Replica Analytics

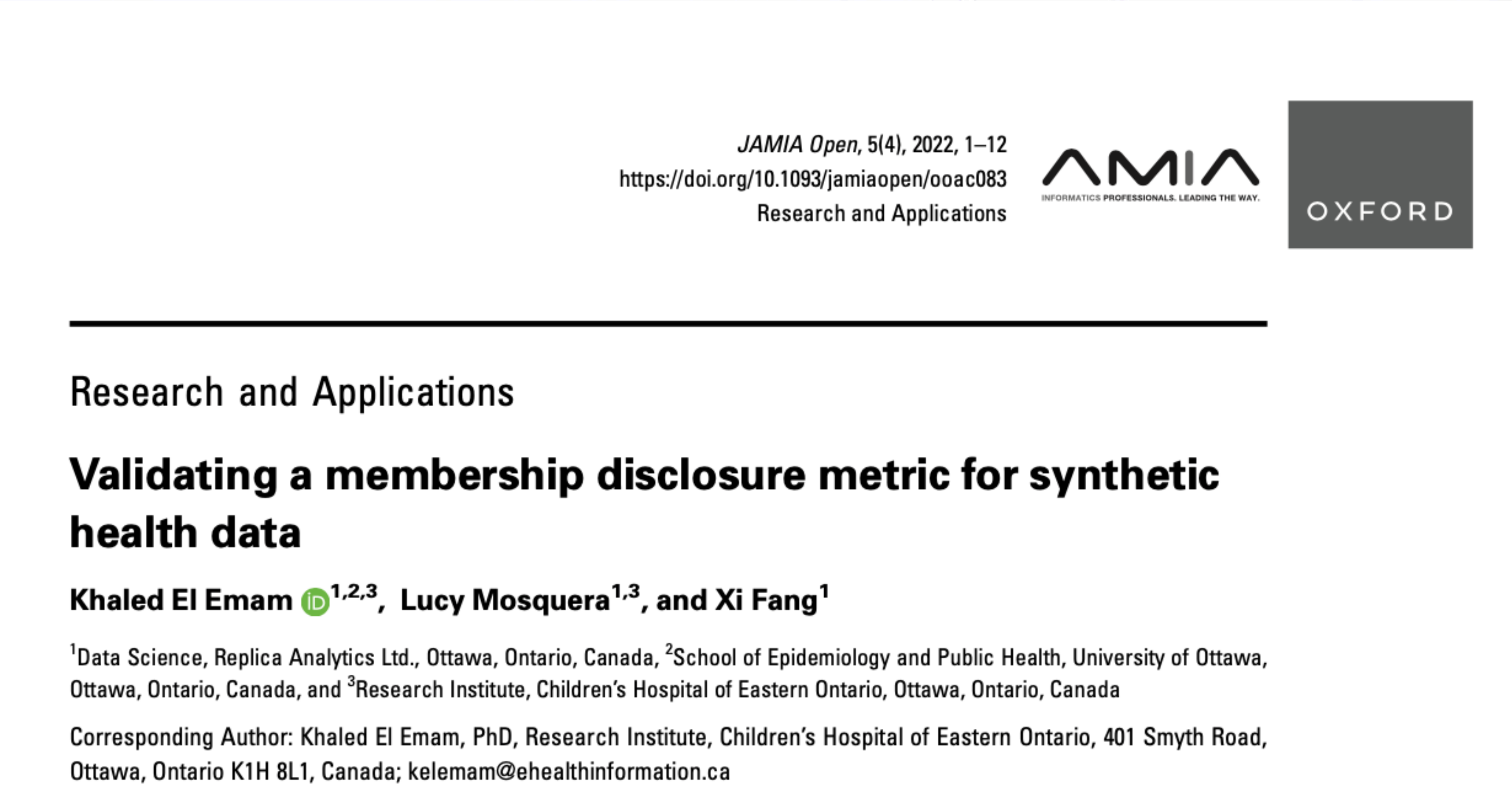
AN AETION COMPANY

Evaluating Privacy Risks in Synthetic Data Using Membership Disclosure

Lucy Mosquera & Xi Fang

Agenda

- Introduction to synthetic data and its privacy risks
- Partitioning method for estimating membership disclosure risk
- Our work assessing how to parameterize the partitioning methods
- Application in clinical trial datasets and optimization of synthesis algorithms



JAMIA Open, 5(4), 2022, 1–12
https://doi.org/10.1093/jamiaopen/ooac083
Research and Applications

Research and Applications

Validating a membership disclosure metric for synthetic health data

Khaled El Emam[1,2,3], **Lucy Mosquera**[1,3], **and Xi Fang**[1]

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

Corresponding Author: Khaled El Emam, PhD, Research Institute, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, Ontario K1H 8L1, Canada; kelemam@ehealthinformation.ca

Introduction to synthetic data and its privacy risks

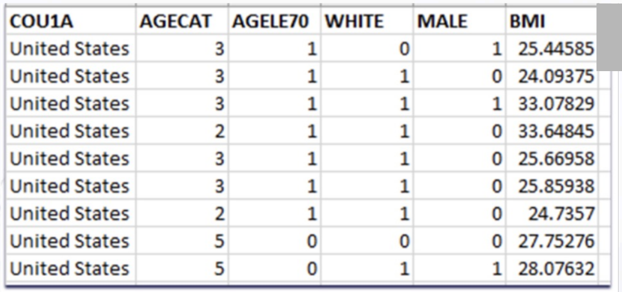
Replica Analytics

AN AETION COMPANY

Synthetic data

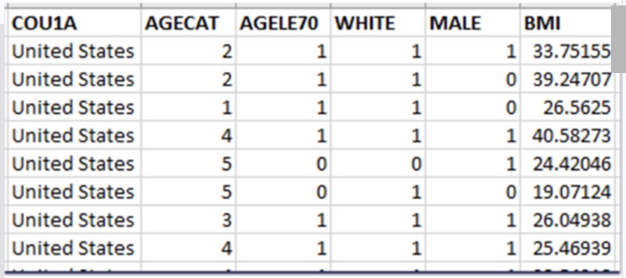
WHAT IT IS

Synthetic data is **generated from real data**, but is not real data.



COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	3	1	0	1	25.44585
United States	3	1	1	0	24.09375
United States	3	1	1	1	33.07829
United States	2	1	1	0	33.64845
United States	3	1	1	0	25.66958
United States	3	1	1	0	25.85938
United States	2	1	1	0	24.7357
United States	5	0	0	0	27.75276
United States	5	0	1	1	28.07632

Real



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

Synthetic

WHY IT MATTERS

It has the **same patterns and statistical properties** as real data.

HOW IT CAN BE USED

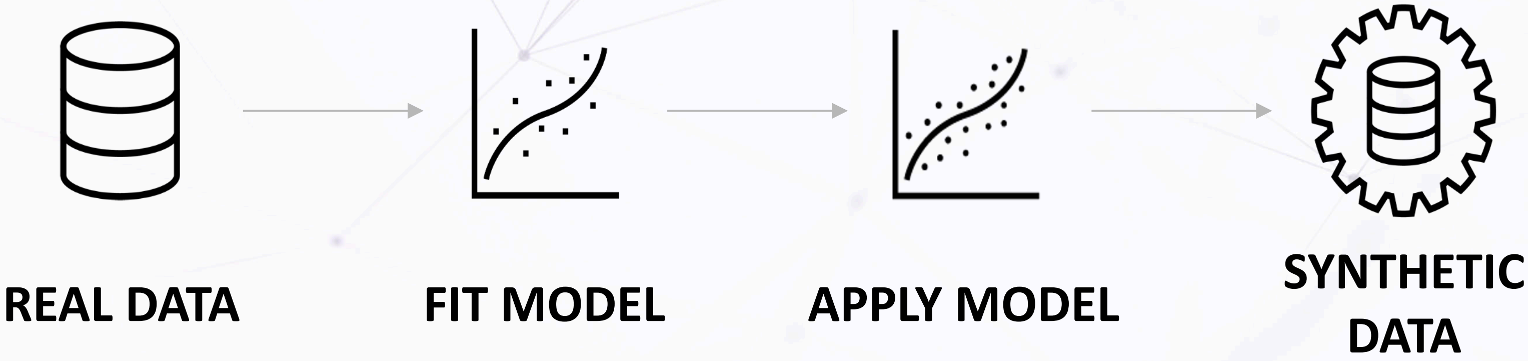
For certain use cases it **can act as a proxy for real data**.

Replica Analytics

AN AETION COMPANY

Synthetic data generation

Machine learning or deep learning models **capture patterns** in the real data, and then **generate new data** from that model.



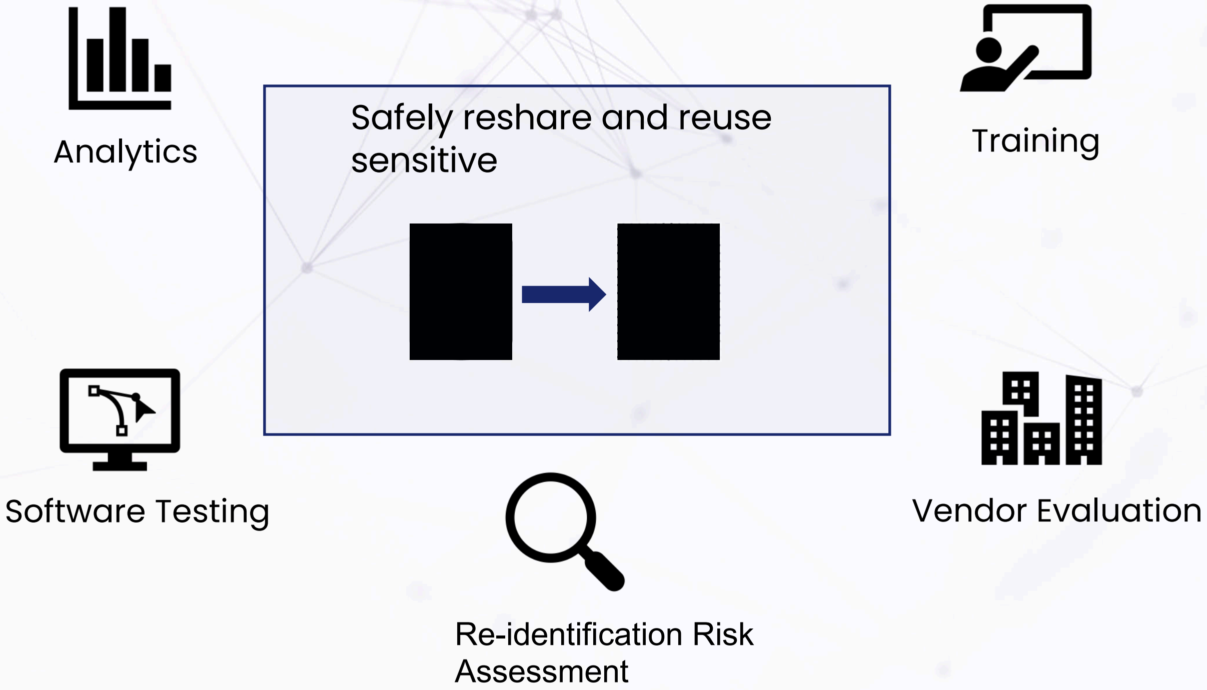
REAL DATA

FIT MODEL

APPLY MODEL

SYNTHETIC DATA

Privacy use cases



Analytics

Safely reshare and reuse sensitive

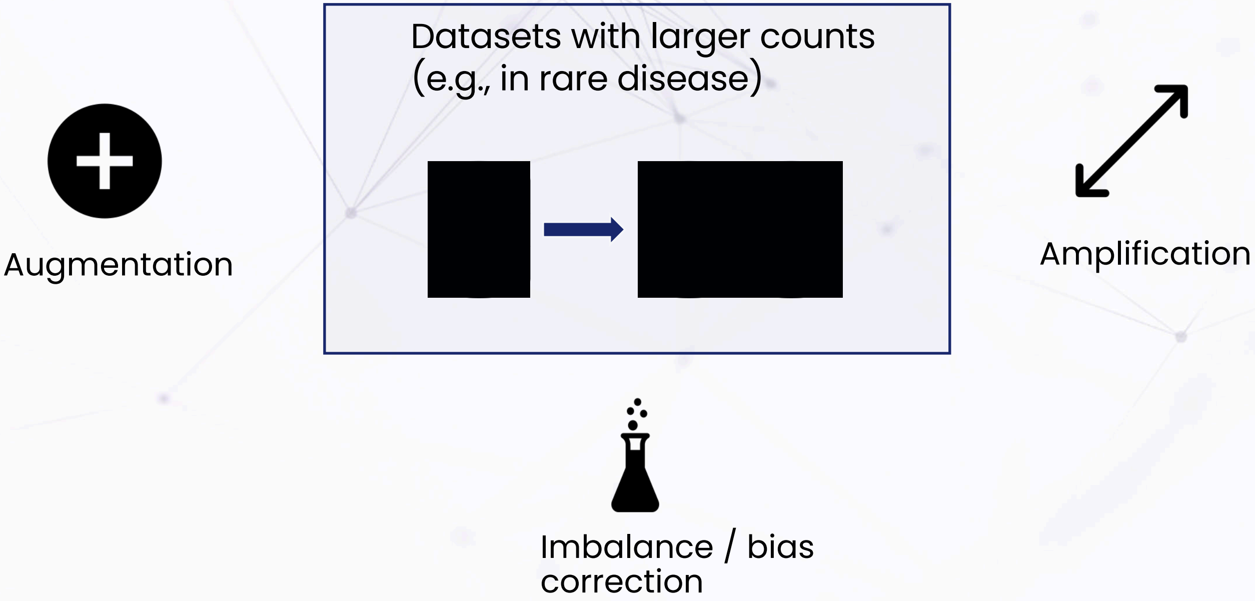
Training

Software Testing

Vendor Evaluation

Re-identification Risk Assessment

Data enhancement use cases



Datasets with larger counts (e.g., in rare disease)

Augmentation

Amplification

Imbalance / bias correction

Replica Analytics

AN AETION COMPANY

Privacy concerns with synthetic data

In general, identity disclosure is not the main type that is of concern

- Unless the generative model has been overfit, in which case many records would just be replicated; but that should not be a common occurrence

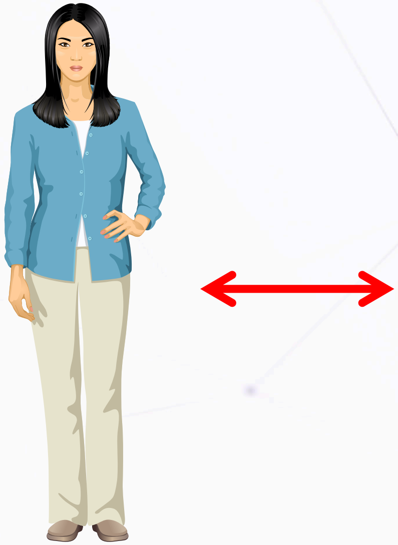
We are concerned with other types of inferences from the dataset:

- Attribution disclosure
- Membership disclosure

Identity disclosure is when a person's identity is assigned to a record

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Attribution disclosure: find a record in the synthetic data similar to a high risk real individual and learn something new about that individual



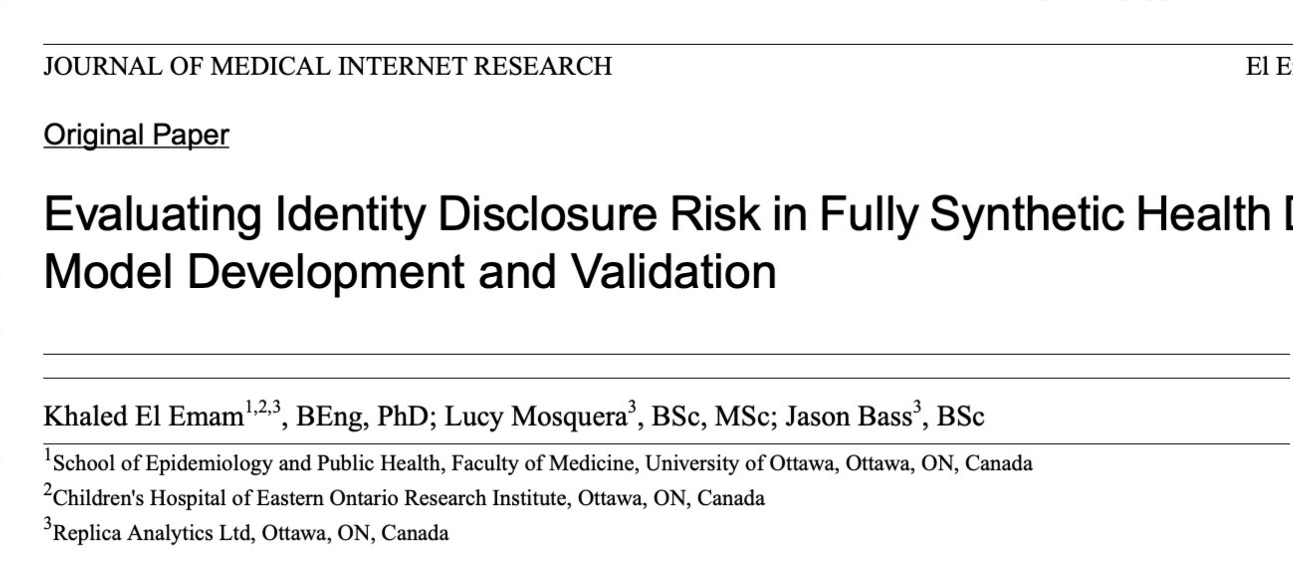
Quasi-identifiers

New Information

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Replica Analytics

AN AETION COMPANY



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

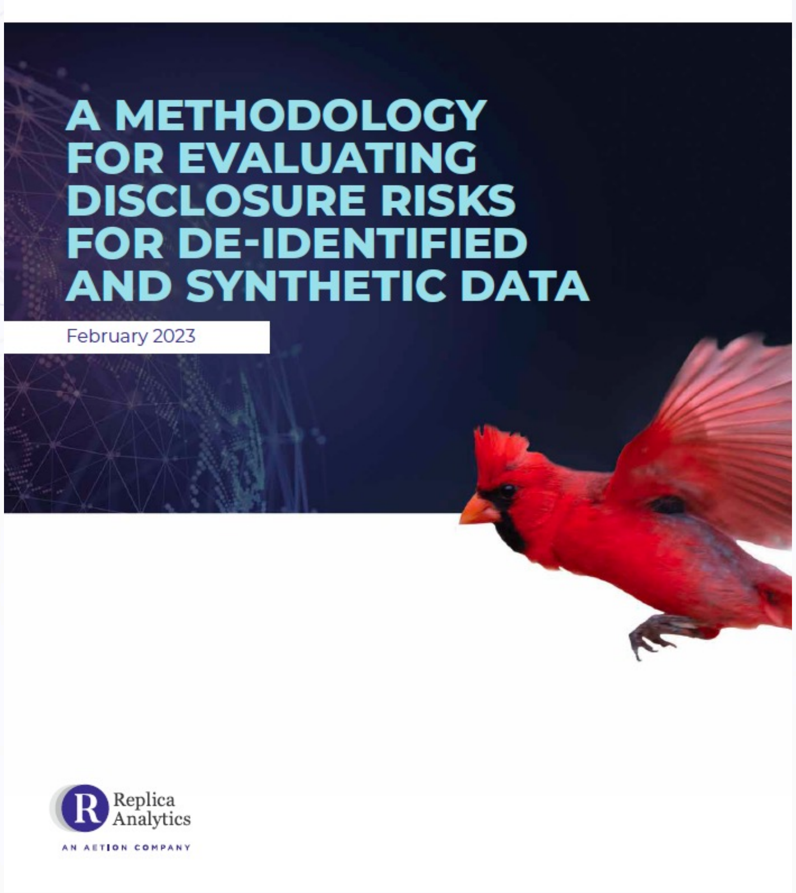
Managing and Regulating Privacy Risks in Synthetic Data

March 30, 2022

Previous work on attribution disclosure in synthetic data

Unified assessment methodology

Upcoming webinar in 2023 will cover our unified comprehensive risk assessment framework for synthetic data



A METHODOLOGY FOR EVALUATING DISCLOSURE RISKS FOR DE-IDENTIFIED AND SYNTHETIC DATA

February 2023

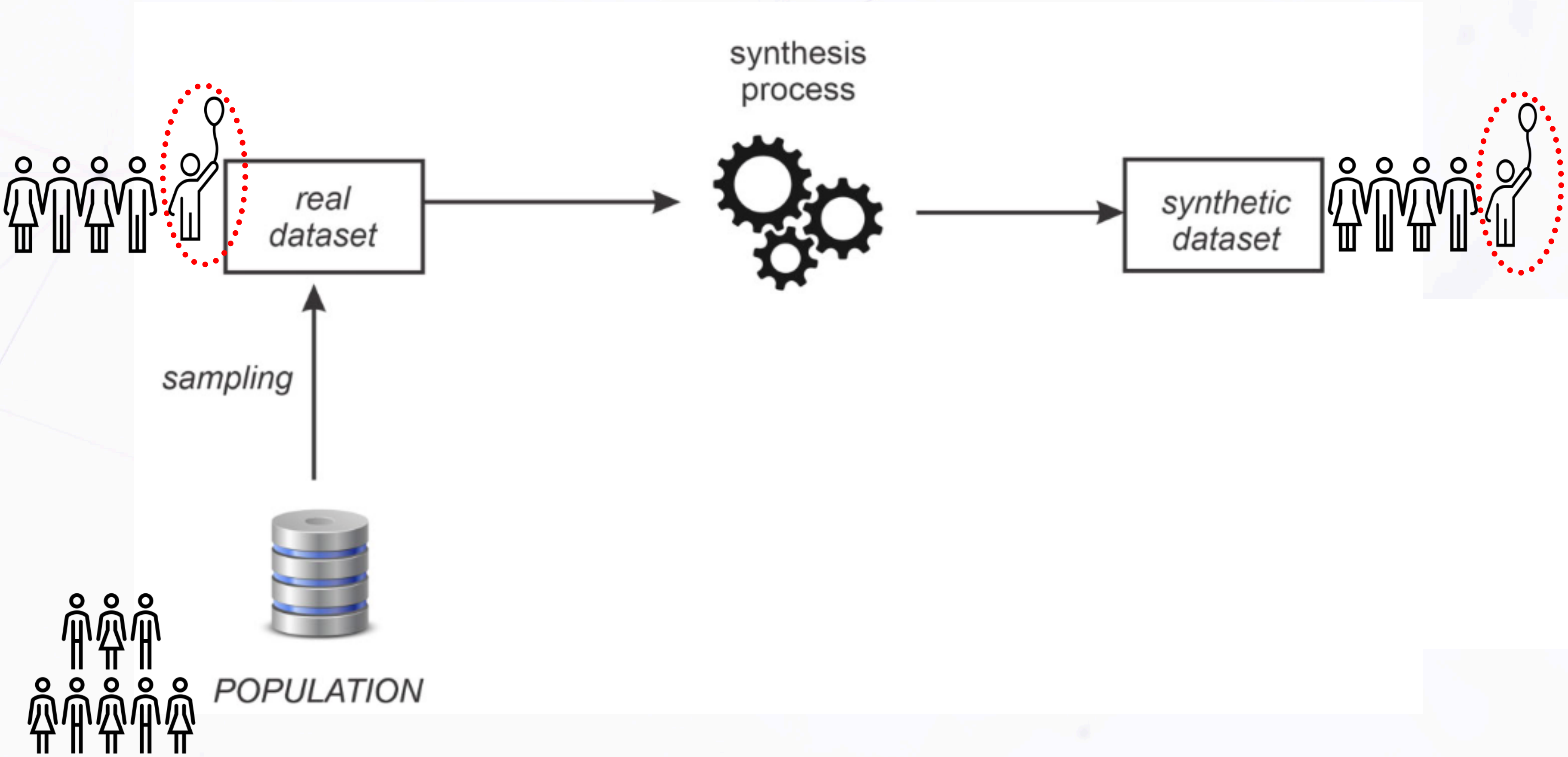
Replica Analytics

AN AETION COMPANY

Membership disclosure

- To what extent an adversary could determine that a target individual is in the training data that was used for training the generative model
- Knowing that someone is in the training dataset may reveal sensitive information about them, for example, if the dataset was about individuals who participated in an HIV study

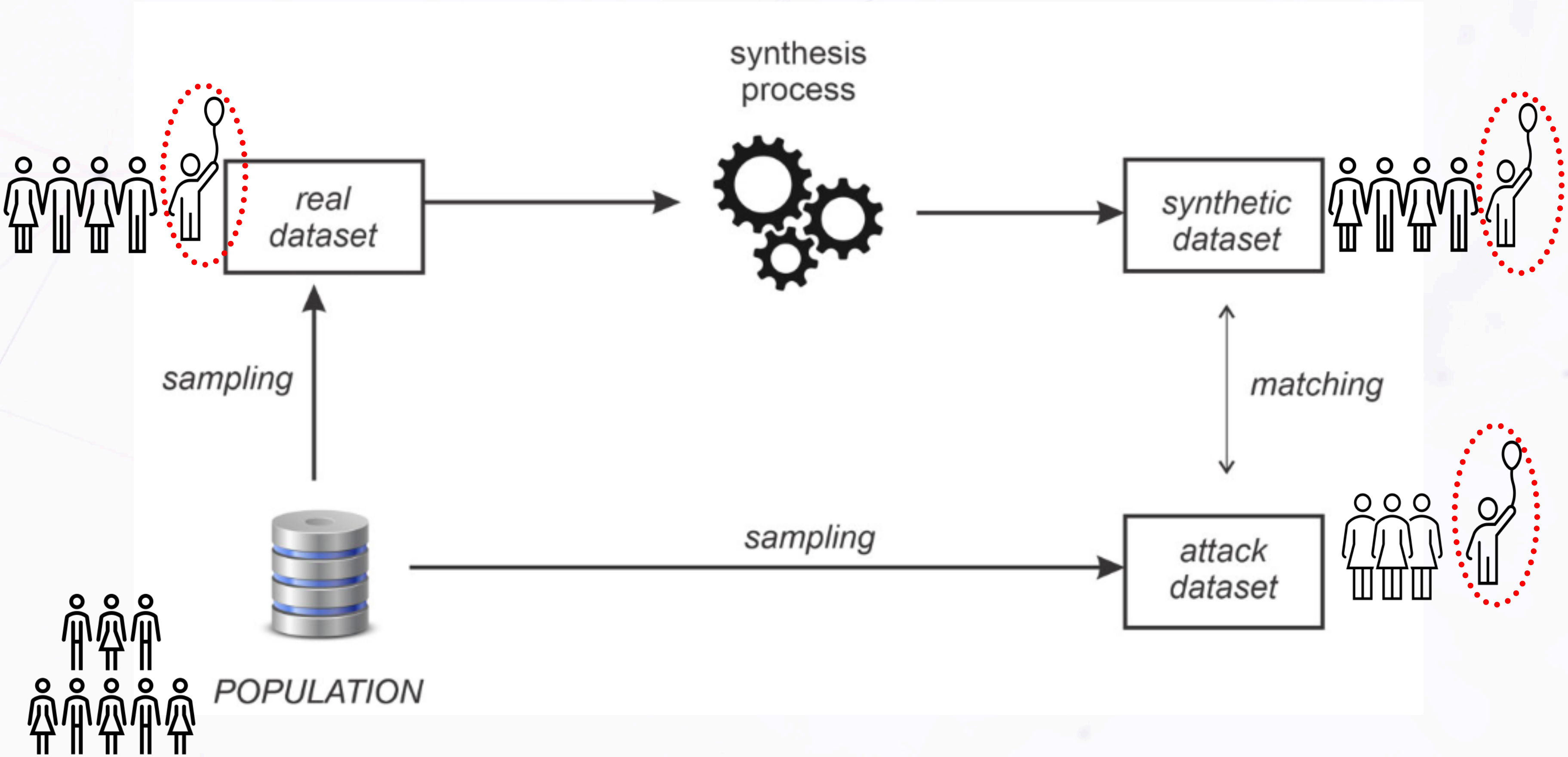
The (ground truth) process for a membership disclosure attack



Replica Analytics

AN AETION COMPANY

The (ground truth) process for a membership disclosure attack

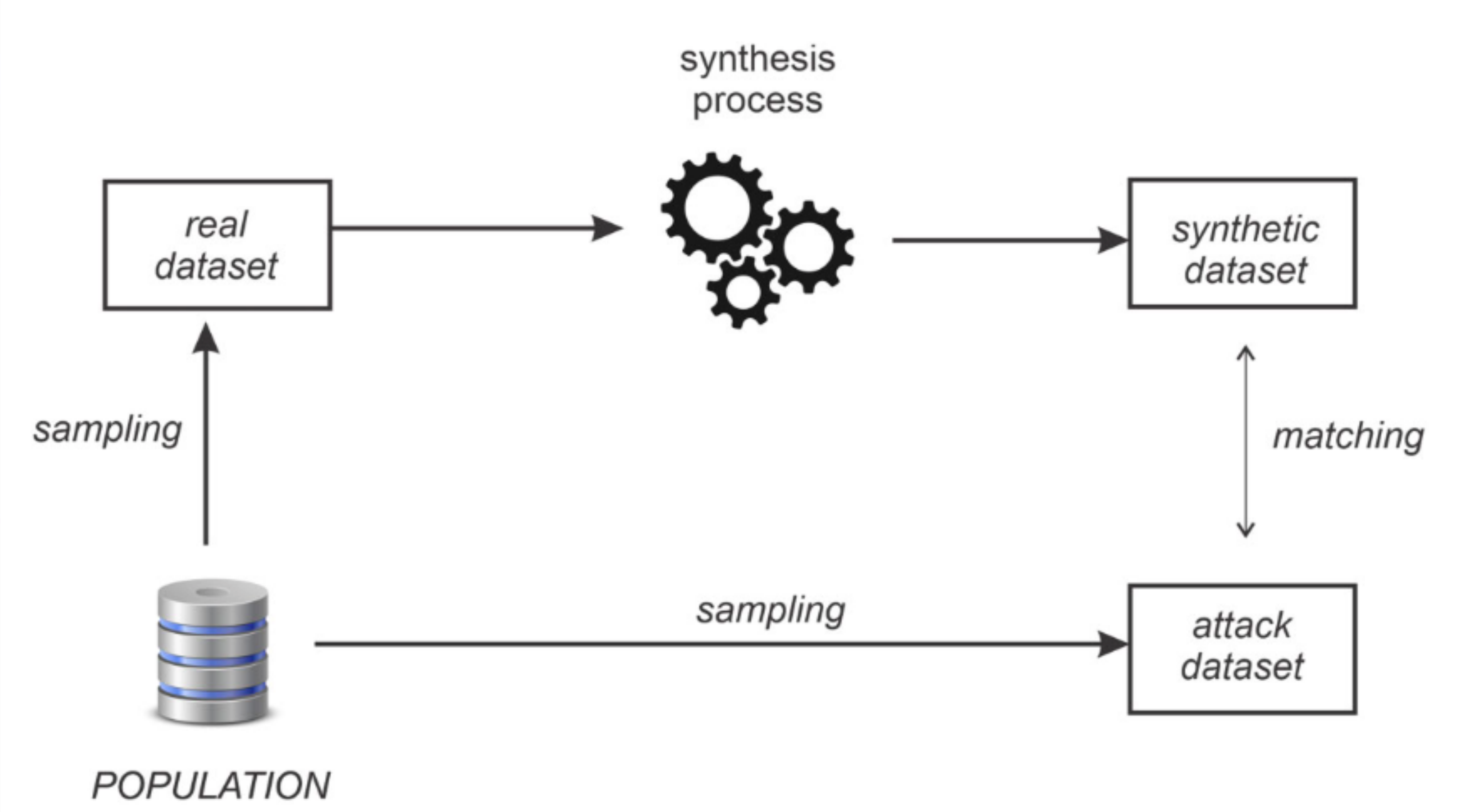


Partitioning method to calculate membership disclosure risks

Replica Analytics

AN AETION COMPANY

The (ground truth) process for a membership disclosure attack

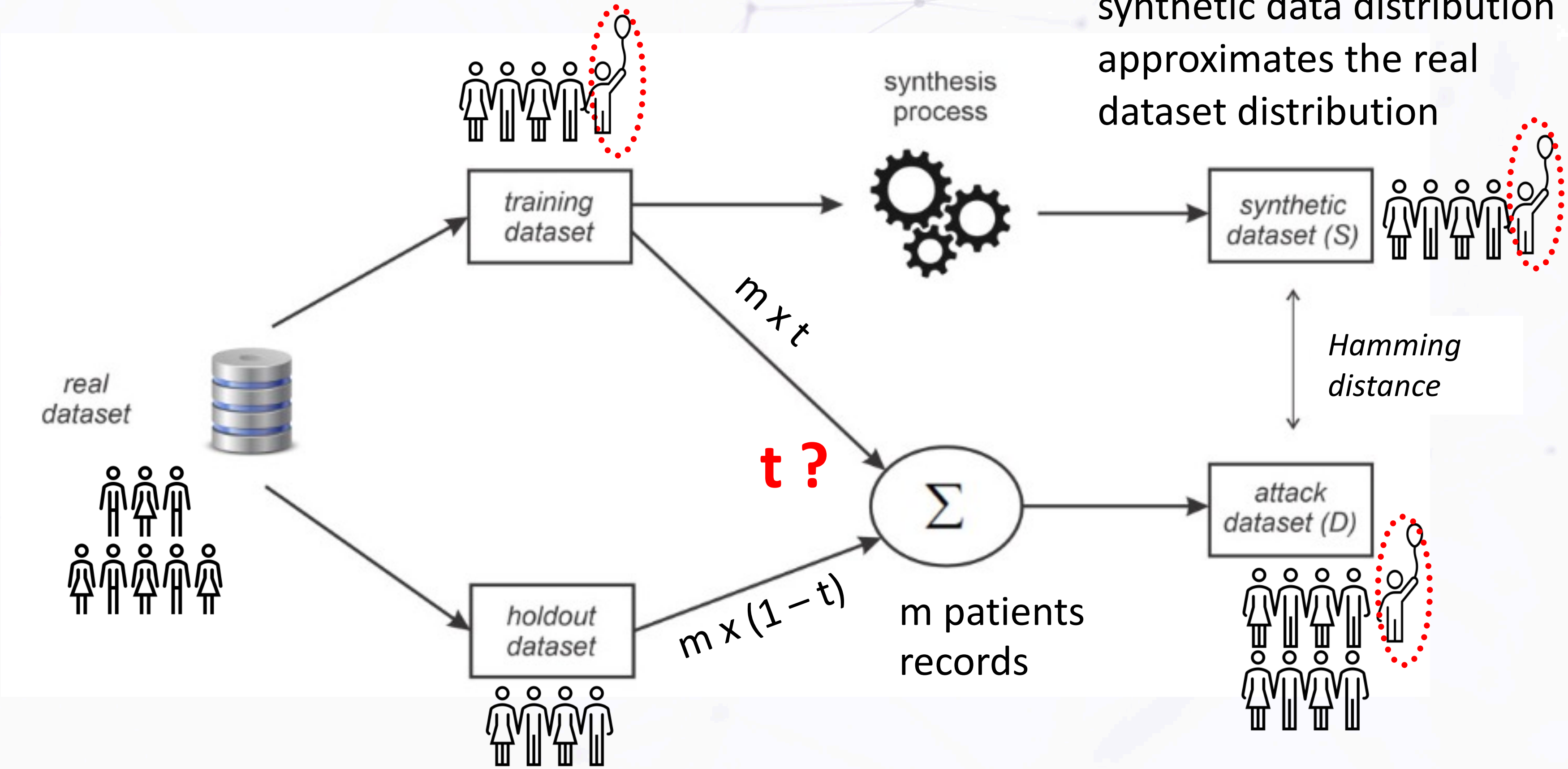


? Does the data custodian have access to the population

The partitioning method

Assumption

synthetic data distribution approximates the real dataset distribution

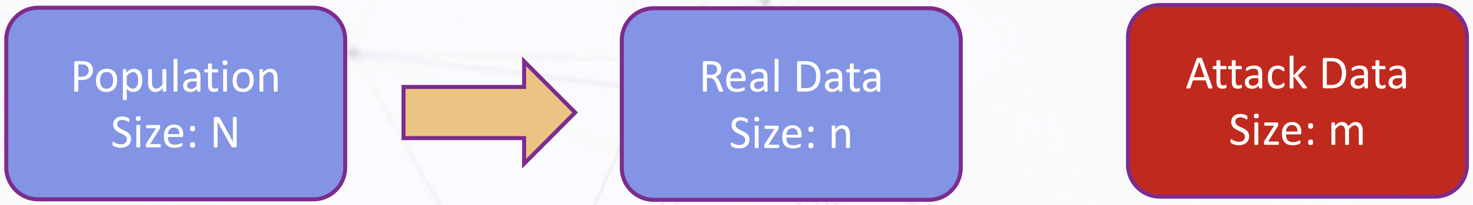


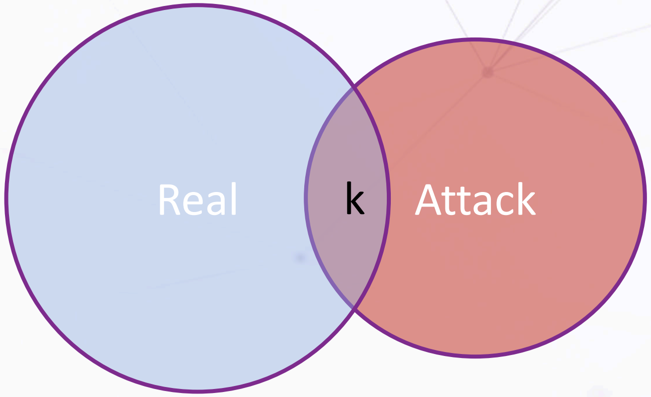
What would be an appropriate value for sampling proportion t? Intuitively, is it 0.5...? Actually, most previous work used 0.5 as the partitioning parameter!

Replica Analytics
AN AETION COMPANY

Parameterizing the partition method

Find t:





Assume there are k individuals in the overlap of real data and attack data, k follows hypergeometric distribution:

$$pr(k = x) = \frac{\binom{N-m}{n-x}\binom{m}{x}}{\binom{N}{n}}$$

Expected value: **mn/N**

Divided by m, the proportion of real records in the attack dataset is: **n/N**

Match synthetic data and attack data

Hamming distance

- The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different.
- Eg. "**kathrin**" and "**kerstin**" is 4.
- In our case, we compare the variables to compute the hamming distance.

Match

- y, record in attack data
- y', record in synthetic data
- L, Hamming distance
- h, pre-defined threshold (cut-off, h = 5, is commonly used in the literature)

$$min_{y'} L(y, y') \leq h$$

Replica Analytics

AN AETION COMPANY

Evaluation metrics

- F1 Score

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$precision = \frac{tp}{(tp + fp)} \qquad recall = \frac{tp}{tp + fn}$$

y, record in attack data
y', record in synthetic data

		Predicted Condition	
		Positive L(y, y’) <=5	Negative L(y, y’)>5
Actual Condition	Positive y in training data	TP	FN
	Negative y not in training data	FP	TN

Simulation study to assess the impact of parameterization

Replica Analytics

AN AETION COMPANY

Given the evaluation metrics, next step is….. **Empirical Demonstration**

Validate t = n/N

- Use the ground truth process to evaluate the membership disclosure risk
- Use the partitioning method to estimate the membership disclosure risk, given various t values between 0 and 1

Estimated membership disclosure risk at t = n/N ≟ **Ground truth membership disclosure risk**

Replica Analytics
AN AETION COMPANY

Data	• Ontario COVID-19 Case dataset • Washington state hospital discharge database • The Canadian Community Health Survey data • The Nexoid COVID-19 behavioral survey
Generative Models	• Sequential tree-based synthesizer (RS) • Generative adversarial network architecture (CTGAN)
Real data size	• 5k, 15k, 25k
Attack data size	• 1k (sufficient records for a stable value of F1)
t	• varied randomly from 0 – 1

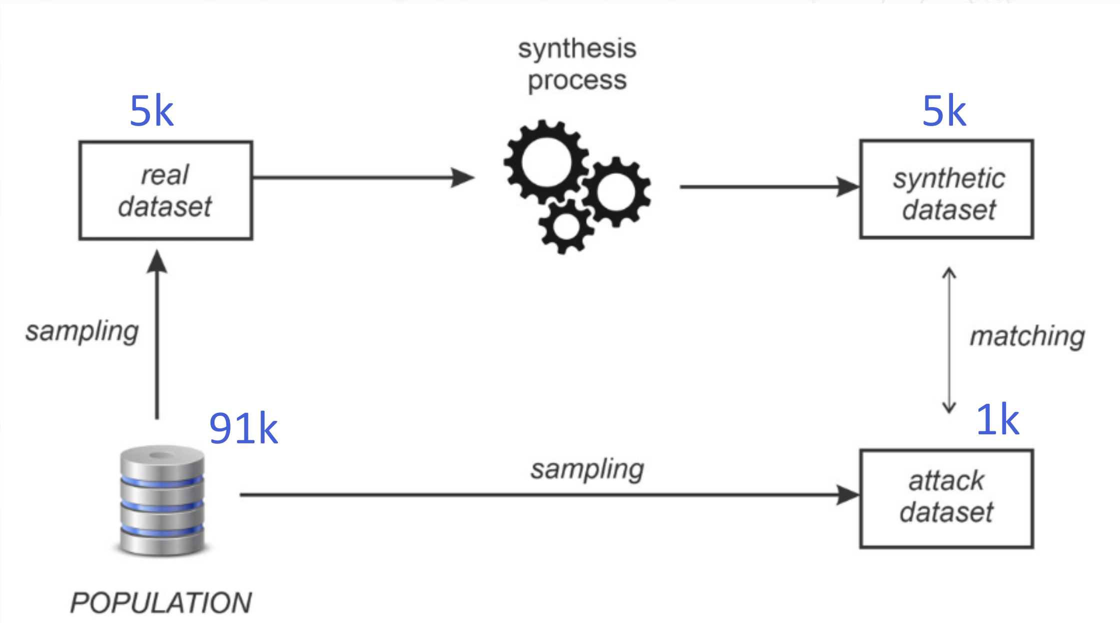
Replica Analytics
AN AETION COMPANY

Given COVID-19 dataset as an example:

Data	Real data size (n)	Population data size (N)	Proposed t (n/N)	Attack data size
COVID-19	5k	91k	0.055	1k
	15k	91k	0.165	1k
	25k	91k	0.276	1k

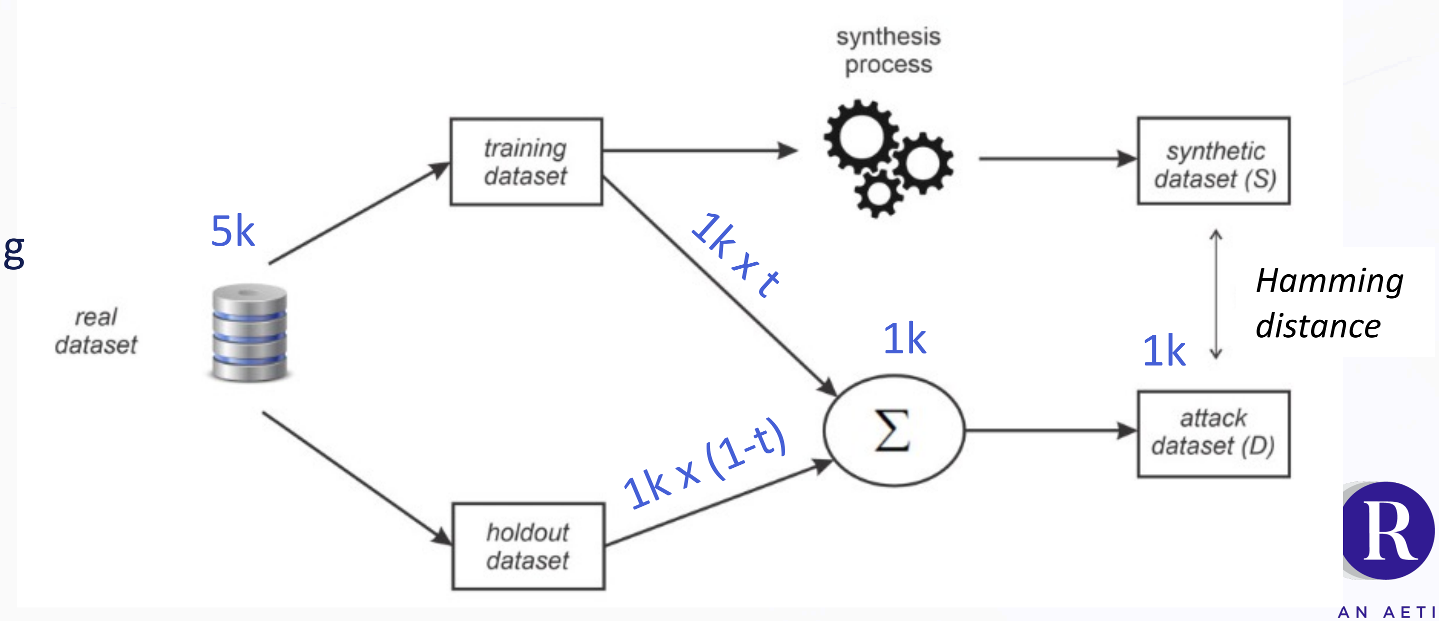
Table 1: The simulation setup of COVID-19 data

Ground Truth

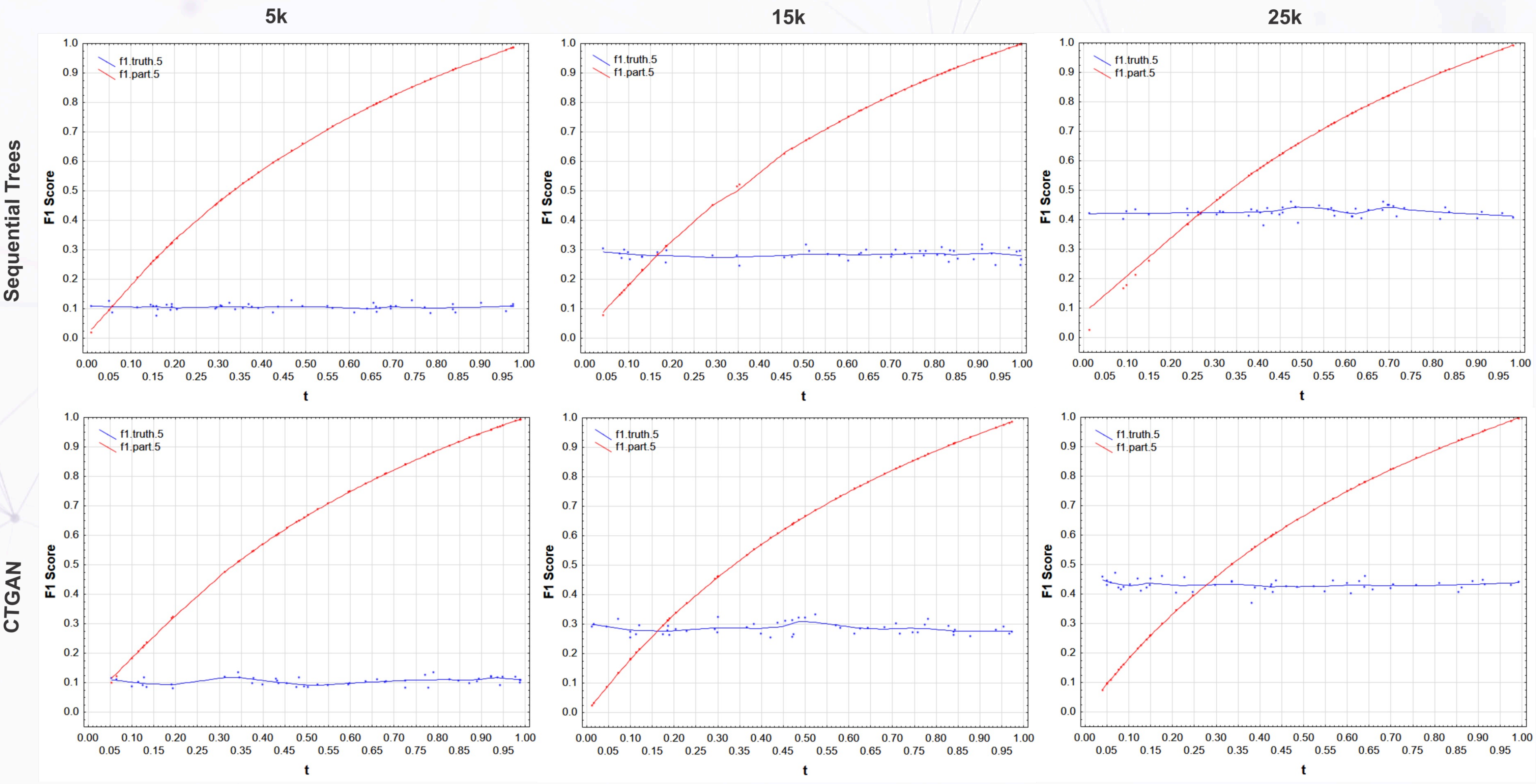


Example of COVID-19 Data Simulation, with real data size = 5k

Partitioning Method



Replica Analytics
AN AETION COMPANY



t = n/N = 0.055

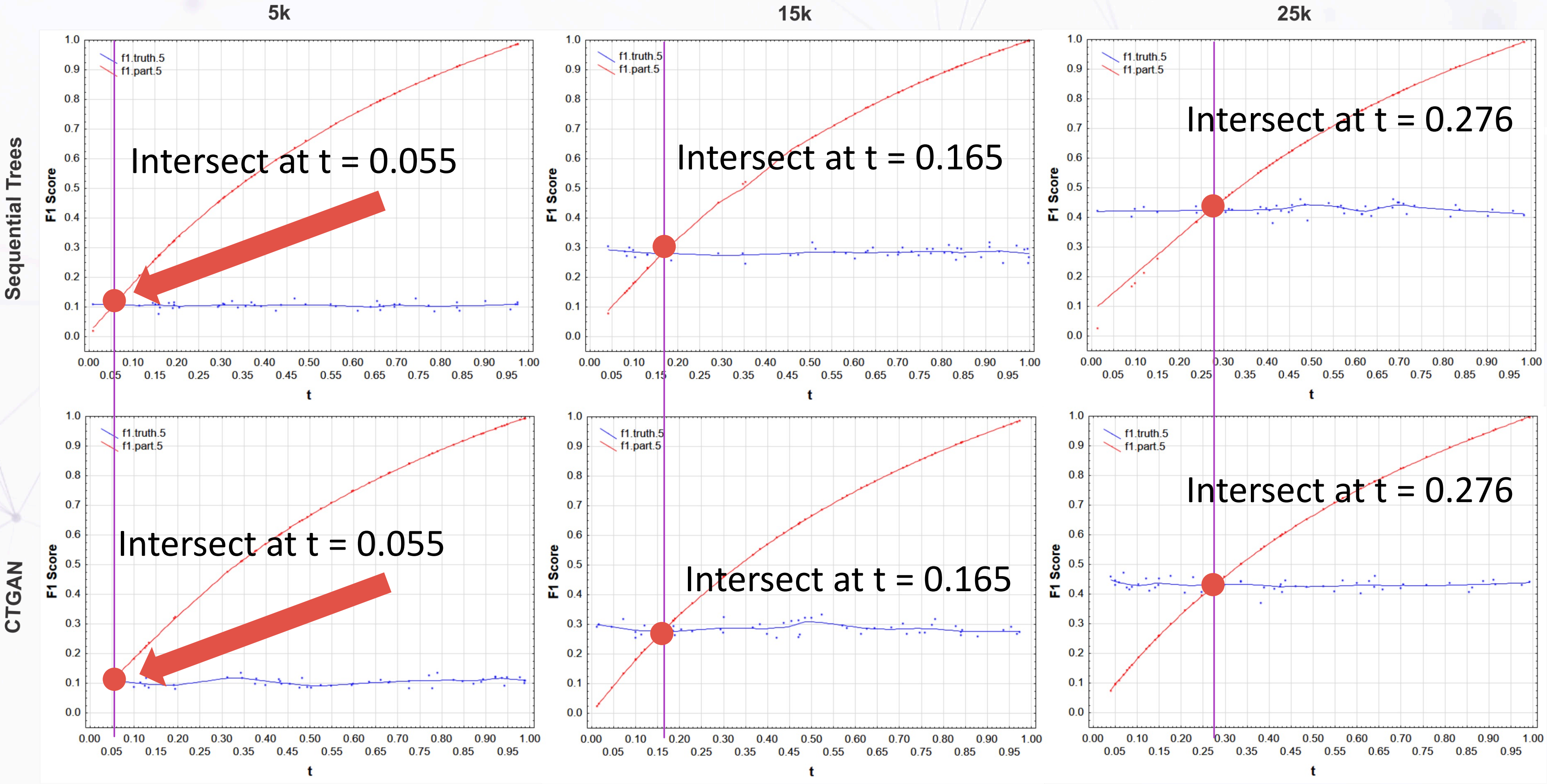
t = n/N = 0.165

t = n/N = 0.276

Figure 1. F1 score results for the COVID-19 dataset showing the ground truth from the simulation and the results using the partition method

Replica Analytics

AN AETION COMPANY



Intersect at t = 0.055

Intersect at t = 0.165

Intersect at t = 0.276

Intersect at t = 0.055

Intersect at t = 0.165

Intersect at t = 0.276

t = n/N = 0.055

t = n/N = 0.165

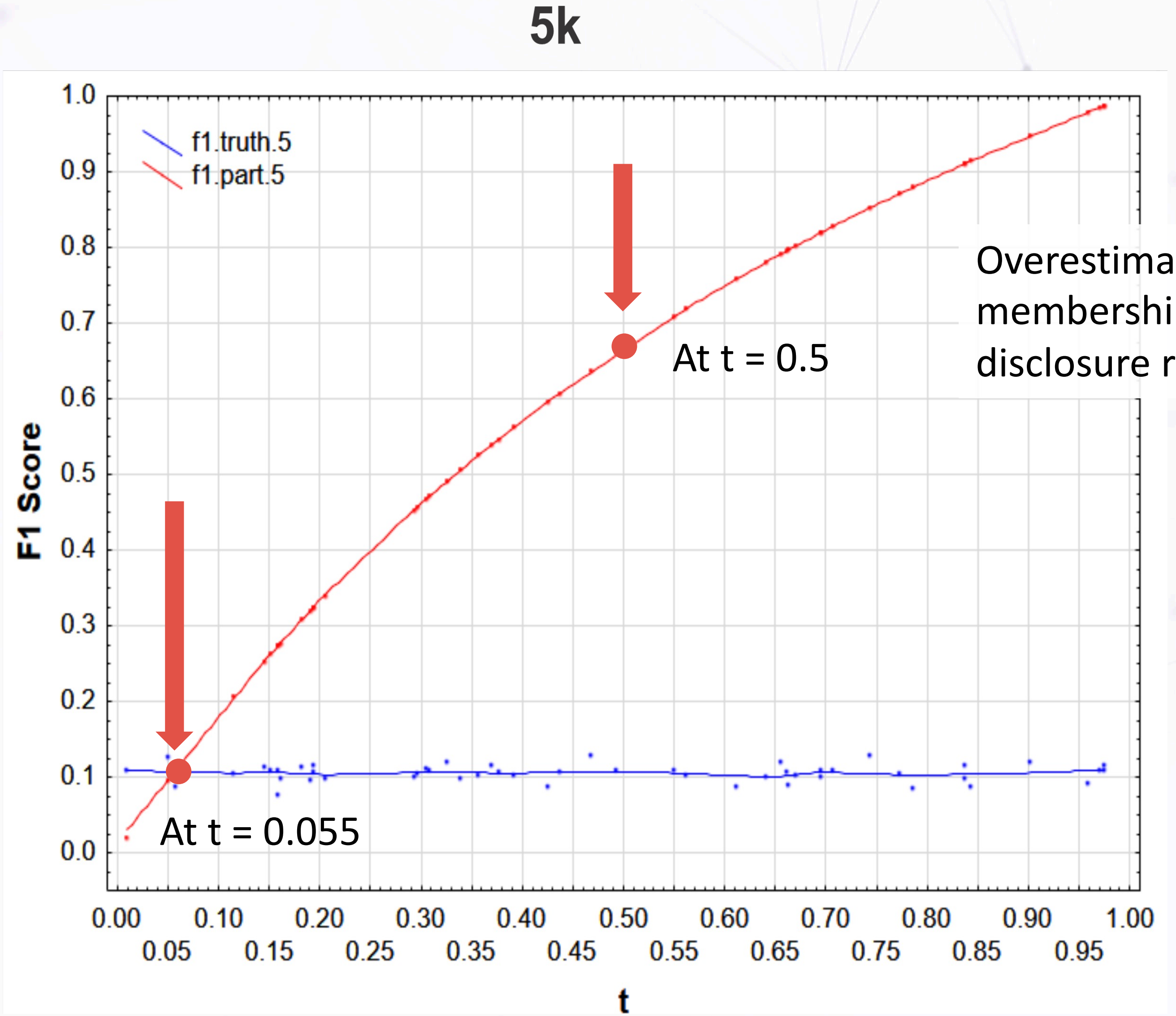
t = n/N = 0.276

Figure 1. F1 score results for the COVID-19 dataset showing the ground truth from the simulation and the results using the partition method

Replica Analytics

AN AETION COMPANY

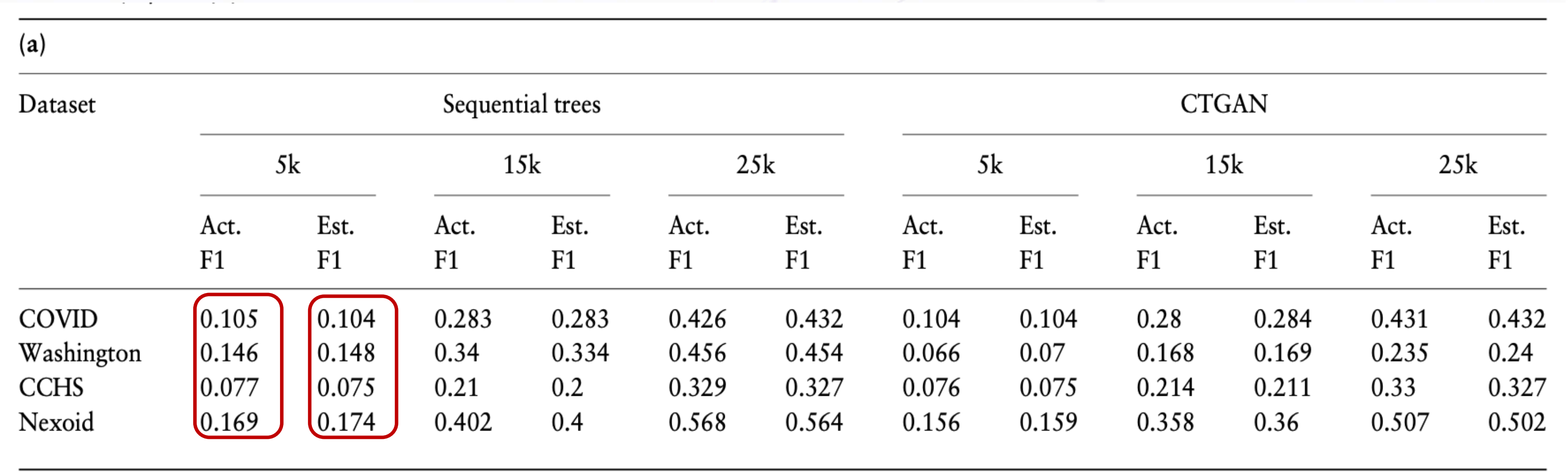
Sequential Trees



At t = 0.5

At t = 0.055

Overestimate the membership disclosure risk!



(a)

Dataset	Sequential trees						CTGAN					
	5k		15k		25k		5k		15k		25k	
	Act. F1	Est. F1	Act. F1	Est. F1	Act. F1	Est. F1	Act. F1	Est. F1	Act. F1	Est. F1	Act. F1	Est. F1
COVID	0.105	0.104	0.283	0.283	0.426	0.432	0.104	0.104	0.28	0.284	0.431	0.432
Washington	0.146	0.148	0.34	0.334	0.456	0.454	0.066	0.07	0.168	0.169	0.235	0.24
CCHS	0.077	0.075	0.21	0.2	0.329	0.327	0.076	0.075	0.214	0.211	0.33	0.327
Nexoid	0.169	0.174	0.402	0.4	0.568	0.564	0.156	0.159	0.358	0.36	0.507	0.502

Table 2: F1 score results.
the ground truth F1 values (from the simulation) versus the F1 values estimated using the partitioning method when t = n/N

Applications of this membership disclosure estimator

How can we assess whether a synthetic dataset has an acceptable membership disclosure risk?

Two challenges with interpreting this membership disclosure estimate in synthetic datasets:

- F1 score can be difficult to interpret:
  - Depends on the distribution of positive classes (proportion of real records in the attack dataset)
  - F1 values won’t have a consistent interpretation with different datasets
- Real sample datasets that are a large proportion of the population will have a higher risk of membership disclosure regardless of the synthesis process

Evaluation metrics

We propose a corrected F1 score relative membership disclosure risk estimate M:

$$M = \frac{F - F_{max}}{1 - F_{max}} \qquad F_{max} = \frac{2 \times {}^{n}/_{N}}{1 + n/N}$$

Where F_{max} is the maximum F1 score that can be achieved if the adversary has no knowledge of the real dataset

- Note: M is undefined when Fmax = 1, no additional improvements are possible

Assessment threshold

Threshold used in the literature is that up to a 20% increase in accuracy over a naïve baseline can be an acceptable threshold for membership disclosure risk

- M<=0.2 is acceptable, M > 0.2 is unacceptable

$$M = \frac{F - F_{max}}{1 - F_{max}}$$

$$F_{max} = \frac{2 \times {}^{n}/_{N}}{1 + n/N}$$

- Negative values indicate decreased accuracy compared to a naïve baseline, meaning the synthesis process lowers membership disclosure risk

Application in clinical trial datasets

We applied the partitioning method in membership disclosure risk evaluation on 7 oncology trial datasets

- Objective: determine what the privacy risks would be for synthetic variants, and whether these risks would be deemed acceptably small.
- Larger picture: growing interest in making clinical trial datasets available (without privacy concerns).

Application methods

- Generative model: Sequential tree-based synthesizer (RS)
- The population size of each trial (N)
  - For each trial, we identified the population by summing up the number of participants of other trials in the same therapeutic area over the same study period and with overlapping geographies from **ClinicalTrials.gov**
- The size of each trial dataset (n).

NIH U.S. National Library of Medicine

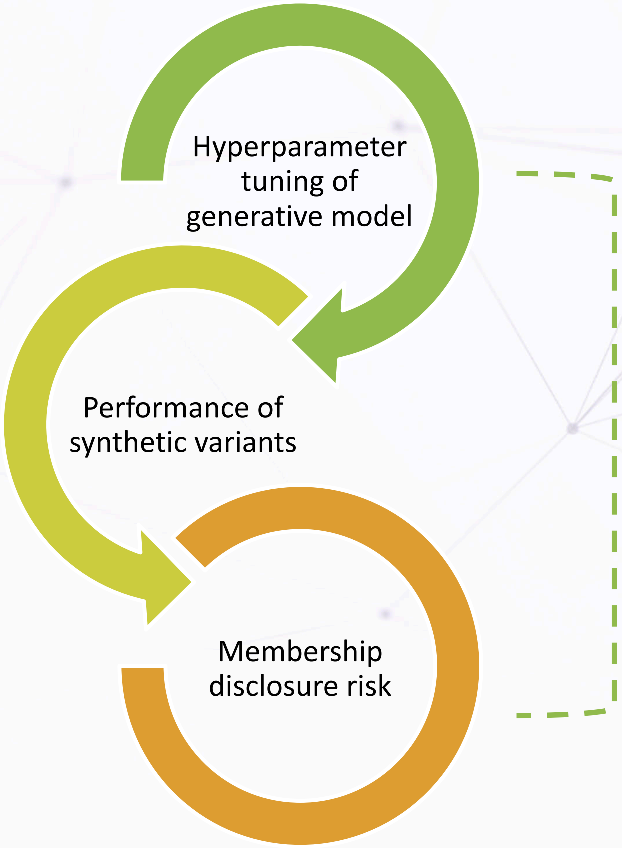
ClinicalTrials.gov

Application results

Data	Dataset size (n)	Population size (N)	M
Trial #1 National Cancer Institute	773	1310	-1.42
Trial #2 Clovis Oncology	367	19255	-0.0137
Trial #3 Sanofi	746	21875	-0.034
Trial #4 Amgen	370	58381	-0.0137
Trial #5 Amgen	520	5868	-0.0947
Trial #6 Amgen	479	16484	-0.0322
Trial #7 NCCTG	1543	27526	0.052

Table 3: Summary of the oncology trials used on the analysis with the study size and the population, as well as the membership disclosure risk.

Application for risk mitigation


Hyperparameter tuning of generative model
Performance of synthetic variants
Membership disclosure risk

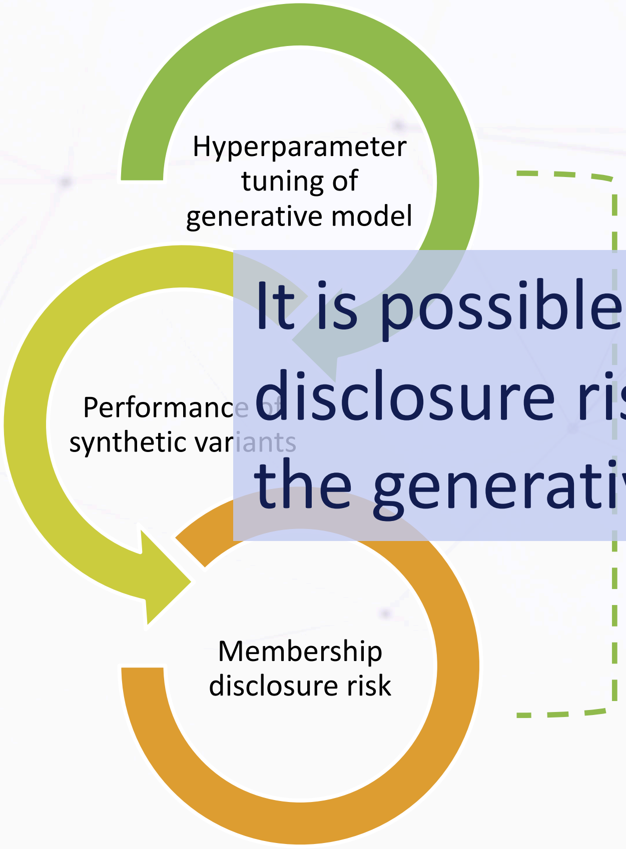
Loss function for hyperparameter tuning:

Risk-utility loss

$$\text{loss}_{\text{RU}} = -\max\left([M > 0.2] \times \left(0.31 + \frac{1}{1 + \mathrm{e}^{(M-1)}}\right), [M \le 0.2]\right) \times U$$

- *U*, the utility metric
- [M>0.2] and [M<=0.2] are Iverson brackets.

Application for risk mitigation


Hyperparameter tuning of generative model
Performance of synthetic variants
Membership disclosure risk

Loss function for hyperparameter tuning:

Risk-utility loss

$$\mathrm{loss}_{\mathrm{RU}} = -\max\left([M > 0.2] \times \left(0.31 + \frac{1}{1 + \mathrm{e}^{(M-1)}}\right), [M \leq 0.2]\right) \times U$$

- *U*, the utility metric
- [M>0.2] and [M<=0.2] are Iverson brackets.

It is possible to ensure the membership disclosure risk is acceptably small within the generative model development!

Replica Analytics

AN AETION COMPANY

Conclusions

Replica Analytics

AN AETION COMPANY

Conclusions

- Our proposed parameterization provides a theoretically and empirically grounded basis for evaluating membership disclosure risk for synthetic data.

- Sequential tree-based synthesizer (RS) produces synthetic oncology clinical trial with low membership disclosure risk, enabling their broader sharing within the research community.

- The risk – utility loss function can optimize for membership disclosure risk within the model development rather than as a post hoc assessment.

Limitations

- We consider the average membership disclosure risk across iterations because of the variation driven by the sampling variability. The average is a good representation of the general membership disclosure risk level, but it does not account for the worse case situation.

- The membership disclosure metric is applicable to tabular data. Our future work should extend these membership disclosure estimators to longitudinal datasets.

- There are other types of privacy risks, all of which should be considered when assessing synthetic data (e.g., attribution risk).

Acknowledgements

Collaborators

CHEO
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

uOttawa

UNIVERSITY OF ALBERTA

Funding

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NSERC
CRSNG

BILL & MELINDA GATES foundation

Computational Resources

Compute Ontario

Replica Analytics
AN AETION COMPANY

Questions?

Replica Analytics

AN AETION COMPANY

Thank you!

Replica Analytics

AN AETION COMPANY