SESSION 5: APPLICATIONS IN COMPLEX HEALTHCARE SETTINGS

GENERATING SYNTHETIC LONGITUDINAL DATA

uOttawa

Replica Analytics
AN AETION COMPANY

Presented by:

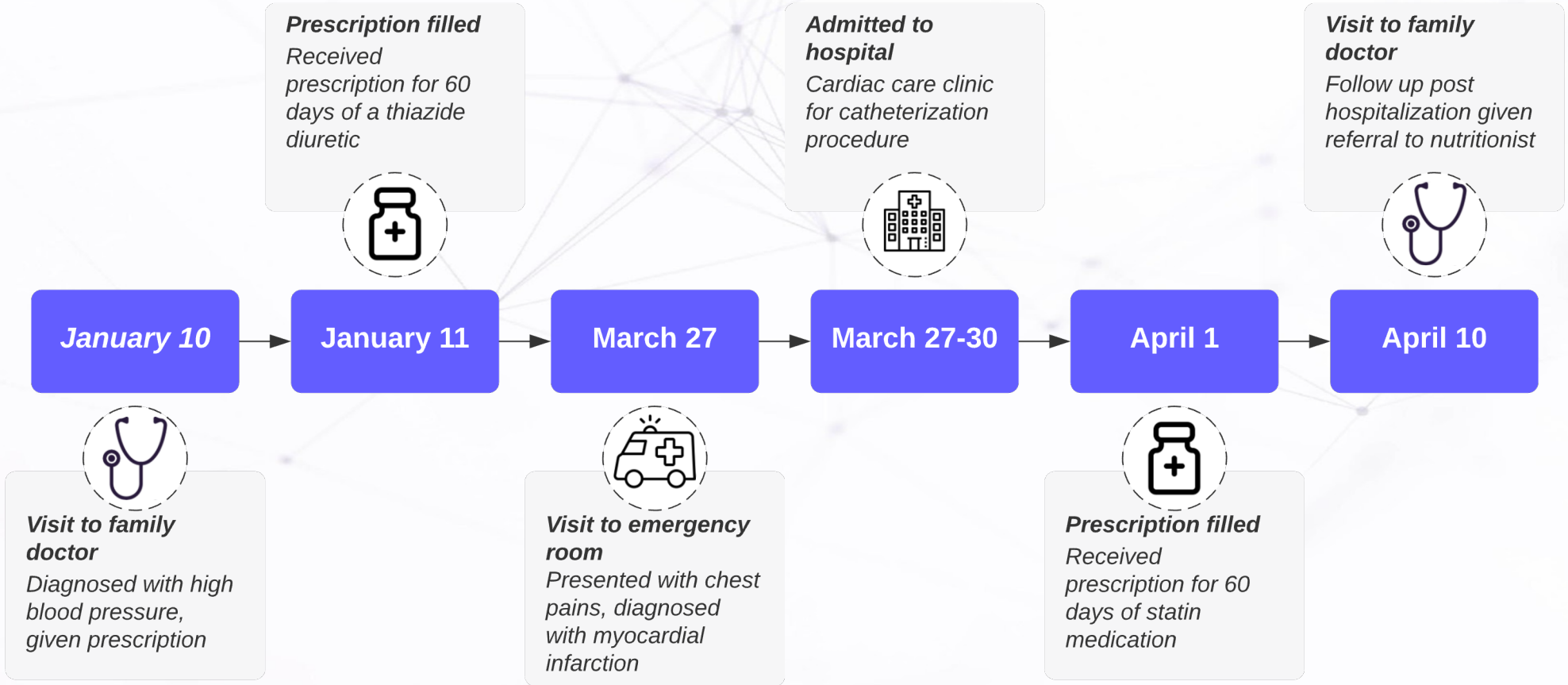
Lucy Mosquera,
Senior Director of Data Science,
Replica Analytics

Agenda

- What is longitudinal health data?
- Why is it difficult to synthesize?
- Case study: large scale single payer EHR synthesis
  - Overview of deep learning model
  - Utility and privacy results

Replica Analytics
AN AETION COMPANY

Longitudinal health data has multiple observations over time



Examples include: electronic medical records, insurance claims data, clinical trial data

Replica Analytics
AN AETION COMPANY

Synthetic data is generated by training a model to learn patterns and relationships in a dataset, then generating new data from that model

The type of model used will determine what patterns and relationships can be preserved in the synthetic data

Replica Analytics
AN AETION COMPANY

What is the challenge with synthesizing longitudinal health data?

Longitudinal health data has many complex relationships we would synthetic data to preserve!

- Long temporal sequences
  - Patients can have hundreds or thousands of visits and interactions with the healthcare system
- Inter-patient variability
  - Different patients even with the same diagnoses can have substantial differences in their data and resource utilization
- High cardinality variables
  - Measures like diagnostic and procedure codes or the prescription drugs received have a high dimensional feature space that can be difficult to model
- Large datasets
  - Potentially huge amounts of data per patient and thousands if not millions of patients can lead to an immense amount of data

Previous work to synthesize longitudinal health data has relied on heavily simplified example datasets (e.g., taking a dozen read points of vital stats during an ICU stay) or pre-specified models built on subject area knowledge (e.g., after a given diagnosis, patients should be treated with X then Y)

Case study: synthesis of single payer health system data

Goal

Assess whether synthetic longitudinal health data could be used to draw the same conclusions as real data when assessing the impact of receipt of two kinds of opioids on health outcomes.

Mosquera *et al.*
BMC Medical Research Methodology (2023) 23:67
https://doi.org/10.1186/s12874-023-01869-w

BMC Medical Research Methodology

RESEARCH Open Access

Check for updates

A method for generating synthetic longitudinal health data

Lucy Mosquera[1,2], Khaled El Emam[1,2,3*], Lei Ding[4], Vishal Sharma[5], Xue Hua Zhang[1], Samer El Kababji[2], Chris Carvalho[6], Brian Hamilton[7], Dan Palfrey[8], Linglong Kong[4], Bei Jiang[4] and Dean T. Eurich[5]

UNIVERSITY OF ALBERTA

Alberta
Freedom To Create. Spirit To Achieve.

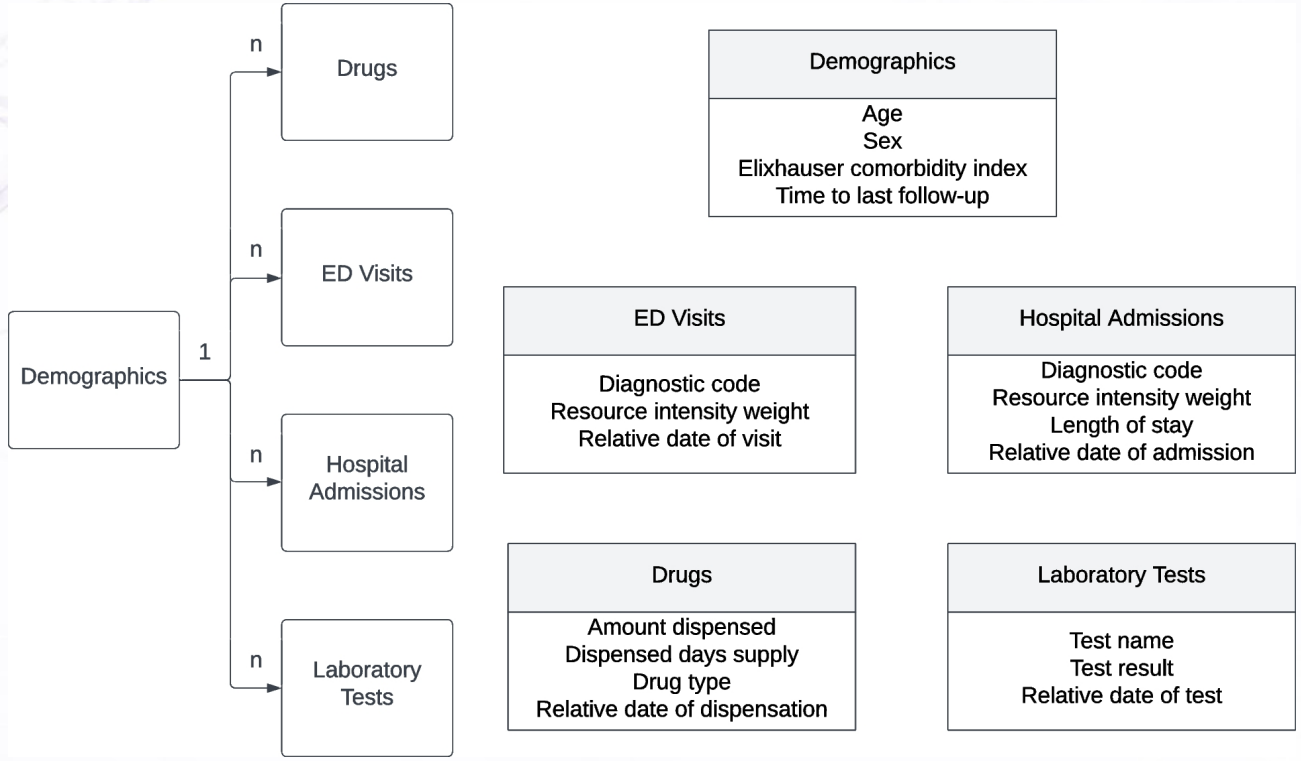
Health City

Replica Analytics
AN AETION COMPANY

Single payer health system data

Data for 100,000 individuals from 2012-2018

Table	# Rows	# Columns
Demographics	100,000	4
Prescriptions	9,975,950	7
ED Visits	1,748,083	5
Hospitalizations	84,669	5
Laboratory tests	2,199,574	3
Coverage	100,000	2
Mortality	4200	6



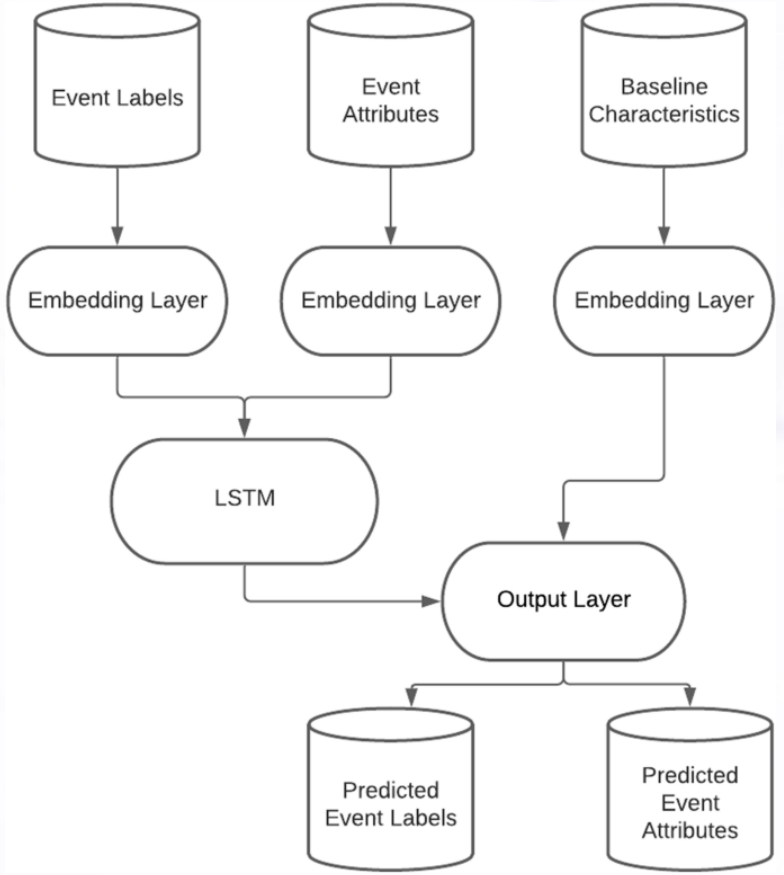
More than 14 million total observations!

Replica Analytics
AN AETION COMPANY

Longitudinal synthesis model

Deep learning model driven by a Long Short-Term Memory (LSTM) model.

- Each patient event is predicted based on the previous events and the baseline characteristics of the patient.
- All features are embedded during training into a multi-dimensional continuous feature space that's easier for the LSTM to learn from
- Predictions include the type or label of each event as well as the associated features. Training can then optimize between learning the event types and event labels
- Comprehensive hyperparameter optimization was conducted using raytune
- Overfitting was prevented using gradient clipping and a validation set to determine when to stop training



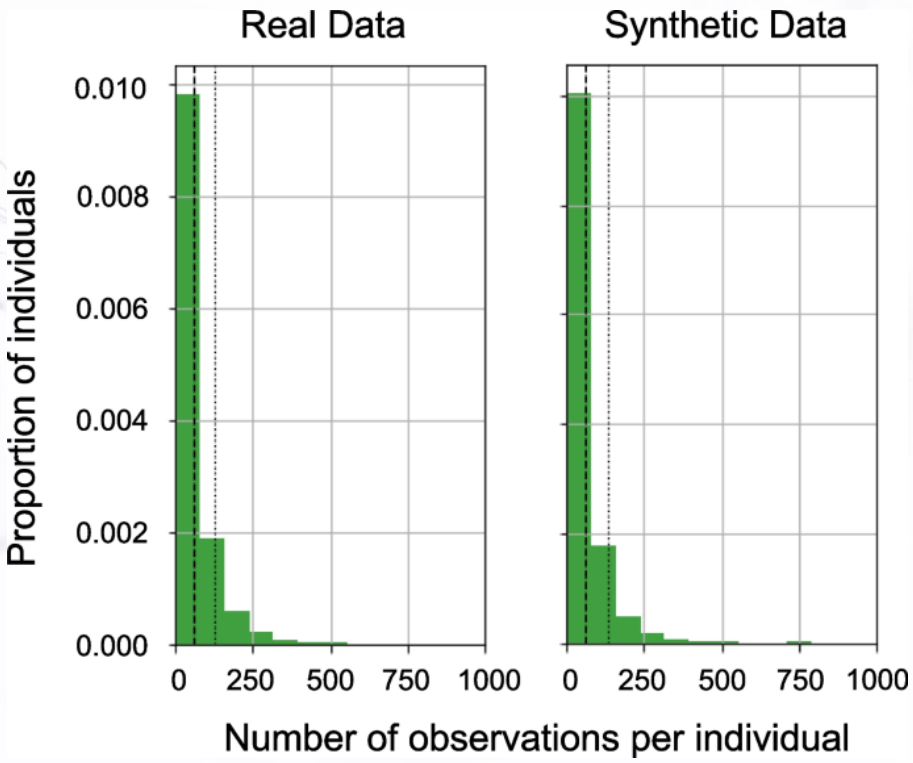
Replica Analytics
AN AETION COMPANY

How similar is the synthetic data to the real?

Assessed similarity using both:

Generic utility assessments compare the real and synthetic datasets without any specific use case in mind

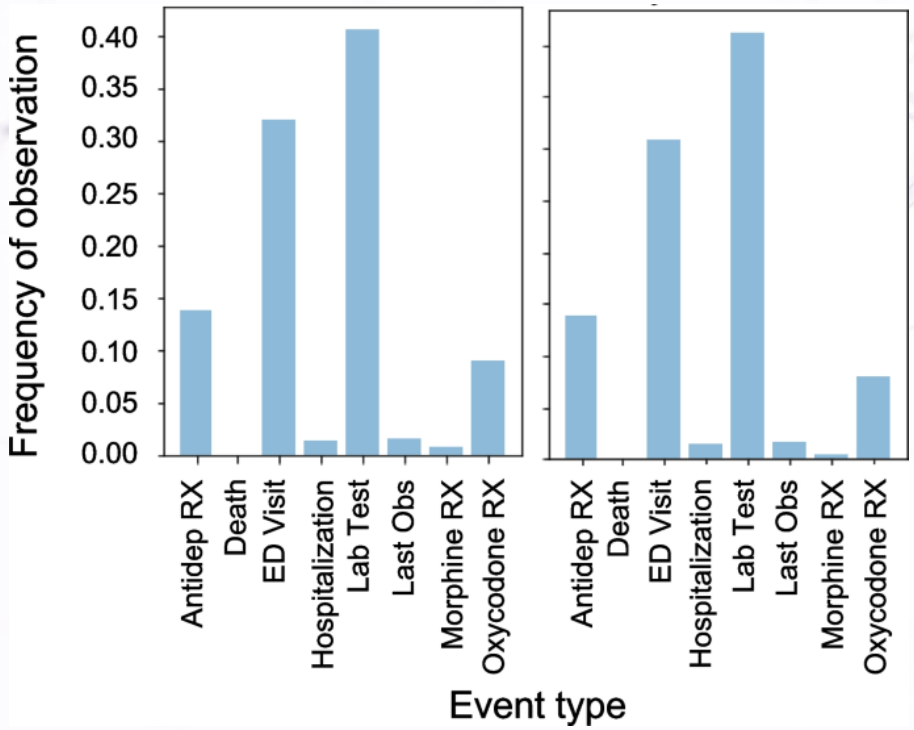
Analysis specific utility assessments that compare the results of a given analysis when conducted using real vs synthetic data



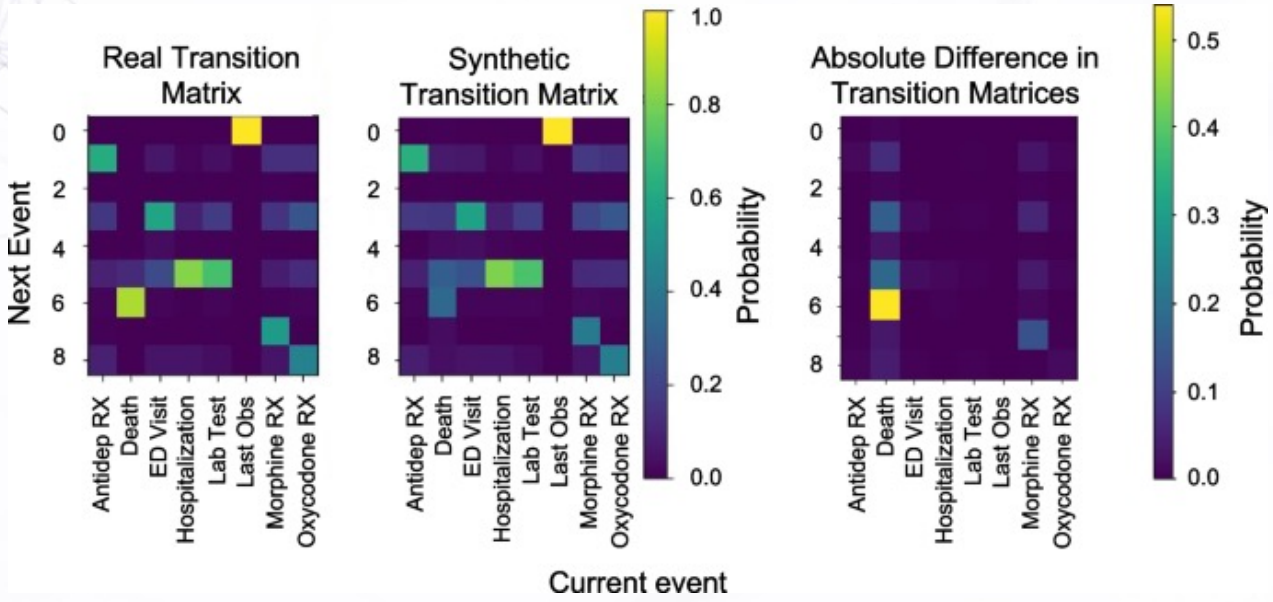
Percent difference in mean total observation per individual : 0.04%

Replica Analytics
AN AETION COMPANY

How similar is the synthetic data to the real?



Hellinger distance between the distribution of event types across all individuals:
0.027



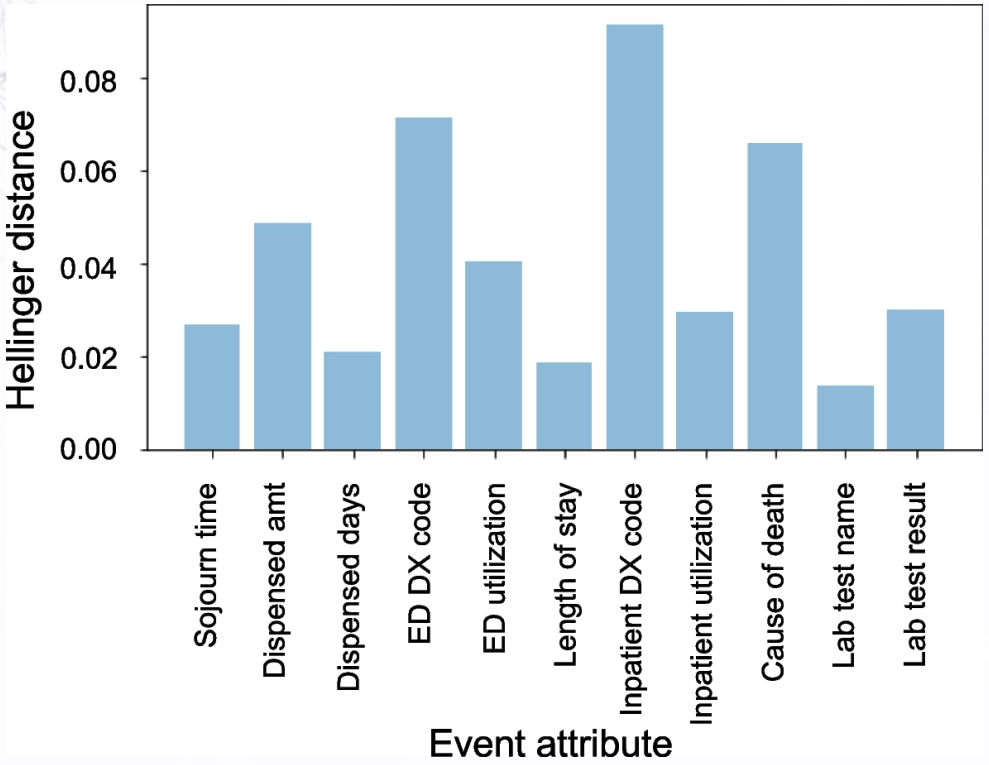
Markov transition matrix comparison:
Mean Hellinger distance 0.0896 (SD: 0.159)

Replica Analytics
AN AETION COMPANY

How similar is the synthetic data to the real?

Mean Hellinger distance between the distribution of event attributes:
0.0417

Overall the generic utility assessments showed that the synthetic data is highly similar to the real data



Replica Analytics
AN AETION COMPANY

How similar is the synthetic data to the real?

With our analysis in mind of comparing outcomes in individuals who were prescribed morphine vs oxycodone

We see very similar distributions between the real and synthetic data in terms of demographics and the rate of prescription of morphine, oxycodone, and antidepressants.

	Real ***n*** = 75,660	Synthetic ***n*** = 75,660	SMD
Age			0.078
Mean (SD)	43.32 (17.87)	44.79 (19.83)	
Median (IQR)	42.00 [27.00]	43.00 [30.00]	
Sex n (%)			0.029
Male	38,623 (51.0)	39,711 (52.5)	
Female	37,037 (49.0)	35,949 (47.5)	
Elixhauser score			0.055
Mean (SD)	0.96 (1.58)	1.05 (1.63)	
Median (IQR)	0.00 [1.00]	0.00 [2.00]	
Opioid Utilization (%)			0.070
Morphine	1758 (2.3)	2649 (3.5)	
Oxycodone	73,902 (97.7)	73,011 (96.5)	
Antidepressant Use	28,224 (37.3)	29,651 (39.2)	0.039

How similar is the synthetic data to the real?

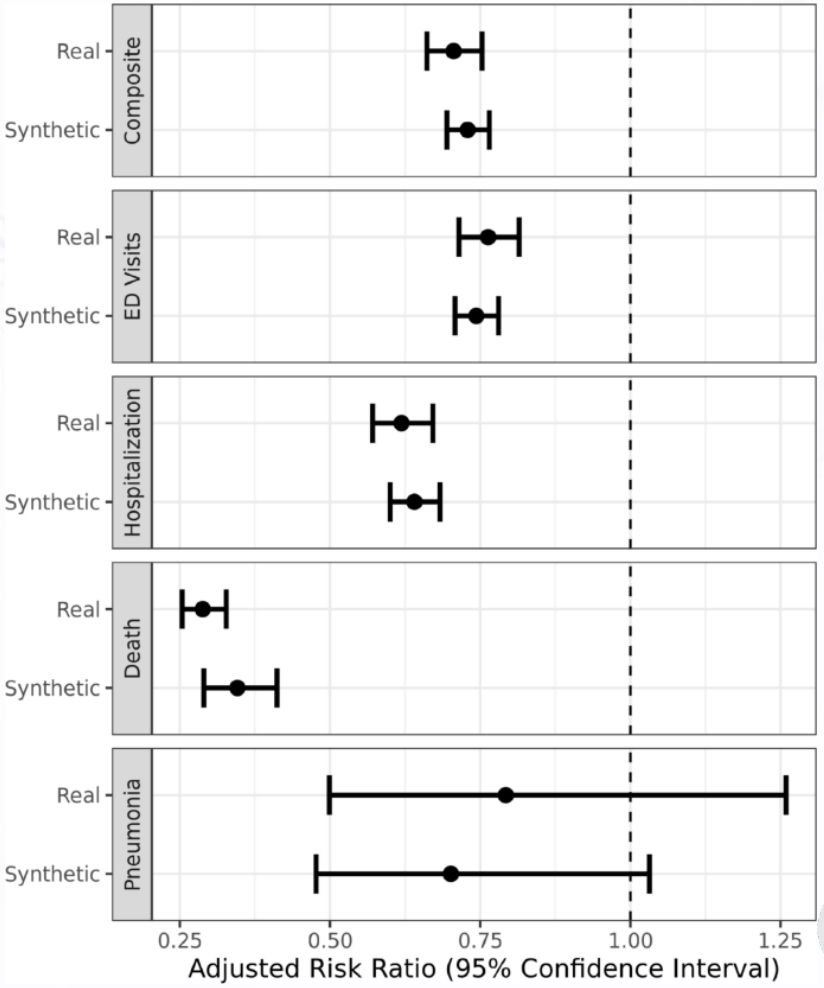
Comparing adjusted hazard ratios from Cox regression models fit on the real and synthetic data:

For the composite endpoint of ED visit, hospitalization, or death:

Real aHR: 0.71 (95% CI 0.66–0.75)
Synthetic aHR: 0.73 95% CI 0.69–0.77

Across all models: average CI overlap 68%

Very similar conclusions can be drawn from models built on the synthetic data compared to the real!



Replica Analytics
AN AETION COMPANY

What about the privacy risks for the synthetic data?

We assessed the risk that an attacker could:

1. Match a synthetic record to a real person in the world present in the training dataset

and

2. Learn something new about the person

In this data this risk was measured to be:

0.001476

Which is very low!

JOURNAL OF MEDICAL INTERNET RESEARCH El Emam et al

Original Paper

Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation

Khaled El Emam[1,2,3], BEng, PhD; Lucy Mosquera[3], BSc, MSc; Jason Bass[3], BSc

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada
[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[3]Replica Analytics Ltd, Ottawa, ON, Canada

Replica Analytics
AN AETION COMPANY

Key Take-Aways

- Longitudinal health data:
  - Is complex and incredibly valuable for research and analytic purposes
  - Can be synthesized using a deep learning architecture
- Synthetic longitudinal health data can be used in analyses to draw similar conclusions as the real data with very low privacy risks meaning synthetic data could be shared to analysts and researchers more readily than real datasets
- Partnerships between health systems, academics, and industry are a powerful way to implement synthetic data solutions

Replica Analytics
AN AETION COMPANY

Replica Analytics

AN AETION COMPANY

Thank You!

Lucy Mosquera: lmosquera@replica-analytics.com