Evaluation of the Privacy Risks in Synthetic Clinical Trial Datasets

Khaled El Emam[1,2,3], Lucy Mosquera[1,3], Xi Fang[3], Samer Kababji[1], Nicholas Mitsakakis[1], Ana-Alicia Beltran-Bless[4], Greg Pond[5], Lisa Vandermeer[4], Dhena Radhakrishnan[1], Mark Clemons[4]

[1] CHEO Research Institute, Ottawa, Canada

[2] School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

[3] Replica Analytics Ltd, Ottawa, Canada

[4] Ottawa Hospital Research Institute, Ottawa, Canada

[5] McMaster University, Hamilton, Canada

Agenda

- Introduction to synthetic data
- Overview of privacy risks
- Our work on membership disclosure metrics
- Membership disclosure risk evaluation on 12 oncology clinical trial datasets

CHEO

Replica Analytics

AN AETION COMPANY

Introduction to synthetic data

Synthetic data

WHAT IT IS

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

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**.

Real

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

Synthetic

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

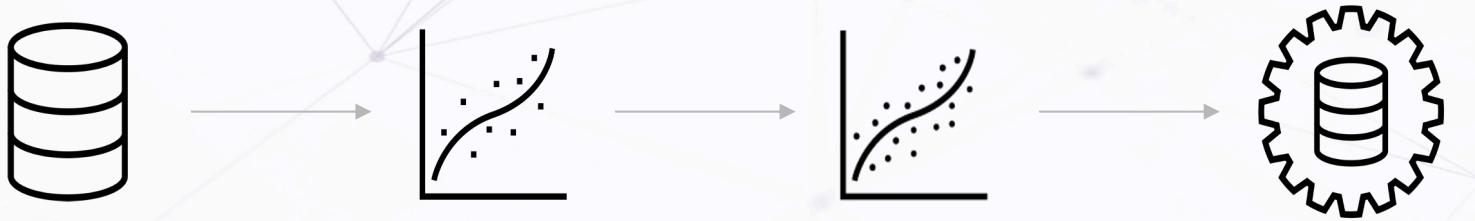
CHEO

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.



CHEO

Replica Analytics

AN AETION COMPANY

Overview of Privacy Risks

Privacy concerns with synthetic data

Identity Disclosure

Year of Birth	Gender	Income
1959	Male	$200k
1959	Male	$220k
1970	Female	$120k

Identity disclosure: The attacker tries to match a record in your dataset to a person in the real world.

Attacker knows:

- David is in the data
- He is male, born in 1959
- His income is $220k

Attribution Disclosure

Year of Birth	Gender	Diagnosis
1959	Male	Prostate Cancer
1959	Male	Prostate Cancer
1970	Female	Breast Cancer

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

Attacker knows:

- Jennifer is in the data
- She is born in 1970

CHEO

Replica Analytics

AN AETION COMPANY

Membership Disclosure Risk Evaluation

Membership disclosure

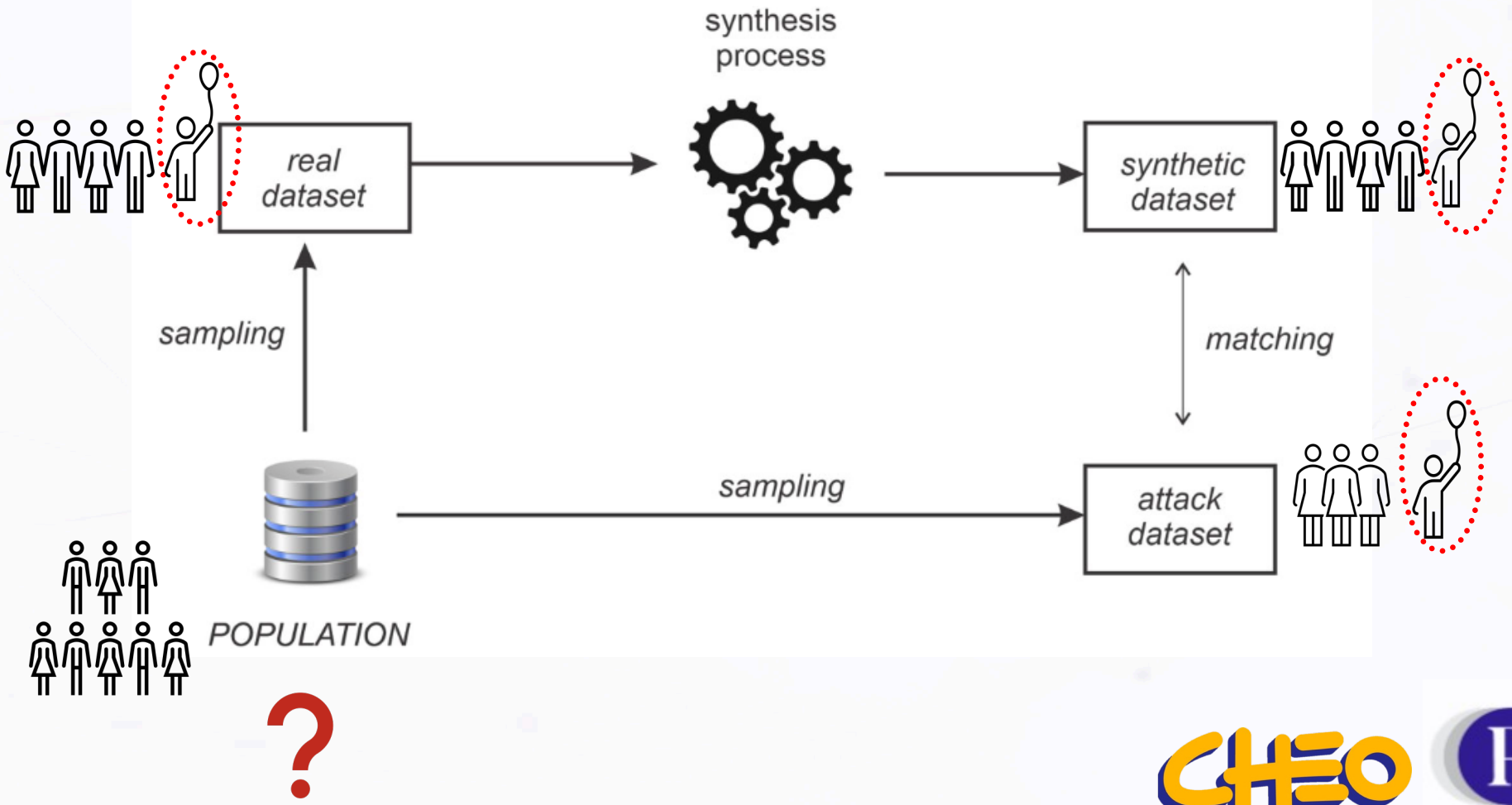
- To what extent an attacker could determine that a target individual is in the training data that was used to train the generative model, then generate the synthetic data
- Knowing that someone is in the training dataset may reveal sensitive information about them, for example, if the dataset was about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

CHEO

Replica Analytics

AN AETION COMPANY

The process (ground truth) for a membership disclosure attack

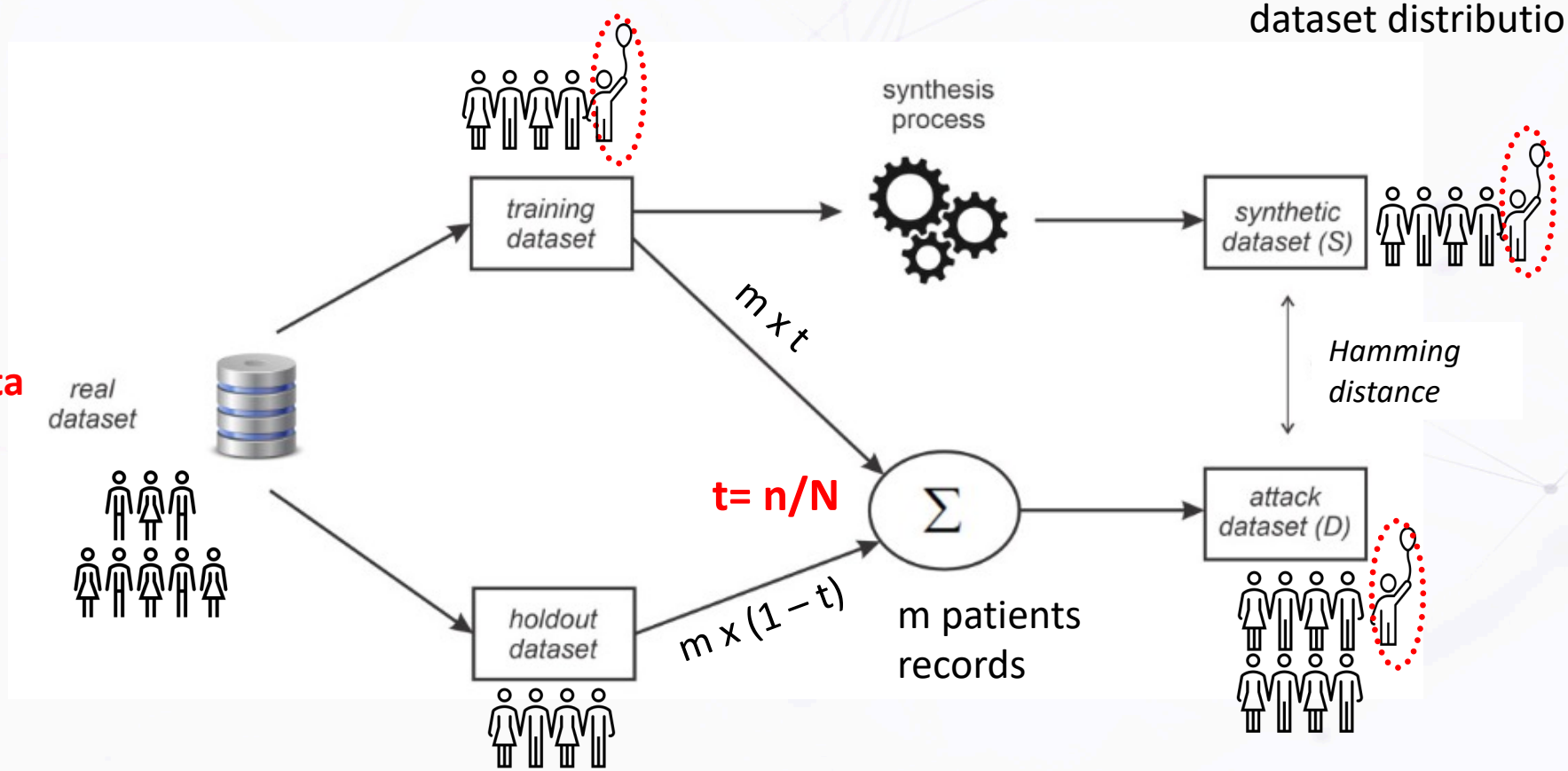


CHEO

Replica Analytics
AN AETION COMPANY

The partitioning method

Assumption
synthetic data distribution approximates the real dataset distribution



Real data size: n

Population size: N

CHEO

Replica Analytics
AN AETION COMPANY

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

Evaluating Privacy Risks in Synthetic Data Using Membership Disclosure

CHEO

Replica Analytics
AN AETION COMPANY

Evaluation metrics

- F1 Score

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad precision = \frac{tp}{(tp + fp)} \qquad recall = \frac{tp}{tp + fn}$$

y, *record in attack data*; **y'**, *record in synthetic data*;
L, *hamming distance*; **h**, *hamming distance threshold*

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

CHEO

Replica Analytics
AN AETION COMPANY

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:

- The 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

CHEO

Replica Analytics

AN AETION COMPANY

Corrected Evaluation Metric

We propose the relative F1 score:

$$F1_{relative} = \frac{F1 - F1_{max}}{1 - F1_{max}}$$

$$F1_{max} = \frac{2 \times {}^{n}/_{N}}{1 + n/N}$$

$F1_{max}$: Maximum F1 score that can be achieved by a naïve attack assuming all the attack records are in the real dataset.

n : Real Dataset Size

N: Population Size

CHEO

Replica Analytics

AN AETION COMPANY

Assessment threshold

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

- $F1_{relative} \leq 0.2$ Acceptable
- $F1_{relative} > 0.2$ Unacceptable
- Negative values indicate decreased accuracy compared to a naïve baseline, meaning the synthesis process lowers membership disclosure risk

CHEO

Replica Analytics

AN AETION COMPANY

Applications of membership disclosure evaluation on 12 oncology trials

CHEO

Replica Analytics

AN AETION COMPANY

Application in clinical trial datasets

We applied the partitioning method in membership disclosure risk evaluation on 12 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).

CHEO

Replica Analytics

AN AETION COMPANY

Application results

Data	Dataset size (n)	Population size (N)	Relative F1
Trial #1 National Cancer Institute	773	1310	-1.44
Trial #2 Clovis Oncology	367	19255	-0.03
Trial #3 Sanofi	746	21875	0.03
Trial #4 Amgen	370	58381	-0.01
Trial #5 Amgen	520	5868	-0.11
Trial #6 Amgen	479	16484	-0.03
Trial #7 NCCTG	1543	27526	0.05
Trial #8 BTA	230	1112	0.18
Trial #9 HER2PLUS	50	2279	-0.11
Trial #10 ILIAD	218	49412	0.02
Trial #11 REACTG	401	6513	0.06
Trial #12 ZOL	211	9076	0.02

CHEO

Replica Analytics

AN AETION COMPANY

Conclusions & Limitations

- Sequential tree-based synthesizer (RS) produces synthetic oncology clinical trial with low membership disclosure risk, enabling their broader sharing within the research community.
- Both the F1 score and relative F1 score are used to evaluate the membership disclosure risk. They are complementary.
- 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).

CHEO

Replica Analytics

AN AETION COMPANY

Acknowledgements

Collaborators

McMaster University

uOttawa

The Ottawa Hospital Research Institute | L'Hôpital d'Ottawa Institut de recherche

Inspired by research. Inspiré par la recherche.
Driven by compassion. Guidé par la compassion.

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

Digital Research Alliance of Canada

CHEO

Replica Analytics
AN AETION COMPANY

Contact Us

Dr. Khaled El Emam

SVP and General Manager at Replica Analytics
Senior Scientist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute
Canada Research Chair in Medical AI at the University of Ottawa

Replica Analytics *https://replica-analytics.com*
CHEO Research Institute *https://www.cheoresearch.ca/*

Email: *kelemam@replica-analytics.com*

CHEO

Replica Analytics
AN AETION COMPANY

Questions?

CHEO

Replica Analytics

AN AETION COMPANY

Thank you!