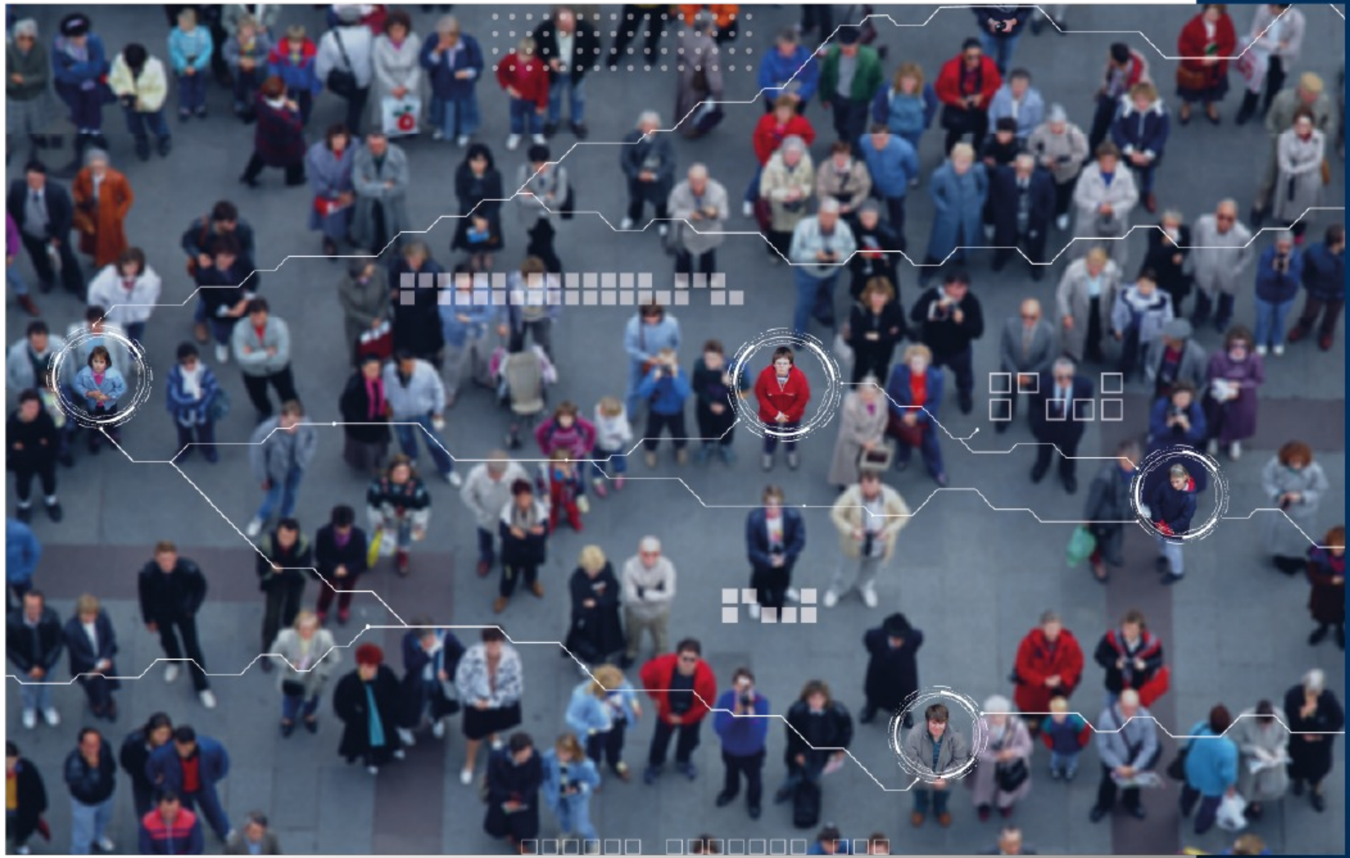


Generating Synthetic Longitudinal Data

A Presentation by Lucy Mosquera of Replica Analytics

Introductions: Amanda Borens

Amanda Borens, MSc
Executive Director of Data Science

- Welcome to the Second 2022 RDCA-DAP webinar
- Place all questions in the Q&A chat box, there will be a Q&A session at the end of the presentation
- Within the Q&A box, please be sure the questions are being sent to "All Panelists" to ensure that they will be seen.
- This presentation is being recorded and will be made available shortly after the presentation

Critical Path Institute is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) and is 54.2% funded by the FDA/HHS, totaling $13,239,950, and 45.8% funded by non-government source(s), totaling $11,196,634. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government.

Presenter: Lucy Mosquera

Replica Analytics

AN AETION COMPANY

RDCA-DAP®

Lucy Mosquera,
Director of Data Science

Lucy Mosquera has a background in biology and mathematics, having done her studies at Queen's University in Kingston and the University of British Columbia. In the past she has provided data management support to clinical trials and observational studies at Kingston General Hospital. She also worked on clinical trial data sharing methods based on homomorphic encryption and secret sharing protocols with various companies.

At Replica Analytics, Lucy is responsible for integrating her subject area expertise in health data into innovative methods for synthetic data generation and the assessment of that data, as well as overseeing our analytics program.

Generating Synthetic Longitudinal Data

Lucy Mosquera
Khaled El Emam

May 18th, 2022

lmosquera@replica-analytics.com
kelemam@replica-analytics.com

Agenda

Introduction to Synthetic Data

1 General description of what synthetic data is

Synthesis of Longitudinal Data

2 An overview of strategies to synthesize longitudinal health data

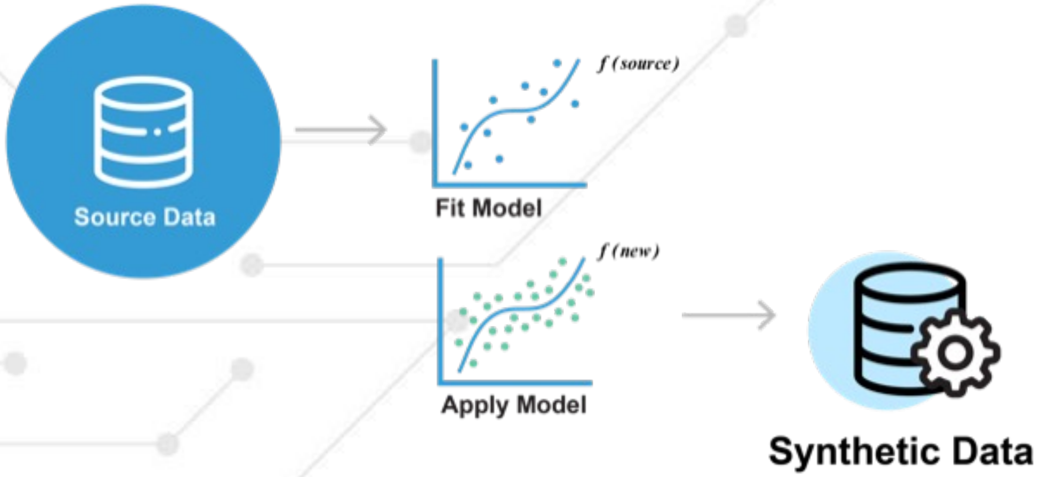
Applications in Rare Disease Data

3 Description of synthetic data use cases in rare diseases

Replica Analytics

AN AETION COMPANY

The Synthesis Process



COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

<u>Additional Clarifications</u>

- The source datasets can be as small as 100 or 150 patients. We have developed generative modeling techniques that will work for small datasets.
- The source datasets can be very large – then it becomes a function of compute capacity that is available.
- It is not necessary to know how the synthetic data will be analyzed to build the generative models. The generative models capture many of the patterns in the source data.

Replica Analytics
AN AETION COMPANY

Use Cases for Synthetic Data

Discover Artificial Intelligence

Discover

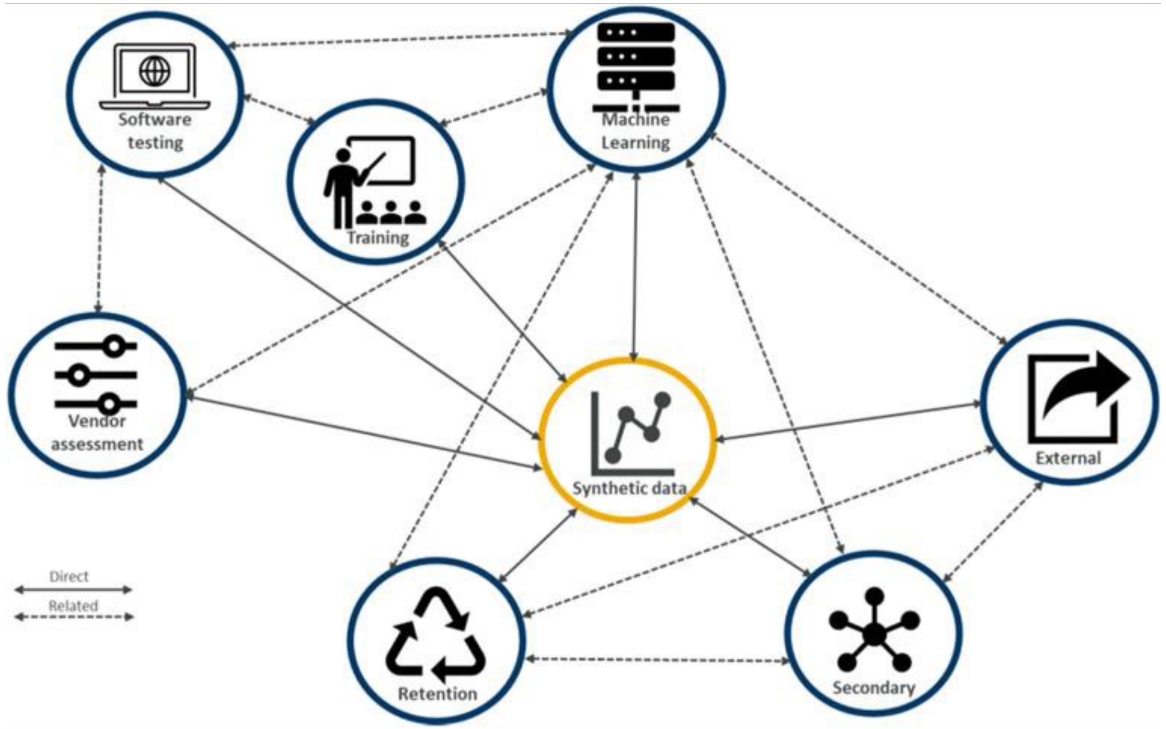
Review

Synthetic data use: exploring use cases to optimise data utility

Stefanie James[1] · Chris Harbron[2] · Janice Branson[3] · Mimmi Sundler[4]

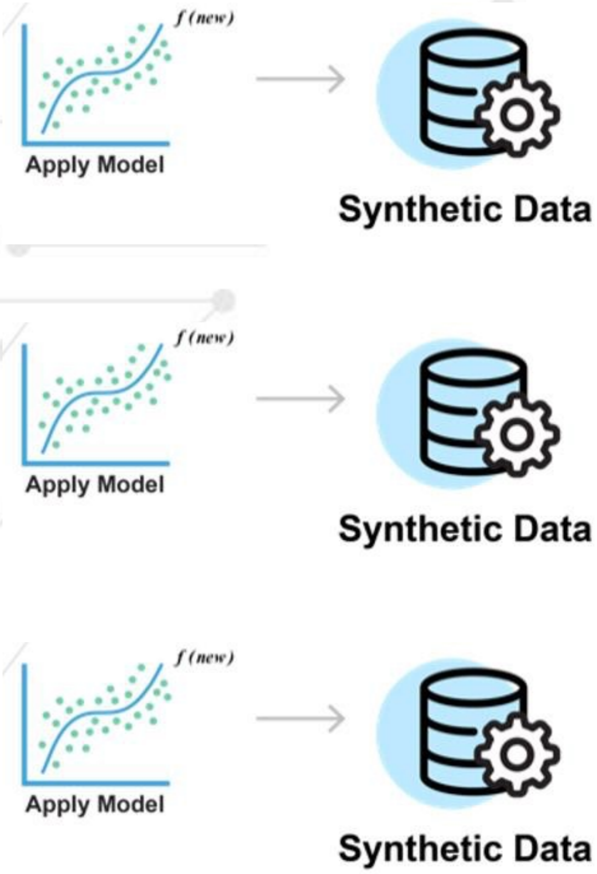
Received: 12 November 2021 / Accepted: 7 December 2021
Published online: 13 December 2021

Can be grouped as:
Privacy use cases
Analytic use cases

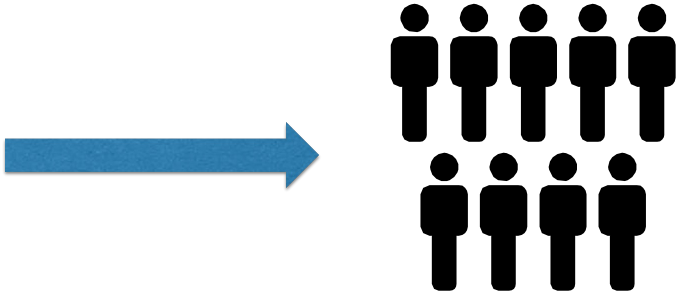


Replica Analytics
AN AETION COMPANY

A simulator exchange allows data to be made available without sharing actual data



Data Consumers

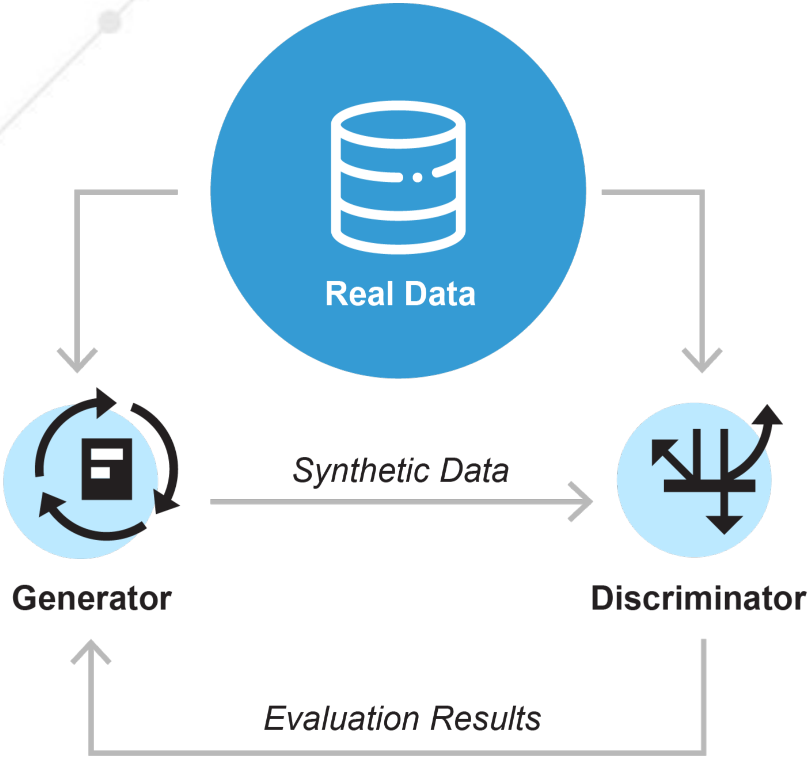


Additional Clarifications

- The simulators would not be given to the data consumers – they would only have access to them through an interface.
- This access would be monitored and throttled to reduce the risk of attacks on the models.
- Data consumers would also need to agree to terms of use around the access to the simulators.

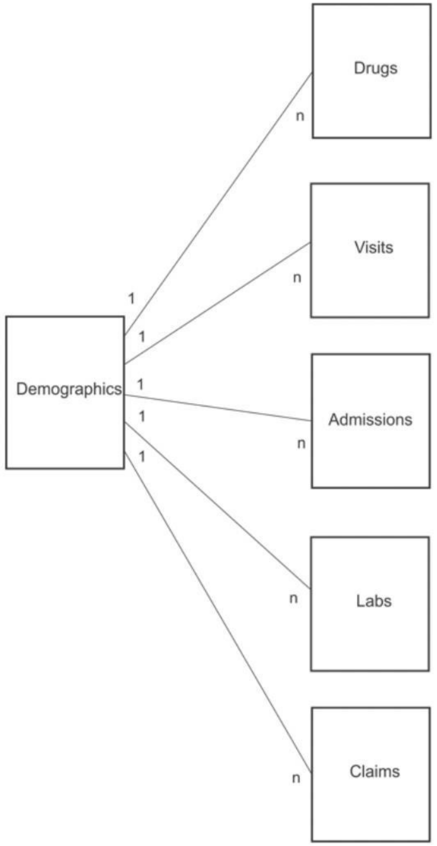
Replica Analytics
AN AETION COMPANY

Training a generative model often uses a discriminator



Replica Analytics
AN AETION COMPANY

Longitudinal Data Model



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

Replica Analytics
AN AETION COMPANY

The complexity of longitudinal data requires a different synthesis approach

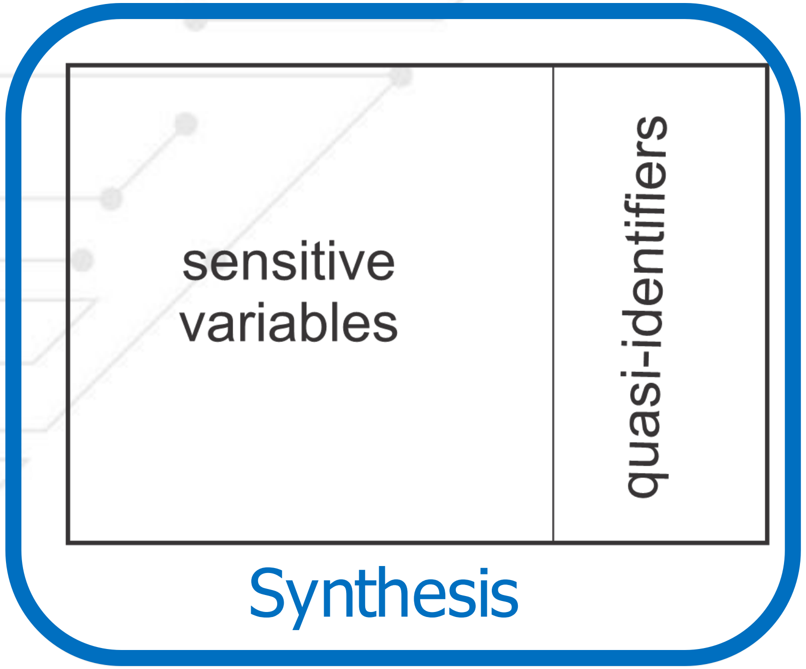
- Features & Cohorts:
  - Define features on the raw longitudinal data and then synthesize the tabular feature dataset
  - Define a cohort on the raw longitudinal data and then synthesize the tabular cohort dataset

- Raw Longitudinal:
  - Fully vs partially synthetic data
  - For RWD we use a hybrid approach of sequential synthesis and recurrent neural network architectures to synthesize these – full synthesis

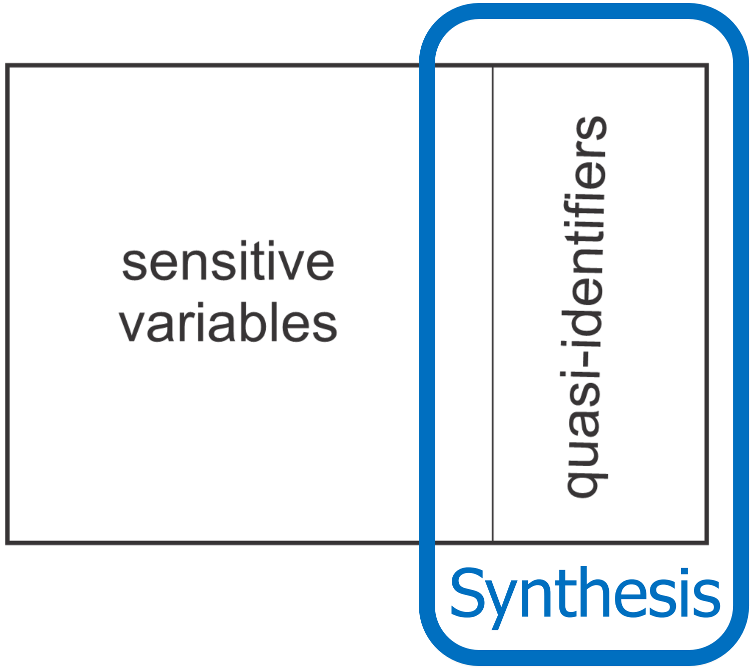
Replica Analytics
AN AETION COMPANY

Two synthesis strategies for raw longitudinal data

Full Synthesis
Synthesize all variables

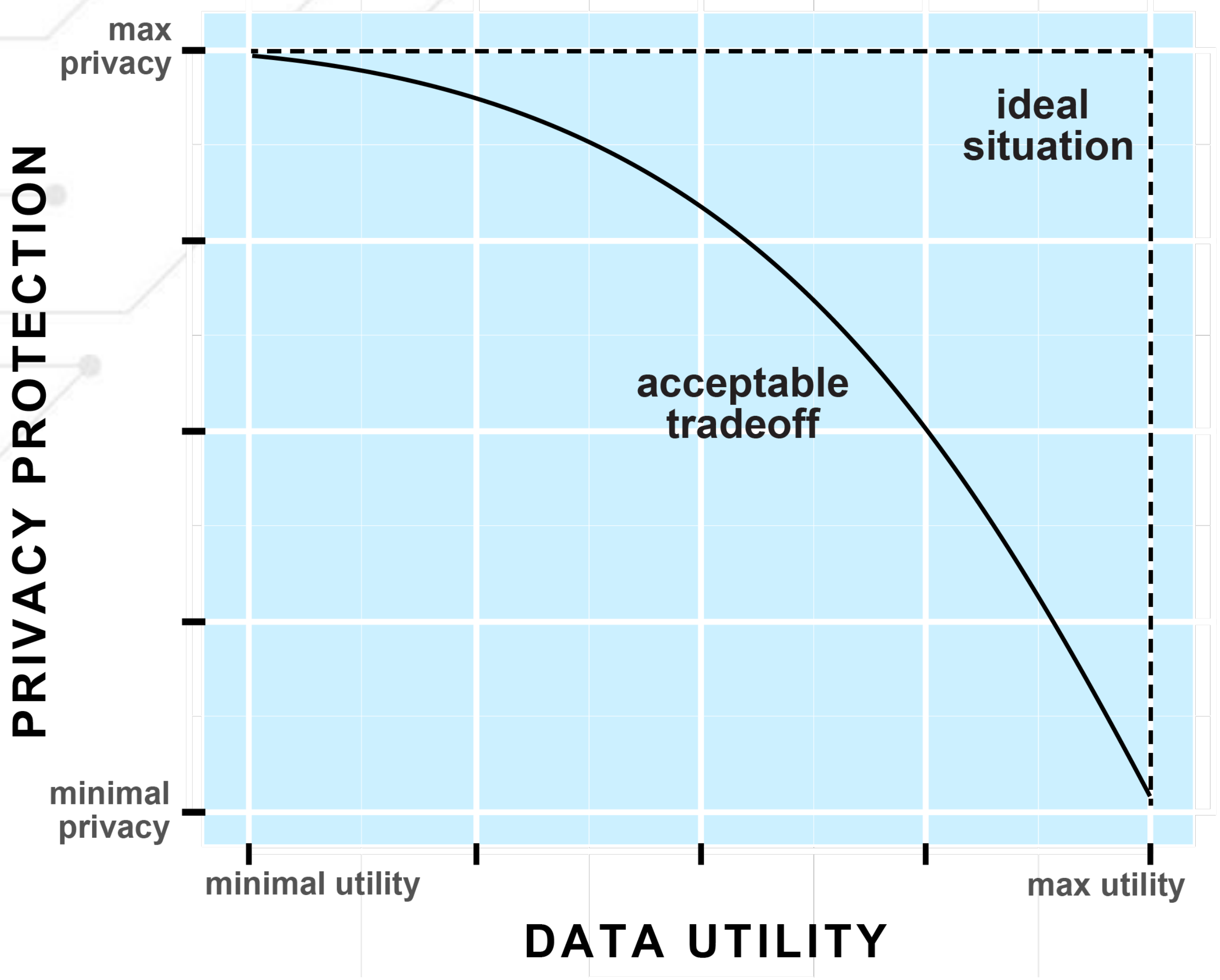


Partial Synthesis
Synthesize quasi-identifiers



Replica Analytics
AN AETION COMPANY

Privacy-Utility Trade-off



Replica Analytics
AN AETION COMPANY

One way to classify utility metrics is as broad and narrow

Broad Metrics → Narrow Metrics

These are generic metrics that are easy to calculate when the generative model is built and synthetic data are synthesized. They are only useful if they are predictive of workload-specific metrics.

These are workload-specific and are what is of most interest to the data users. However, all the possible workloads will not be known in advance and therefore we have to consider representative workloads when developing and evaluating utility metrics.

Replica Analytics
AN AETION COMPANY

Examples of Broad Metrics

- Comparison of the number of events per patient
  - Number of certain types of events (e.g., prescriptions) per patient
  - Limit the above to a certain time interval
- Comparison of the overall frequency of events
- Comparisons of event distributions across classes of events using univariate distribution comparison metrics
- Evaluation of the k-order transition matrices among events or classes of events

Attribution Disclosure: Find a similar record in the synthetic data and learn something new

Quasi-identifiers

Sensitive variables

Sex	Year of Birth	NDC
Male	1985	009-0031
Male	1988	0023-3670
Male	1982	0074-5182
Female	1983	0078-0379
Female	**1989**	**65862-403**
Male	1981	55714-4446
Male	1982	55714-4402
Female	1987	55566-2110
Male	1981	55289-324
Female	1986	54868-6348
Male	1980	53808-0540

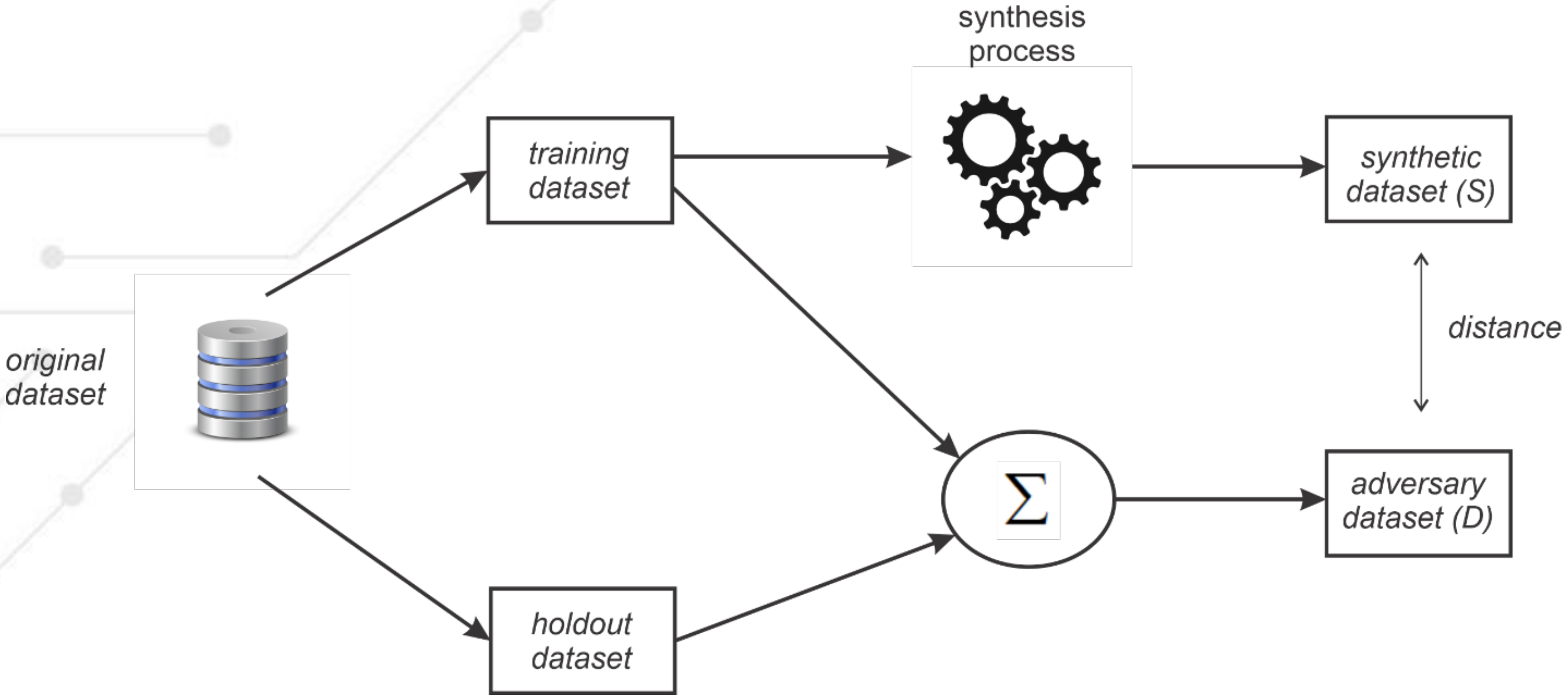
Attribution Risk Results

Published risk assessment results for synthetic data generated using sequential tree synthesis method:

	Synthetic Data Risk	
	Population-to-sample risk	Sample-to-population risk
Washington State Inpatient Database	0.00056	0.0197
Canadian COVID-19 cases	0.0043	0.0086

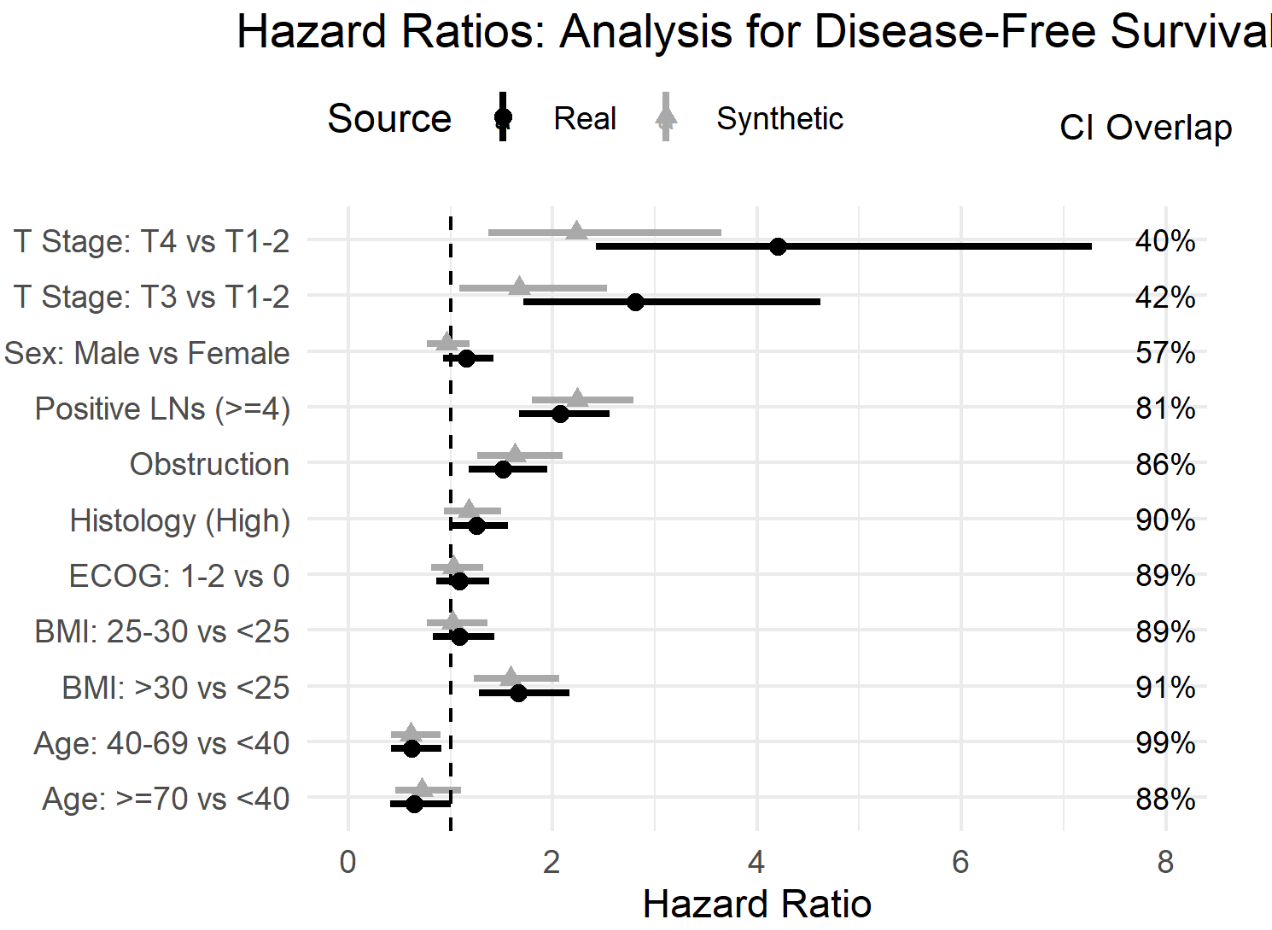
El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. J Med Internet Res 2020;22(11):e23139, doi: 10.2196/23139.

Membership disclosure: is the distance between S and D predictive of which records are in the training dataset



Replica Analytics
AN AETION COMPANY

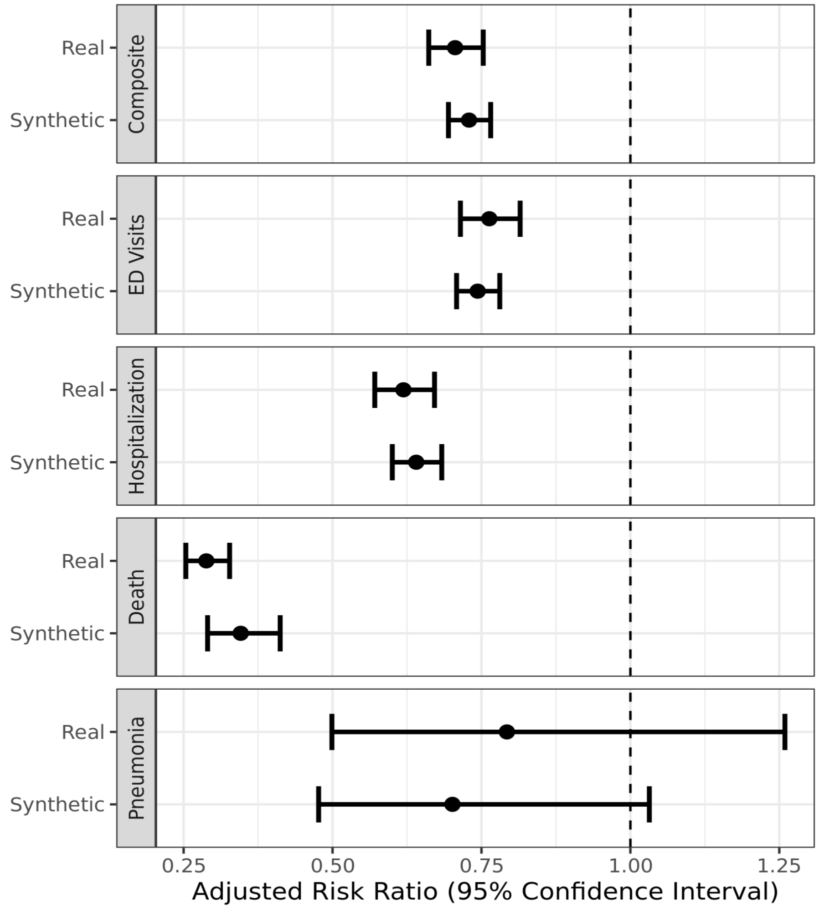
Analysis Specific Utility: Adjusted model of impact of bowel obstruction on DFS



Azizi Z, Zheng C, Mosquera L GOING-FWD Collaborators, *et al*. Can synthetic data be a proxy for real clinical trial data? A validation study *BMJ Open* 2021;**11:**e043497. doi: 10.1136/bmjopen-2020-043497

Replica Analytics
AN AETION COMPANY

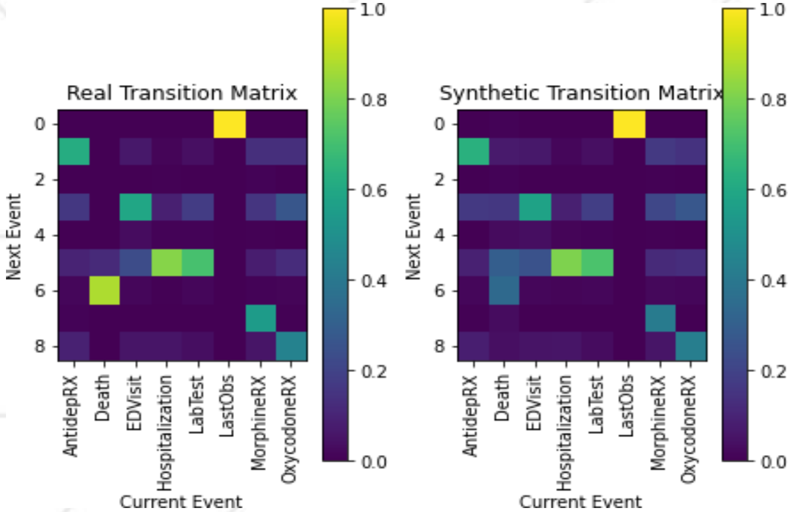
Analysis Specific Utility Results: Adjusted Cox regression

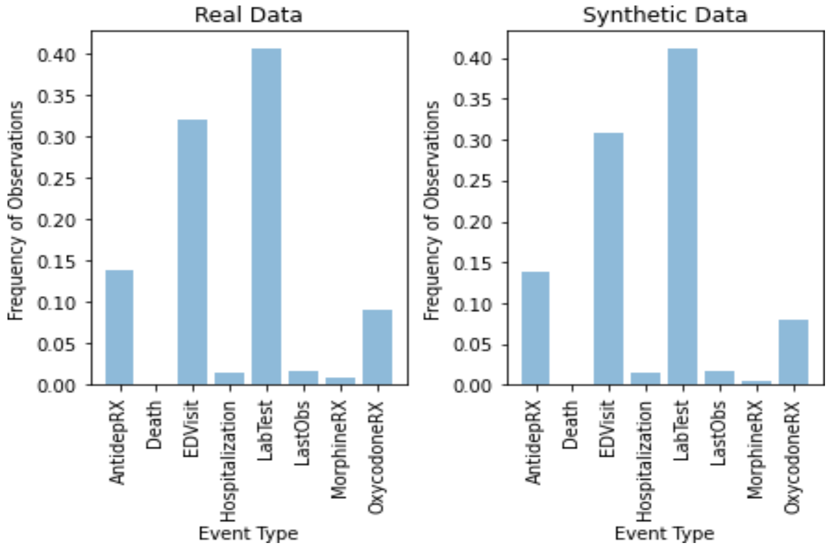


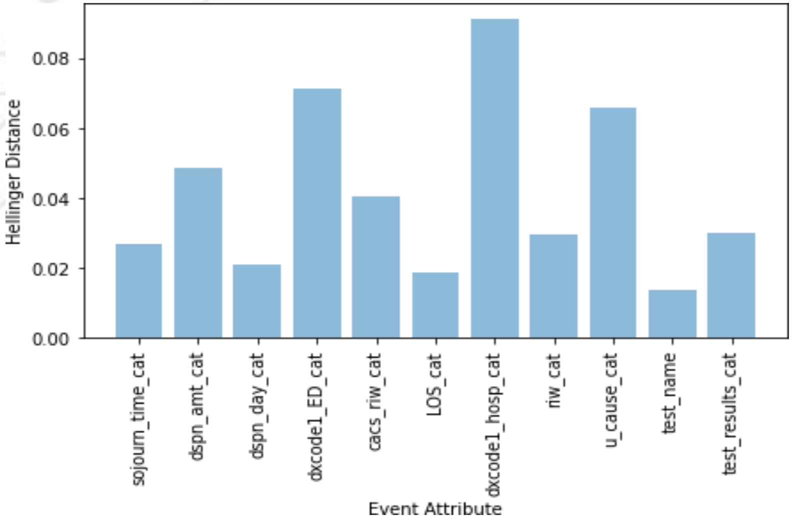
Note: Adjusted estimates include the following co-variates: age, sex, antidepressant use, Elixhauser score, ALT, eGFR, HCT; Opioid 1 served as the reference group

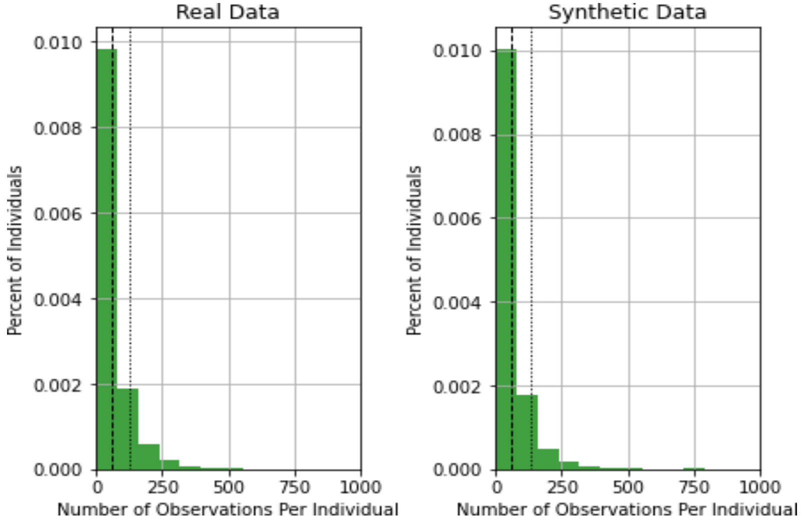
Replica Analytics
AN AETION COMPANY

Broad Utility Results









Replica Analytics
AN AETION COMPANY

Rare Diseases

Data synthesis can be used to meet two needs for rare disease data:

- Mitigating privacy risks to facilitate data sharing in difficult to anonymize small datasets; supporting open data initiatives
- Amplifying and augmenting existing small datasets

Replica Analytics

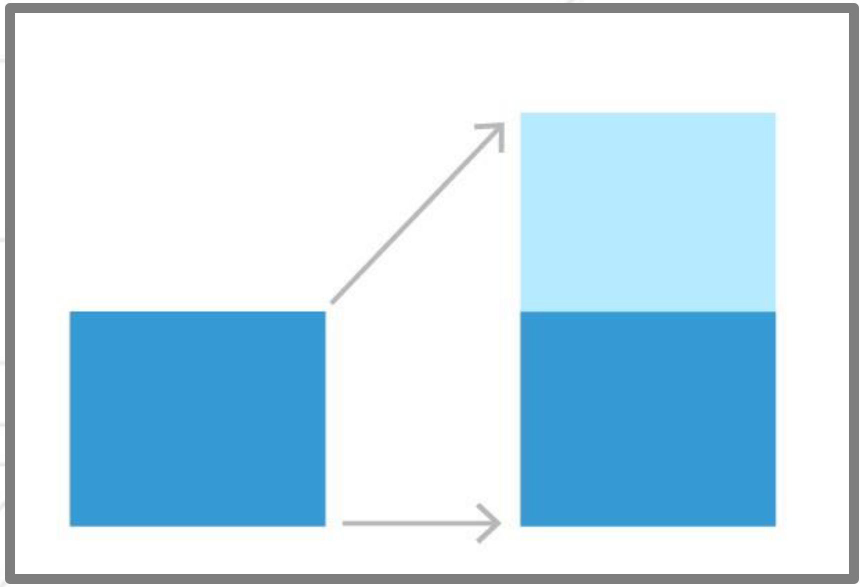
AN AETION COMPANY

Privacy Risks in Rare Disease Data

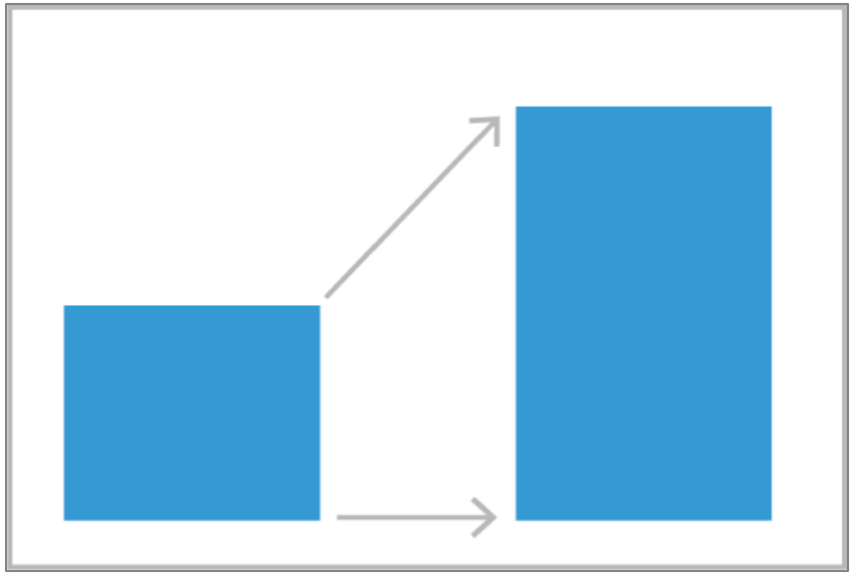
- Rare disease datasets can be difficult to anonymize due to small sample sizes, heterogeneity among patients
- Synthesis mitigates privacy risks without compromising data utility to the same degree

Replica Analytics
AN AETION COMPANY

Data Augmentation vs Data Amplification: Two different approaches for getting more data



(a) Augmentation

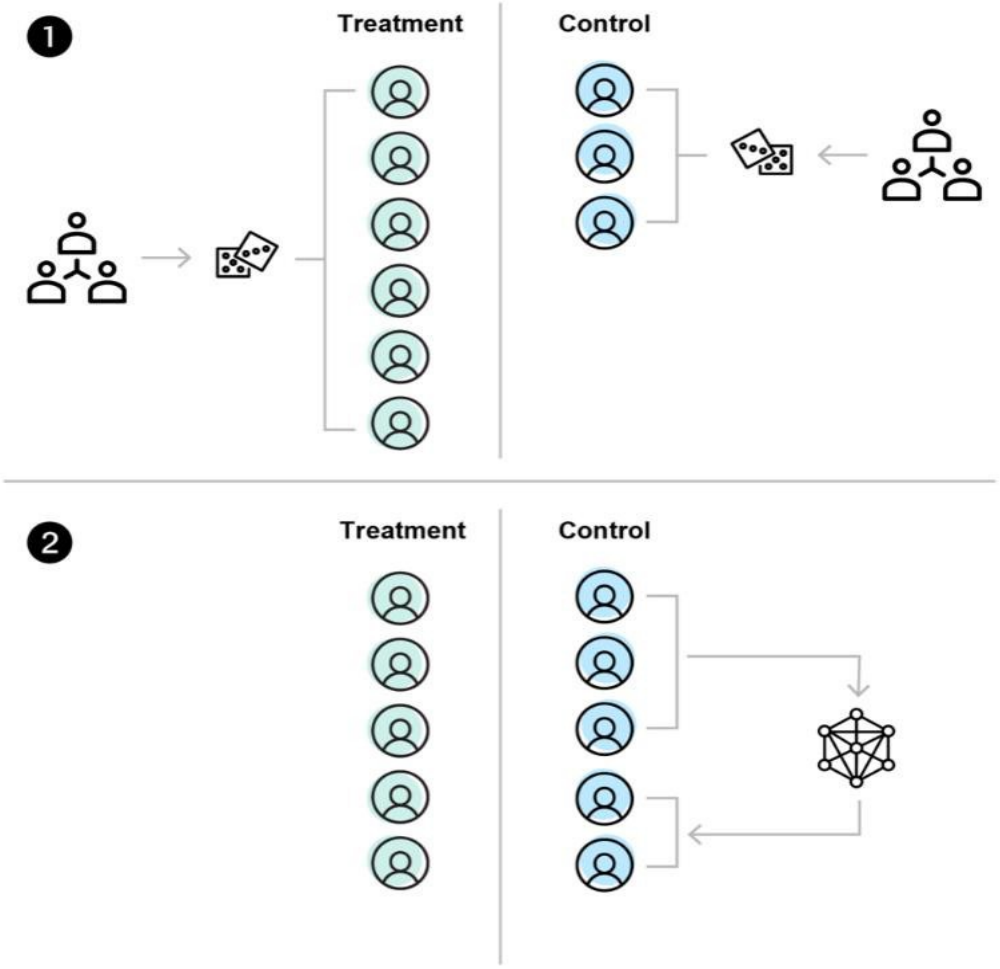


(b) Amplification

Replica Analytics
AN AETION COMPANY

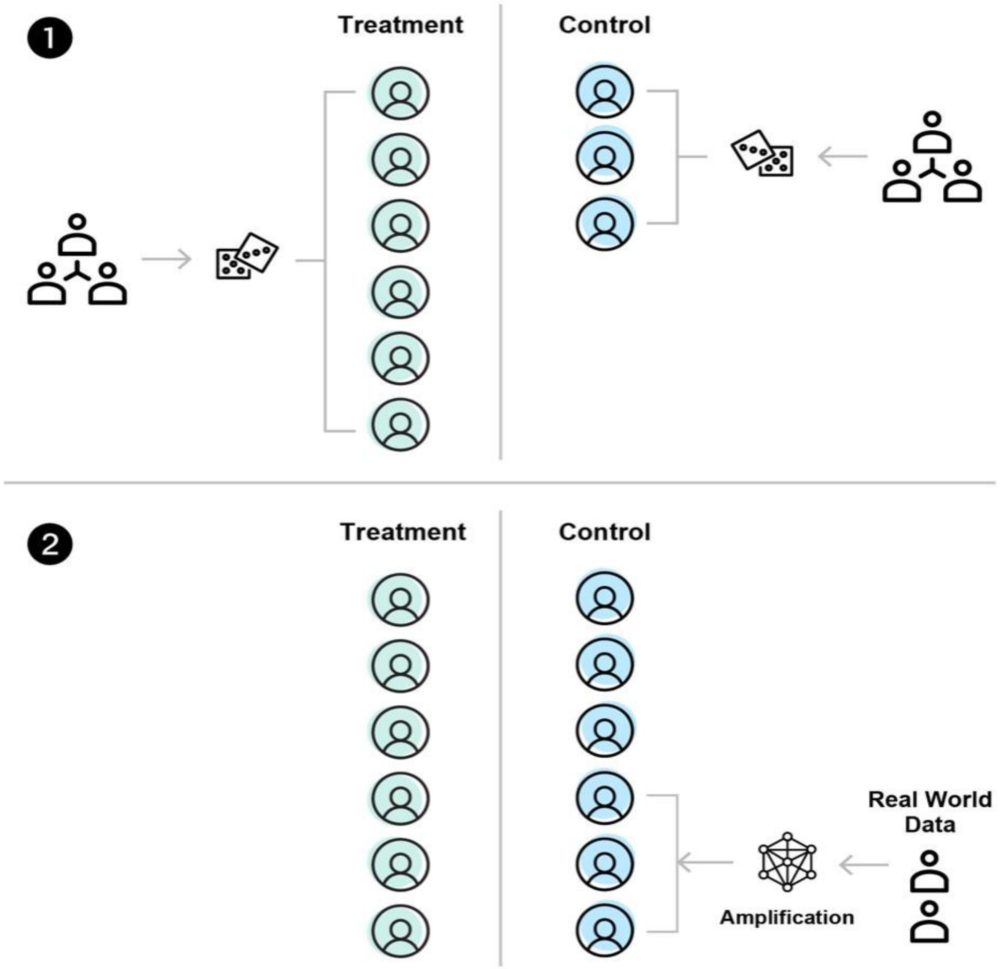
Virtual Patients

Virtual patients can be simulated to reduce recruitment or to rescue studies with low recruitment or high attrition



Virtual Patients

Real-world data can be amplified to create synthetic external controls, especially when there are insufficient RWD or RWD diversity



Replica Analytics
AN AETION COMPANY

Valid Inferences on Synthetic Data

Analyses conducted on synthetic data can produce valid statistical inferences by using multiple imputation framework

(a) **Single analysis of synthetic data**

(b) **Multiple imputation analysis of synthetic data**

Replica Analytics
AN AETION COMPANY

Question and Answer panel

Amanda Borens, MSc
Executive Director of
Data Science

Lucy Mosquera
Director of Data Science

Jeff Barrett, PhD,
FCP, Senior Vice President;
RDCA-DAP Lead

Khaled El Emam, PhD
Co-Founder and GM

RDCA-DAP®

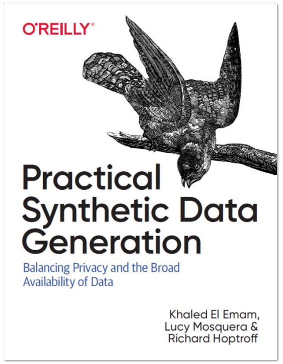
Replica Analytics
AN AETION COMPANY

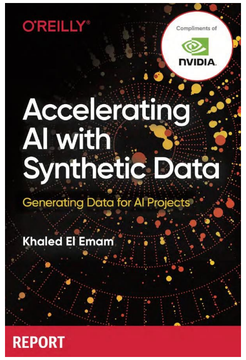
To Learn More

- Join our mailing list: https://bit.ly/3gRVAIi
- Follow us on Linkedin: https://bit.ly/2XS3KHF
- Listen to our comprehensive on-line tutorials on data synthesis:

  https://bit.ly/2TXI0Jy

- Read our introductory report and book on the topic



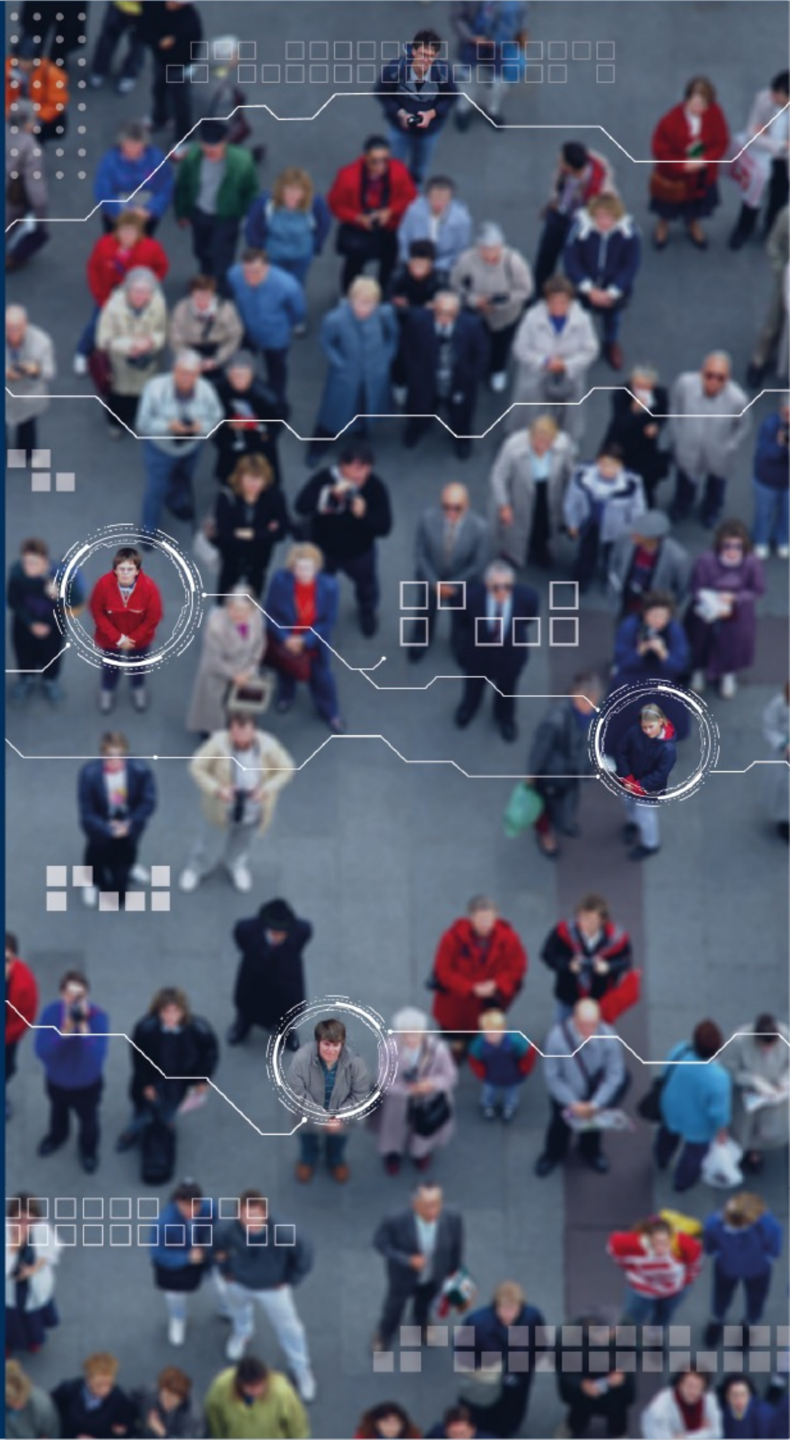


Replica Analytics
AN AETION COMPANY

RDCA-**DAP**®

Rare Disease Cures Accelerator
Data and Analytics Platform

Thank You!



RDCA-DAP®
Rare Disease Cures Accelerator
Data and Analytics Platform

Contact Us

Scan the QR code to access RDCA-DAP

Web

c-path.org/rdca-dap

Email

rdcadap@c-path.org

Data Contributions

Alex Bétourné, Scientific Director,
RDCA-DAP | abetourne@c-path.org