a significant barrier to the health innovation community. Data accessibility is a difficult endeavor and many organizations that have shown interest in health data have not been able to harness the potential power of this data.

New methods, such as synthetic data generation, have the potential to unlock the historically siloed and difficult-to-access data sets, and provide a channel for readily available access in a secure and reliable manner. Synthetic data is not considered to be personal information because there is no one-to-one mapping between the synthetic records and real people.

In this R&D project, we wanted to answer the question of whether synthetic data can be analytically useful and at the same time protect patient privacy. We synthesized a complex longitudinal health system dataset and evaluated its utility in a typical series of epidemiological analyses.

Objectives

The benefits of demonstrating that synthetic data can replicate estimates derived from longitudinal real-world -individual level data are:

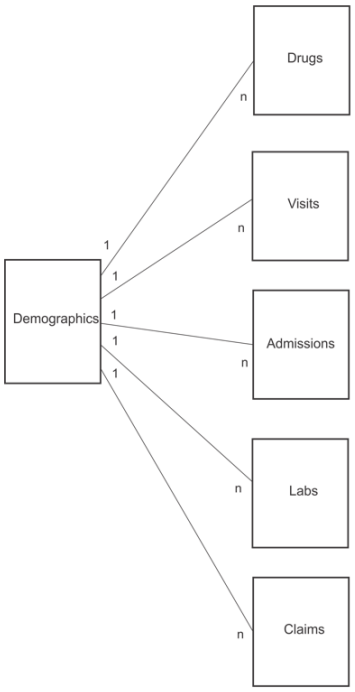
1) It would allow data to be made portable and, as a result, these datasets could be shared with researchers and industry partners outside of their respective jurisdictions without concerns related to the HIA or privacy - facilitating other research which is currently very difficult or not happening outside of the jurisdictions.
2) By allowing this data to be portable, researchers would be able to combine individual level data from multiple jurisdictions to conduct analyses at the individual level. This would be particularly appealing for exposures or endpoints that have low frequencies in any one population as the combination of the populations would increase statistical power and allow for the identification of rare but important exposures or events in populations. This also overcomes the current approach of running underpowered models within each jurisdiction and then combining the results through meta-analytical techniques.
3) From an economic development perspective, synthetic data enables investigation from the business and health innovation community to accelerate activities and levels the playing field for smaller companies or start-ups to enable product development, evaluate product market fit, and related commercialization activities.
4) Positive results also reduce the privacy and security risk and can attract organizations seeking access to health data to the province and country, leading to related economic development outcomes.

The project was performed as a collaboration among multiple organizations, including IHE, Health City, Replica Analytics, University of Alberta, and Alberta Innovates. Along the way, consultations with the provincial privacy commissioner were key to ensuring that the approach taken was transparent and benefited from regulatory feedback.

The Project and Data

The specific project to kick-off this effort focused on a single complex health dataset that combined administrative and clinical information. The dataset that needed to be synthesized consisted of 300,000 patient records covering multiple events per patient. Some patients had a handful of events and others had tens of thousands of events over a seven-year period. The domains covered in the data included drugs, laboratory results, emergency department visits, hospital admissions, and doctors' visits.

Data synthesis proceeds in two general steps. The first is to train a generative model on the original data. This training captures the patterns in the original dataset. Then the generative model is used to synthesize a new dataset that is based on the patterns that were captured during training. The generated data comes directly from the model and is not derived from the original data.



Demographics
Age
Sex
Time to last day of follow-up available
Comorbidity score (elixhauser)

Drugs
Dispensed amount quantity
Relative dispensed time in days
Dispensed day supply quantity
Morphine use (binary)
Oxycodone use (binary)
Antidepressant use (binary)

Visits (ED)
Relative admission time in days
Problem code 1
Problem code 2
Resource intensity weights

Admissions (Hospital)
Relative time admitted in days
LOS
Diagnosis code 1
Diagnosis code 2
Resource intensity weight

Lab
Test name
Test result (integer)
Relative time in days lab taken

Claims
Primary diagnosis code
Provide specialty
Relative service event start date

The objective was to create a generative model from this dataset such that the synthesized data would replicate analytic patterns. Specifically, the synthetic data was assessed based on how well models of mortality and other events (such as hospitalization) agreed with models built using the real data.

Solution & Outcomes

A deep learning model was developed to generate synthetic data. The model captured the baseline characteristics of the patients as well as their sequence of events and event attributes. Novel approaches were needed to address the heterogeneity in the data, and to leverage the history of events to generate valid subsequent events for each patient.

To validate the approach, some general comparisons of the original and synthetic data were performed. The comparisons showed that both datasets were quite similar. In addition, Cox regression models to predict various outcomes were also developed. The real and synthetic model were very similar with high confidence interval overlap (see the summary in the sidebar).

To address privacy concerns, a privacy risk assessment was performed to evaluate the likelihood that records in the synthetic dataset can be matched with real individuals. The results of that assessment demonstrated that the meaningful identity disclosure privacy risks were below commonly used risk thresholds by approximately an order of magnitude (see the summary in the sidebar).

Overall, we were able to demonstrate that a deep learning generative model can capture the key characteristics of a complex longitudinal health dataset and generate realistic synthetic variants. The synthetic variants had an acceptably low identity disclosure risk.

This approach allows data users to access the synthetic data with minimal constraints, but still provide privacy protection. As the technology is scaled, this will provide a means to rapidly make data available to a broad community of users and drive innovation within the province.

How Safe is Synthetic Data?

As part of this project, we evaluated the privacy risks of the generated synthetic data. The focus was estimating the probability to which a synthetic record can be correctly linked to a real person *and* learning something new about that person from the synthetic data. We determined that this probability was 0.0019 (less than half a percent), which is very low and lower than generally accepted thresholds for privacy risks. In general, it is always good to be prudent and evaluate the privacy risks in synthetic data, and we have a well-documented methodology for doing so.

Reference

[1] K. El Emam, L. Mosquera, and J. Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” JMIR, vol. 22, no. 11, Nov. 2020. [Online]. Available: https://www.jmir.org/2020/11/e23139.

How Good is the Quality of Synthetic Data?

The synthetic data was evaluated using multiple approaches. The first is to use generic metrics that compare the structure of the synthetic data to the real data. For example, we compare whether the number and type of events in the synthetic data are similar to those in the real data. The second approach is to perform a substantive epidemiological analysis on both datasets and see if we would draw the same conclusions. In the current project we built models to predict mortality and hospitalization.

Our results indicated that the synthetic data was similar to the real data structurally, and the substantive conclusions that would be drawn from the statistical models would be the same, even for complex multivariate models. This gives us some assurance that the synthetic data can be a useful proxy for real data in a number of specific use cases.

Reference

[1] L. Mosquera et al., "A Method for Generating Synthetic Longitudinal Health Data", (submitted for publication), Nov. 2020.

Future Opportunities

We see synthetic data as a precursor 'funnel' to *bona fide* health system data. Many health data requests are not developed or robust enough for direct access to health data. Yet, it is difficult for potential users to have sufficient data literacy without access to actual data (chicken/egg problem). We believe synthetic data addresses this problem by enabling users to better understand and refine their project or study parameters enabling them to have a well-developed data request prior to using actual health data.

Deploying synthetic data sets in areas of public health concerns will enable provincial health systems to freely partner with academic institutions, institutes or philanthropic organizations, or industry to broaden their talent pool of data scientists as well as data sets (such as social indicators of health).

A similar access challenge exists with artificial intelligence and machine learning projects. These projects often do not have a targeted question with defined parameters established but rather, are exploratory in nature and look for compelling artifacts or correlations in the data that one would not typically predict. However, securing ethics approval for such studies can be challenging. From an academic perspective this restricts the ability of exploratory data science research and limits our ability to train the next generation of health data scientists. Synthetic data sets would offer a safe way to make health data broadly available to many researchers and students for training and data literacy.

The concept of virtual clinical trials has been around for some time but here too, personal health data privacy is also a concern. The ability to leverage synthetic data techniques to better design clinical trials by testing hypothesis on synthesized data sets has the potential to enable rapid study design before ever touching a patient. The ability to integrate patient-facing input such as apps or other forms of data capture while maintaining privacy is a significant opportunity to transform how clinical trials are conducted.

Health City

Health City is a Canadian not-for-profit Corporation that works with clinicians, innovators, philanthropic organizations, and companies to develop new pathways of care that can drive better health outcomes and economic development in the health sector. Our focus is on transforming innovations from our health sector into solutions that have commercial application and global relevance, adopting them for impact locally and scaling them for export to global markets.

For more information, visit www.edmontonhealthcity.ca.