Synthetic Data

Dean T Eurich
Program Lead, Clinical Epidemiology, School of Public Health
Elected Member Royal Society of Canada College of New Scholars, Artists and Scientists
Professor, School of Public Health

Sept 13, 2021

Research Interest & Need for Synthetic Data

Research Interest Overview

- Health Services Research
  - Improving the efficiency and effectiveness of health professionals and the health care system, through changes to practice and policy.
  - Major areas of interest are chronic diseases (diabetes, CVD, inflammatory conditions), as well as infectious diseases (pneumonia and IPD)
  - Substantial research in FN and sensitive data around FN communities and chronic disease care
    - Clinical trials
    - Quality improvement programs
    - DATA – LOTS OF IT
      - Small clinical datasets and registries collected from primary sources (e.g. medical charts, devices, aps)
      - BIG DATA – secondary use of large national and international datasets on entire populations (e.g., GPRD data in the UK, HMO databases from the US, Provincial administrative databases in AB, SK, ON, BC, MB, QC).

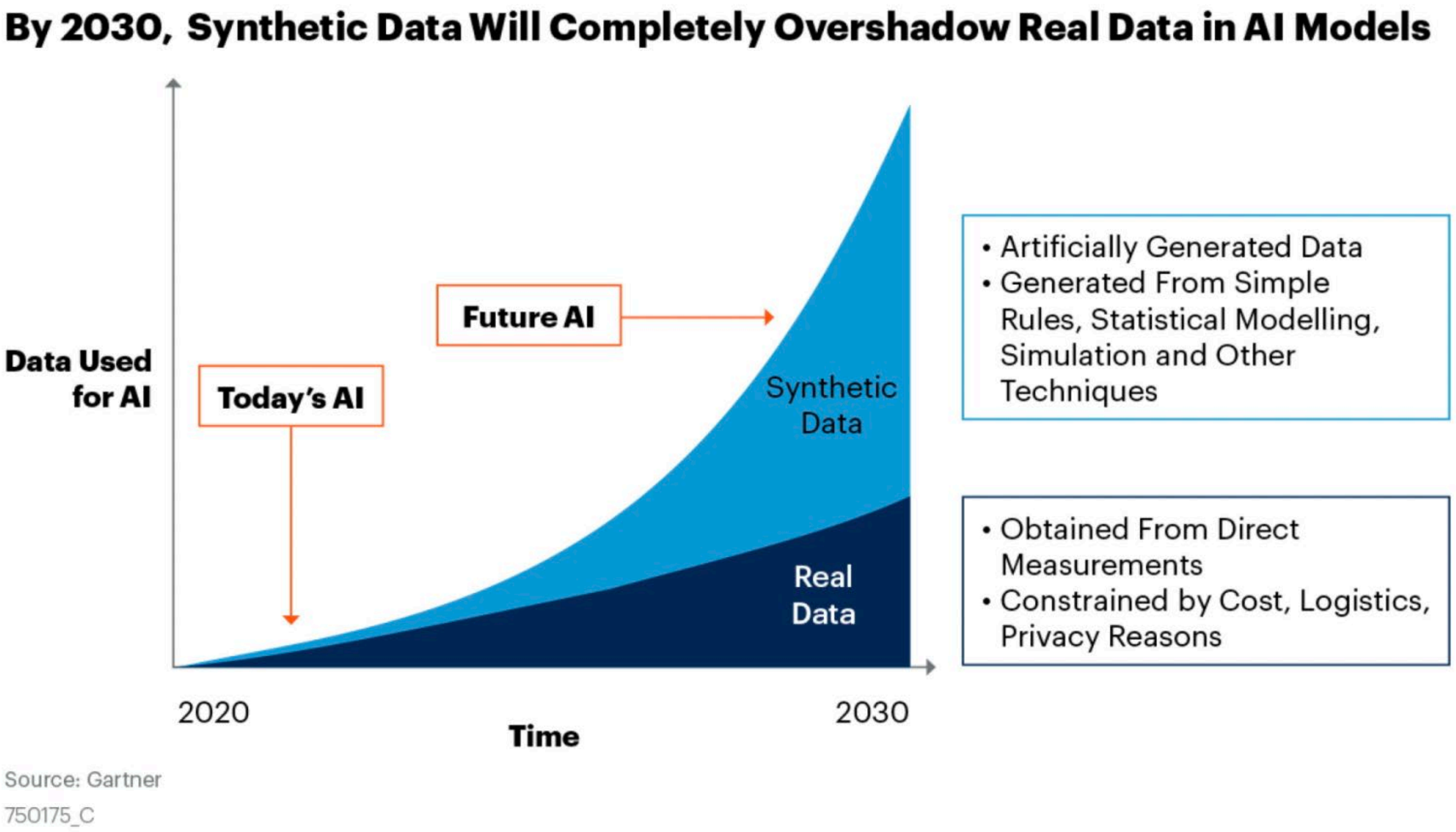
Major Issues around Health Data

- Increasingly becoming more difficult (and slow!) to access data sources due to legal and privacy concerns
  - Diseases are changing – more and more ‘niche and rare’ diseases
    - Difficult to work with small populations and data
    - Increased privacy concerns with limited number of patients in a population
  - Data is changing – no longer just ‘health data’
    - Environmental, educational, justice, phone apps, shopping habits – all of this data is increasingly being used in health services research
    - Although datasets are often huge, the risk of privacy is high because so many datapoints are being collected on the individual
  - Data Sharing – extremely difficult in todays legal and privacy frameworks
    - E.g. Alberta Health data, ON ICES data
  - Training – most currently trained data scientists are unprepared for ‘real world data”

Why Synthetic Data

- Facilitates data sharing when privacy is a concern
  - Allows external researchers to gain access to data more quickly
    - Could share data outside our “borders”
    - MUST occur if we are to truly evaluate rare disease not just nationally but internationally
  - Allows for cloud based computing systems
    - Currently not allowed in Ab as AWS, Azure, etc servers are out of province or out of country
    - Increase use of more advanced analytics (e.g., ML approaches to big data)
  - Allows sharing of data for training purposes within the data owner’s organization and potentially elsewhere
  - 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 populations of interest based on key outcomes or traits (e.g., rare diseases)

Synthetic Data is Coming Fast



Gartner

Attributes for Good Synthetic Data

- Speed - needs to overcome our slow process for data access
  - Data pipelines/systems but be in place to generate the synthetic data
- Accurate!
    - What does that mean!!
    - So many different metrics being applied in the synthetic world
  - At a minimum:
    1. Fidelity at the individual sample level (e.g., synthetic data should not include prostate cancer in a female patient),
    2. Fidelity at the population level (e.g., marginal and joint distributions of features)
    3. Correct classification
        - Comorbidities – difficult as 1000's of codes to classify disease
        - Health care utilization – difficult as some have no healthcare use, others >5,000's encounters!
        - Timing – correct timing of events (e.g., certain drugs only used after certain diagnoses)
    4. Privacy!!

BMC Medical Research Methodology **volume 20**, Article number: 108 (2020)

Synthetic Data Initiative in AB

Synthetic Data

- Objectives:
  - Develop and evaluate 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

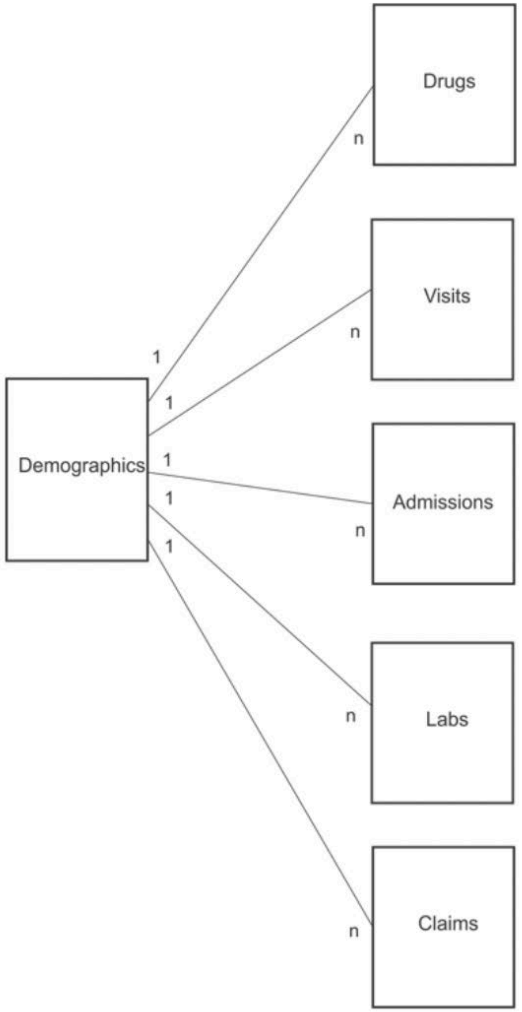
Core Project Team
Project Manager - Health City
Data Custodian/Investigator – U of A
Privacy Specialist –Alberta innovatesData
Project Advisor & Analyst - IHE
Synthetic Data Vendor – Replica Analytics

Synthetic Data Analysis

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

How Complex was the Data Synthesized & Methods Used?

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 (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

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.

Methods – Analytical Comparison

- Using Cox proportional hazards regression models, 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.
- 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).

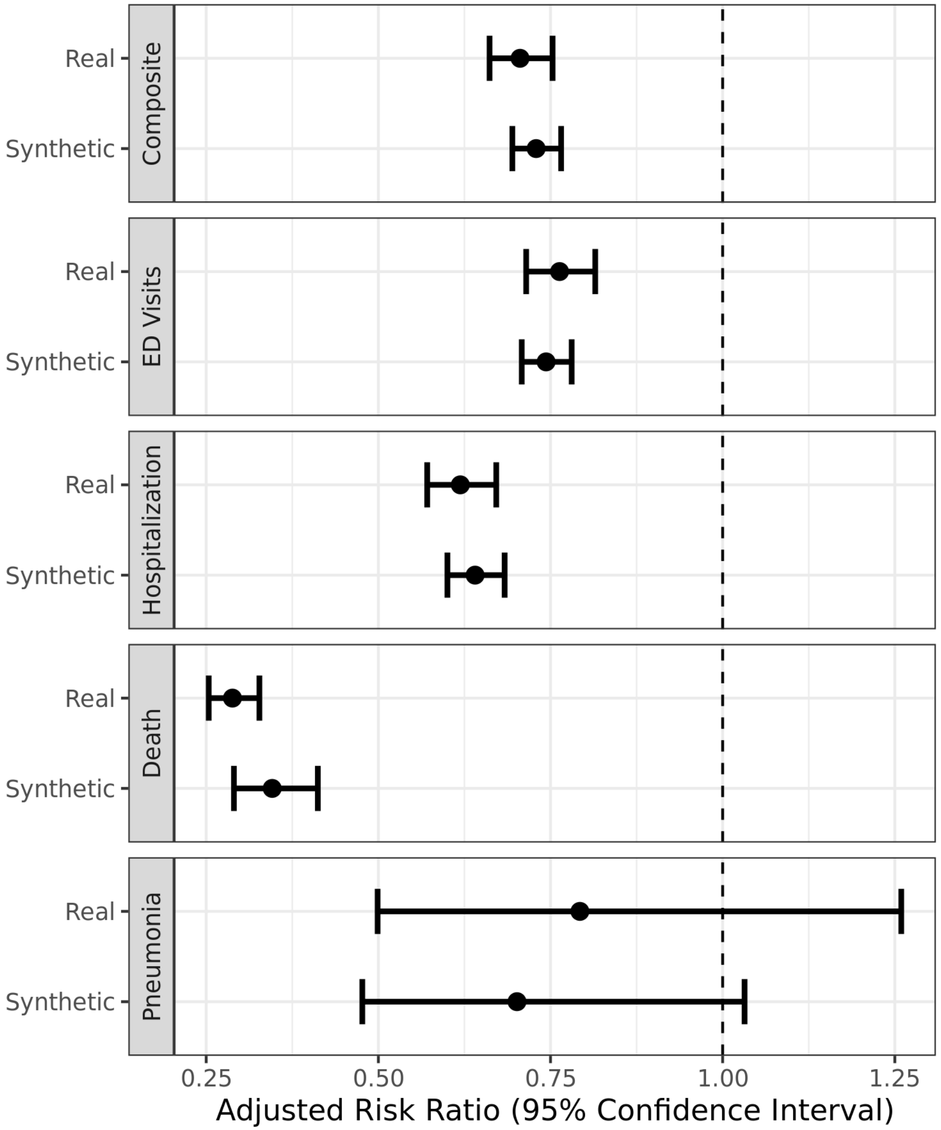
How Good is the Synthetic Data?

Results

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

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

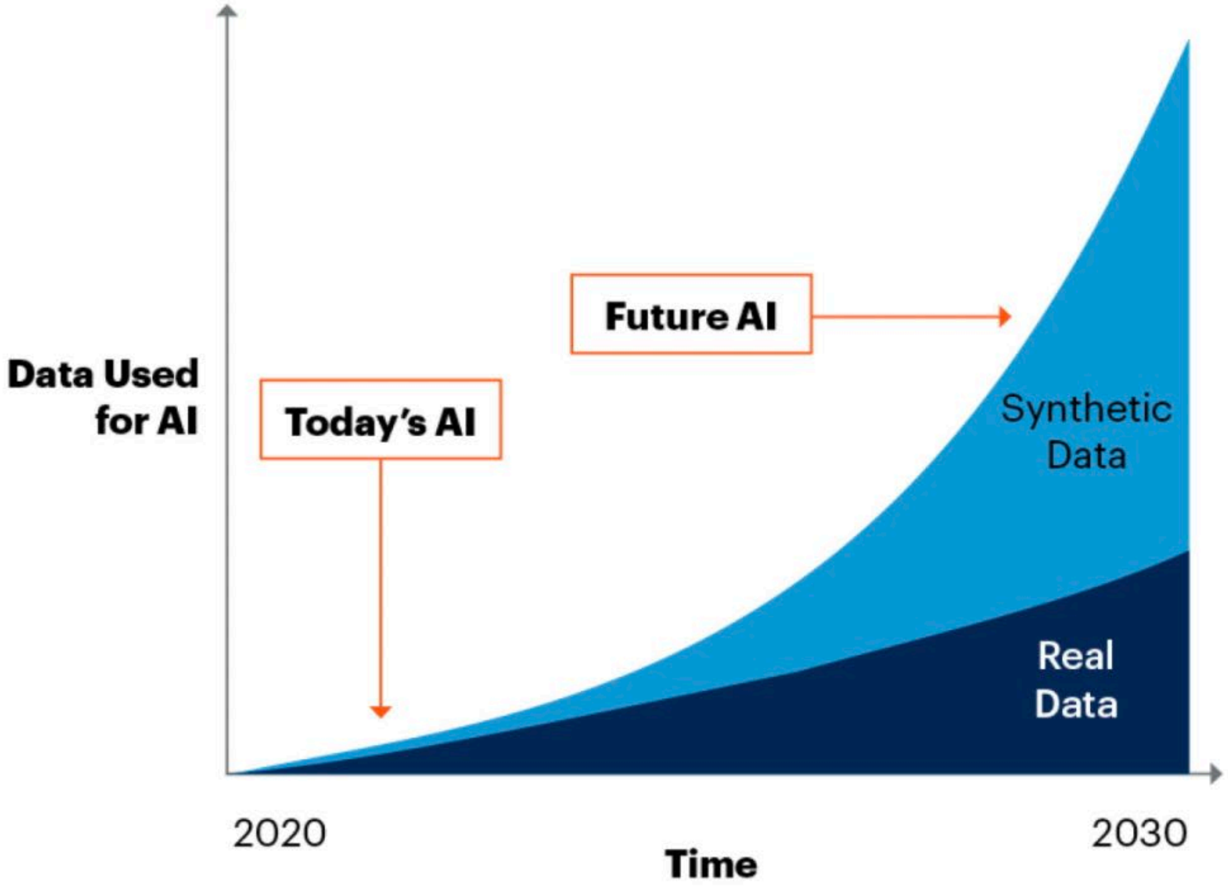


Results – Outcome Comparison

	Real Data N = 75,660	Synthetic Data N = 75,660
Time to Death, Days (mean (SD))	1,474.48 (772.23)	1,077.88 (722.44)
Death N (%)	3,299 (4.4)	1,440 (1.9)
Hospitalization N (%)	22,495 (29.7)	21,582 (28.5)
Emergency Room Visits N (%)	64,376 (85.1)	65,193 (86.2)
Composite Endpoint N(%)	64,848 (85.7)	65,497 (86.6)
Hospitalization Related to Pneumonia N (%)	505 (2.2)	472 (2.2)

Future Utility of Synthetic Data?

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

Source: Gartner

750175_C

Gartner

Future of Synthetic Data

- If synthetic data is shown to be "as good as" real data….the applications will be widespread in health research
  - Training
  - Data sharing – researchers/cooperation's
  - Modeling simulate potential impacts of drugs, devices, policies in populations; simulating clinical trials, etc
  - Discovery research
  - Harness the potential of AI and ML in health

Acknowledgements

Alberta Health Services

Alberta

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF ALBERTA

Alberta Diabetes INSTITUTE

Mila

Thank You