Alberta Synthetic Data Research Project

Multi-party collaboration between Health City, IHE, Alberta Innovates, University of Alberta & Replica Analytics

July 7, 2021
Agenda

- Overview of the Alberta Synthetic Data Research Project
- What is synthetic data? What are some of the use cases for synthetic data in Alberta?
  - How complex was the health system data synthesized?
  - How was the data synthesized?
  - How good is the quality of synthetic data?
  - What are the privacy risks of synthetic data?
Research Project Overview

- **Opportunity:**
  - Synthetic data has potential to support greater sharing and utilization of Alberta data assets, including with organizations outside of the research community (e.g., life sciences industry).

- **Objectives:**
  - Develop and interrogate a synthetic dataset in order to understand opportunities and limitations:
  - Explore processes for generating synthetic data that is representative of an existing Alberta health dataset
  - Identify any key privacy and security concerns of key groups in Alberta
  - Analyze and validate the synthetic data set to understand how representative it is of the original data set to understand future utility.
Project Partners/Structure

**Project Steering Committee**
Project Sponsors - Health City & IHE
Strategic Partner - Alberta Innovates

**Data Stakeholders Group**
- OIPC
- Alberta Health
- Alberta Health Services
- University of Alberta

**Core Project Team**
- Project Manager - Health City
- Data Custodian/Investigator – U of A
- Privacy Specialist – Alberta Innovates
- Project Advisor & Analyst - IHE
- Synthetic Data Vendor – Replica Analytics
**Aim:** To evaluate the ability of synthetic data to simulate actual anonymized patient-level data by replicating analyses and comparing outcomes (death, ED Visits, hospitalizations) in a typical time to event study.
What is Synthetic Data?
Synthetic Data

---

**Source Data**

- Fit Model
- Apply Model

**Synthetic Data**

<table>
<thead>
<tr>
<th>COU1A</th>
<th>AGECAT</th>
<th>AGELE70</th>
<th>WHITE</th>
<th>MALE</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>33.75155</td>
</tr>
<tr>
<td>United States</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>39.24707</td>
</tr>
<tr>
<td>United States</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>26.5625</td>
</tr>
<tr>
<td>United States</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>40.58273</td>
</tr>
<tr>
<td>United States</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>24.42046</td>
</tr>
<tr>
<td>United States</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>19.07124</td>
</tr>
<tr>
<td>United States</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>26.04938</td>
</tr>
<tr>
<td>United States</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25.46939</td>
</tr>
</tbody>
</table>
Use Cases for Synthetic Data

- Facilitates data sharing when privacy is a concern
  - Allows external researchers to gain access to data more quickly
  - Allows sharing of data for training purposes within the data owner’s organization
  - Allows collaboration with external vendors for technology evaluations

- Data amplification and augmentation
  - Generate additional observations at low cost, accelerating the use of machine learning tools
  - Can be applied to amplify the presence of subgroups of interest based on key outcomes or traits
How Complex was the Data Synthesized?
Methods - Data Reduction

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
- Relative admission time in days
- Problem code 1
- Problem code 2
- Resource intensity weights

Admissions
- Relative time admitted in days
- LOS
- Diagnosis code 1
- Diagnosis code 2
- Resource intensity weight

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

Claims
- Primary diagnosis code
- Provider specialty
- Relative service event start date
Methods – Patient Selection & Outcomes

• A random subset of ~ 80,000 subjects who received a dispensation for Opioid 1 or Opioid 2 between Jan 1, 2016 and Dec 31, 2017, 18 years of age and over were included in our analyses.

• Our primary outcome was a composite endpoint of time to all-cause emergency department visit, hospitalization, or death during the follow-up.
  • The secondary outcomes included each component of the composite endpoint separately, as well as to evaluate cause specific admissions to hospital for pneumonia (J14.9) as a prototypical example of a cause specific endpoint.
How was the Data Synthesized?
Context & Approach

• Synthesis of a complex dataset:
  • Five correlated event sequences (admissions, emergency visits, claims, drugs, and lab data)
  • Large variation in the number of events per patient
  • Heterogeneous variables types
  • Variation in the number and type of attributes per event

• No known state-of-the-art methods could be used to create a fully synthetic version of data this complexity

• Developed a novel deep learning model in collaboration with a consultant at MILA in Montreal
Deep Learning Model

• Trained to predict next event for each patient based on their previous events and baseline traits
• Model updates based on loss observed in training data. Training stops when loss on a hold out sample is minimized.

Note: red denotes real data
Synthetic Data Generation

- Sequential decision trees are used to generate synthetic baseline characteristics and first.
- Repeatedly fed into deep learning model to synthesize events throughout the study period

Note: blue denotes synthetic data
How Good is the Synthetic Data?
Generic Utility
Results Summary

- Clockwise from top right:
  - Comparison of event distribution
  - Comparison of sequence lengths per individual
  - Hellinger distance of event attributes
    - 0 = no difference in distribution
    - 1 = biggest possible difference in distribution
  - Comparison of Markov transition matrices
Methods – Analytical Comparison

• Using Cox proportional hazards regression models, unadjusted and adjusted hazard ratios (HRs) and 95% CIs were calculated to assess the risk associated with either Opioid 1 or Opioid 2 and our outcomes of interest in both the synthetic and real data separately.

• All subjects were prospectively followed until outcome of interest, death, or censoring defined as the date of termination of Alberta Health coverage or 31 March 2018, providing a maximum follow-up of 2 years.

• Finally, the estimates derived from the real and synthetic datasets were directly statistically compare. Opioid 1 served as the reference group for all estimates.

• Potential confounding variables included in all multivariate models included age, sex, Elixhauser comorbidity score, use of antidepressant medications, and our 3 laboratory variables (ALT, eGFR, HCT). All analyses were performed using STATA/MP 15.1 (StataCorp., College Station, TX).
### Results

<table>
<thead>
<tr>
<th></th>
<th>Real N = 75,660</th>
<th>Synthetic N = 75,660</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>43.32 (17.87)</td>
<td>44.79 (19.83)</td>
</tr>
<tr>
<td><strong>Sex = Male</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>37,037 (49.0)</td>
<td>35,949 (47.5)</td>
</tr>
<tr>
<td><strong>Elixhauser Score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>0.96 (1.58)</td>
<td>1.05 (1.63)</td>
</tr>
<tr>
<td><strong>ALT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>31.67 (63.90)</td>
<td>40.72 (111.92)</td>
</tr>
<tr>
<td><strong>GFR</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>85.82 (23.56)</td>
<td>83.11 (25.05)</td>
</tr>
<tr>
<td><strong>HCT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>0.42 (0.05)</td>
<td>0.41 (0.06)</td>
</tr>
<tr>
<td><strong>Opioid</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 N (%)</td>
<td>1,758 (2.3)</td>
<td>2,649 (3.5)</td>
</tr>
<tr>
<td>Group 2 N (%)</td>
<td>73,902 (97.7)</td>
<td>73,011 (96.5)</td>
</tr>
<tr>
<td><strong>Antidepressant</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>28,224 (37.3)</td>
<td>29,651 (39.2)</td>
</tr>
</tbody>
</table>
Results – Adjusted Cox Regression

Note: Adjusted estimates include the following co-variates: age, sex, antidepressant use, Elixhauser score, ALT, eGFR, HCT; Opioid 1 served as the reference group.
## Results – Outcome Comparison

<table>
<thead>
<tr>
<th></th>
<th>Real Data ( N = 75,660 )</th>
<th>Synthetic Data ( N = 75,660 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to Death, Days (mean (SD))</td>
<td>1,474.48 (772.23)</td>
<td>1,077.88 (722.44)</td>
</tr>
<tr>
<td>Death</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>3,299 (4.4)</td>
<td>1,440 (1.9)</td>
</tr>
<tr>
<td>Hospitalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>22,495 (29.7)</td>
<td>21,582 (28.5)</td>
</tr>
<tr>
<td>Emergency Room Visits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>64,376 (85.1)</td>
<td>65,193 (86.2)</td>
</tr>
<tr>
<td>Composite Endpoint</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(%)</td>
<td>64,848 (85.7)</td>
<td>65,497 (86.6)</td>
</tr>
<tr>
<td>Hospitalization Related to Pneumonia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N (%)</td>
<td>505 (2.2)</td>
<td>472 (2.2)</td>
</tr>
</tbody>
</table>
What are the Privacy Risks?
Privacy Scan

- AHS is custodian of original, non-identifying data
- Project under auspices of existing REB approval, amended to include creation of synthetic data
- Agreement in place between AHS and researcher, per HIA
- Agreement in place between researcher and Replica Analytics that includes clauses for confidentiality, security, monitoring, incident response and prohibition against attempting to re-identify data subjects
- Synthetic data set tested against risk of re-identification
  - Privacy Assurance test found the risk of reidentification to be quite low
- Synthetic data remains in protected environment, in custody of researcher
- This project relied on HIA research provisions
Information Flow

1. Original data from Alberta Health Services
2. Non-identifying HI
3. HIA s. 54
4. Synthetic data

REB Approval & Research Agreement

Researcher

Synthetic Data Summit | 24
Acknowledgements
Thank You