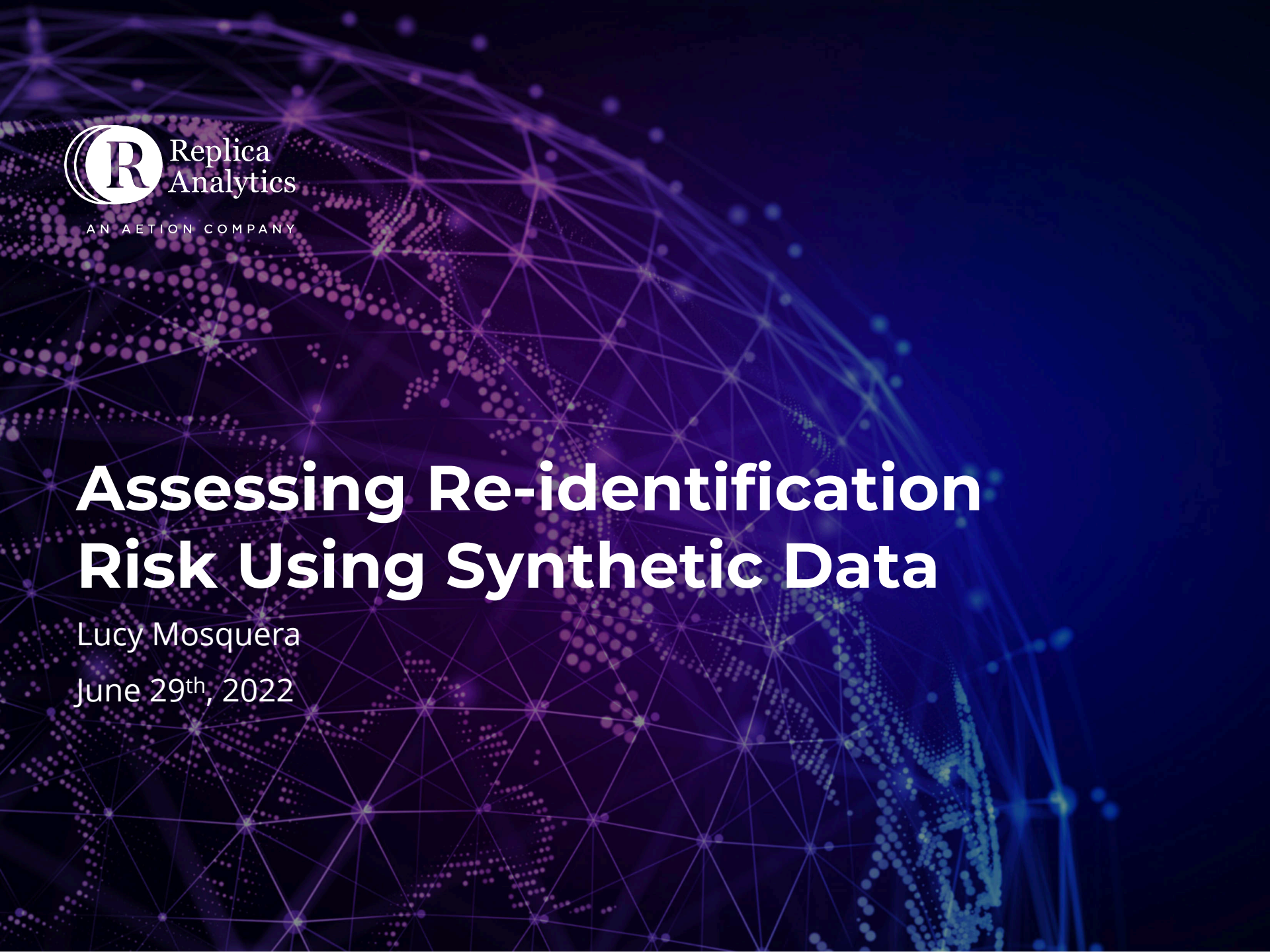
ASSESSING RE-IDENTIFICATION RISK USING SYNTHETIC DATA

JUNE 29, 2022 | 11AM ET

Presented by:

Lucy Mosquera, Director,
Data Science, Replica Analytics



Assessing Re-identification Risk Using Synthetic Data

Lucy Mosquera

June 29th, 2022

Agenda

1. Introduction to re-identification risk
2. Our risk estimator and an intro to synthetic data
3. Performance results for this estimator

Replica Analytics
AN AETION COMPANY

Acknowledgements

This work was conducted with our valued colleagues:

Yangdi Jiang, Bei Jiang, and Linglong Kong

Need for Non-identifiable Data

Privacy regulations such as:

- GDPR in the EEA
- CPPA (C-27) in Canada
- The Privacy Act in Canada
- HIPAA in the United States

Thus far, there is no legislative requirement to obtain additional data subject consent / authorization to use and disclose data for secondary purposes that is deemed to be non-identifiable

Different jurisdictions may use different definitions or thresholds for what is or isn't acceptable; as a general trend the acceptable thresholds are getting stricter over time

Replica Analytics

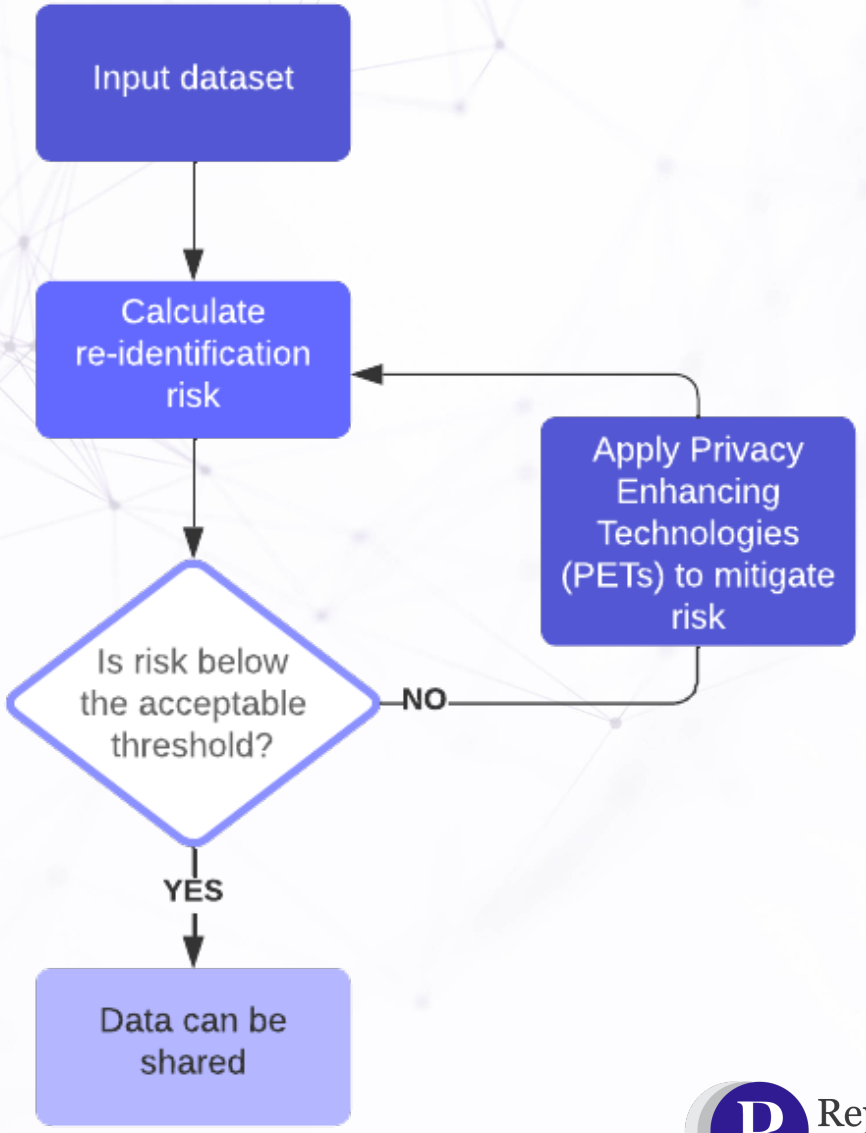
AN AETION COMPANY

Privacy Focused Data Sharing

Privacy Enhancing Technologies (PETs) include:

- Risk-based de-identification (Expert Determination)
- Synthetic data generation
- Homomorphic encryption
- Federated analysis

Different PETs may assess risk using different metrics (e.g., attribution disclosure in synthetic data or delta value in homomorphic encryption).



Replica Analytics
AN AETION COMPANY

Assessing Risk in Synthetic Datasets

JOURNAL OF MEDICAL INTERNET RESEARCH

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

Replica Analytics
AN AETION COMPANY

Re-identification Risk

- Re-identification risk is the probability of being able to correctly match a record in a microdata sample to a real person
- In order to share a dataset, data custodians typically show that the re-identification risk is below an accepted threshold
- Can be expressed as maximum risk, average risk, or uniqueness; this work focuses on average risk

Replica Analytics
AN AETION COMPANY

Microdata

Sex	Year of Birth	NDC
Female	1983	0078-0379
Female	1989	65862-403
Male	1981	55714-4446

Quasi-identifiers

Step 1: Identify the quasi-identifiers in the microdata

Replica Analytics
AN AETION COMPANY

Population

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

Microdata

Sex	Year of Birth	NDC
Female	1983	0078-0379
Female	1989	65862-403
Male	1981	55714-4446

Step 2: Compare microdata records to population using quasi-identifiers

Replica Analytics
AN AETION COMPANY

Population

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

Microdata

Sex	Year of Birth	NDC
Female	1983	0078-0379
Female	1989	65862-403
Male	1981	55714-4446

Step 2: Compare microdata records to population using quasi-identifiers

Replica Analytics
AN AETION COMPANY

Population

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

Microdata

Sex	Year of Birth	NDC
Female	1983	0078-0379
Female	1989	65862-403
Male	1981	55714-4446

Records with the same values for a set of quasi-identifiers are called an equivalence class

Replica Analytics
AN AETION COMPANY

Population

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

Microdata

Sex	Year of Birth	NDC
Female	1983	0078-0379
Female	1989	65862-403
Male	1981	55714-4446

Step 3: Calculate risk for each record in the microdata as 1 divided by the number of records that match in the population; then average across all records

Risk = 1/3 (1/1 + 1/1 + 1/2) = 0.83

Replica Analytics

AN AETION COMPANY

Risk Estimation in Practice

Typically data custodians will not have access to the population level dataset so re-identification risk cannot be calculated empirically

Instead re-identification risk is estimated by making certain assumptions about the population

Existing methods use a variety of estimation techniques including:

- Using microdata sample entropy
- Bayesian methods
- Hypothesis testing

Replica Analytics
AN AETION COMPANY

Challenges of Risk Estimation

Risk estimation is affected by:

- What proportion of the population data is present in the microdata (i.e., sampling fraction)
- How many quasi-identifiers there are, and the overall number of equivalence classes present
- What controls will be implemented when the data are shared

Some risk estimators make strong assumptions, for example, about:

- The independence of quasi-identifiers
- Sample proportions seen in the microdata are the same in the population
- Equivalence class size distribution in the population following a particular distribution

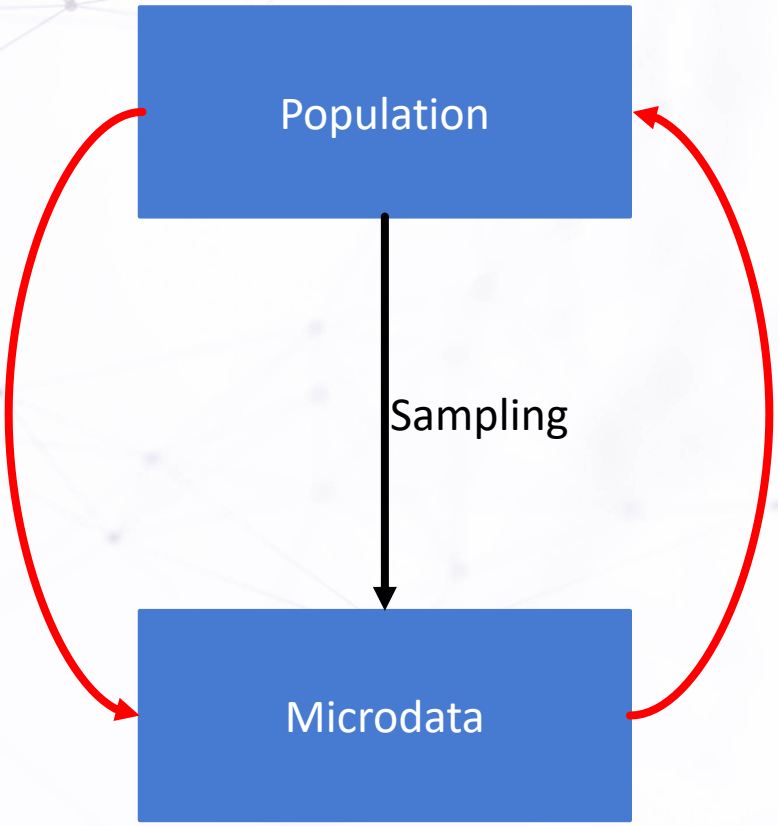
These assumptions can lead to substantial over or under estimation of risk depending on whether or not they are true in real datasets

Direction of Attack

The previous example illustrated a sample to population re-identification attack.

Comprehensive risk assessments will also take into account population to sample attacks.

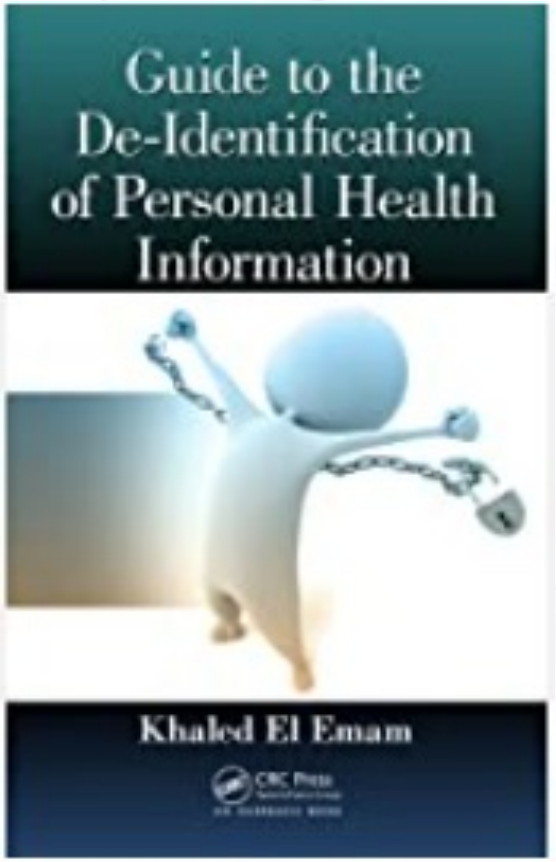
Risk in population to sample attacks is driven by the equivalence class sizes in the sample dataset



Replica Analytics
AN AETION COMPANY

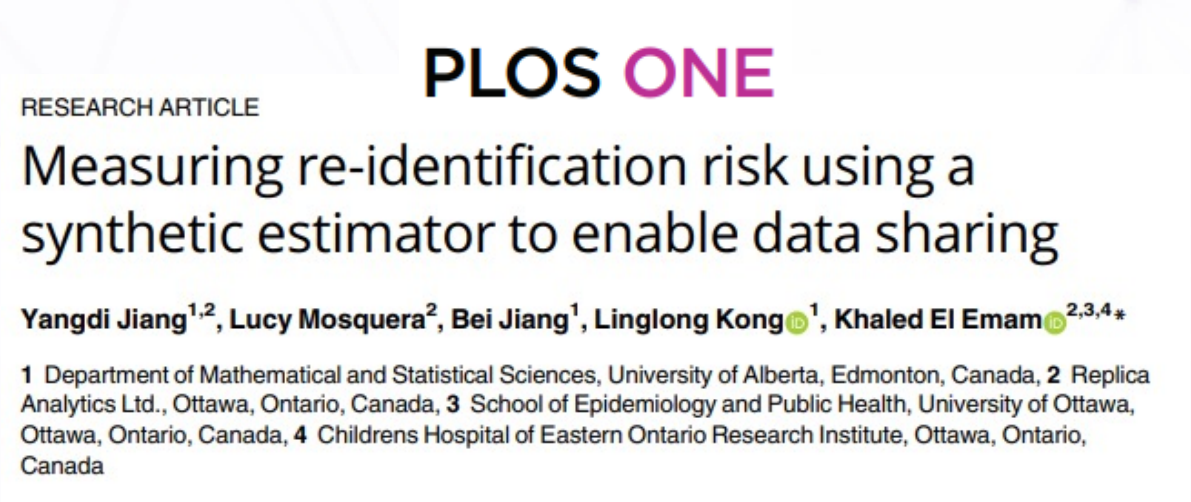
Learn More

More information on re-identification risk assessment strategies and how to anonymize data can be found in:



Replica Analytics
AN AETION COMPANY

Our Work



PLOS ONE

RESEARCH ARTICLE

Measuring re-identification risk using a synthetic estimator to enable data sharing

Yangdi Jiang[1,2], Lucy Mosquera[2], Bei Jiang[1], Linglong Kong[1], Khaled El Emam[2,3,4]*

1 Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada, **2** Replica Analytics Ltd., Ottawa, Ontario, Canada, **3** School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada, **4** Childrens Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada

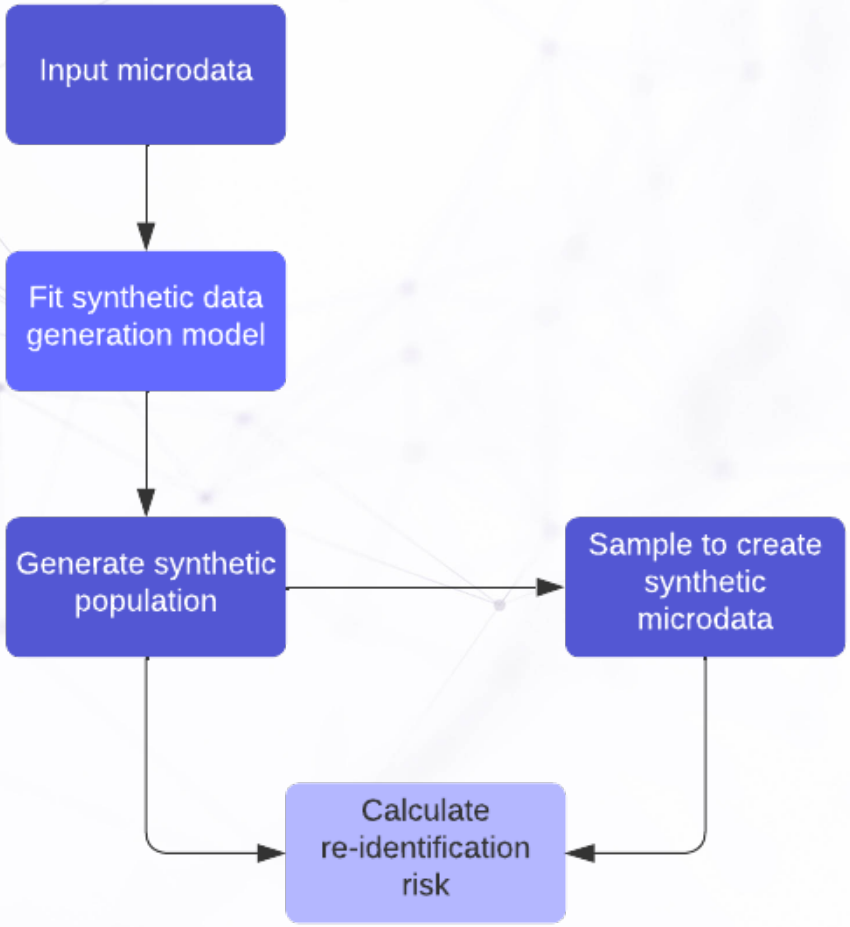
Proposes a new re-identification risk estimator and compares using simulation its performance to 3 popular risk estimators across 4 datasets for a variety of sampling fractions and true risk values.

Replica Analytics

AN AETION COMPANY

Our Estimator

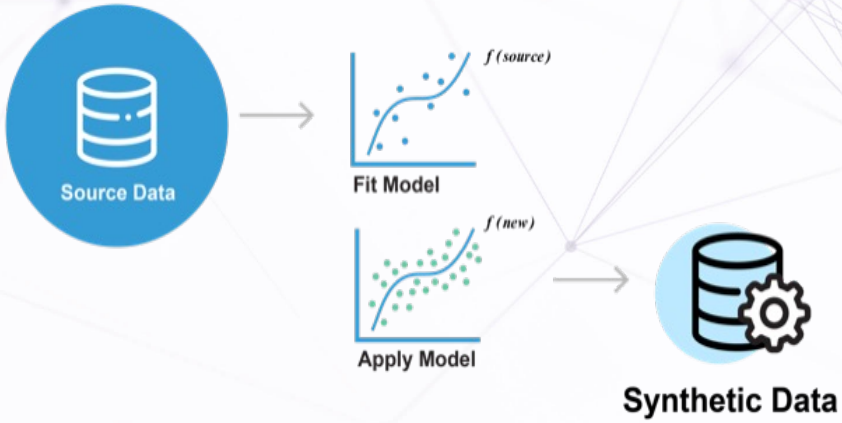
- Our estimator* uses synthetic data generation to create the ‘missing’ population dataset
- This allows the re-identification risk to be calculated empirically
- To compute, the data custodian must:
  - Identify the quasi-identifiers in the dataset
  - Estimate the size of the population



*Patent pending

Replica Analytics
AN AETION COMPANY

Synthetic Data



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

Replica Analytics

AN AETION COMPANY

Synthetic Data Use Cases

Discover Artificial Intelligence

Discover

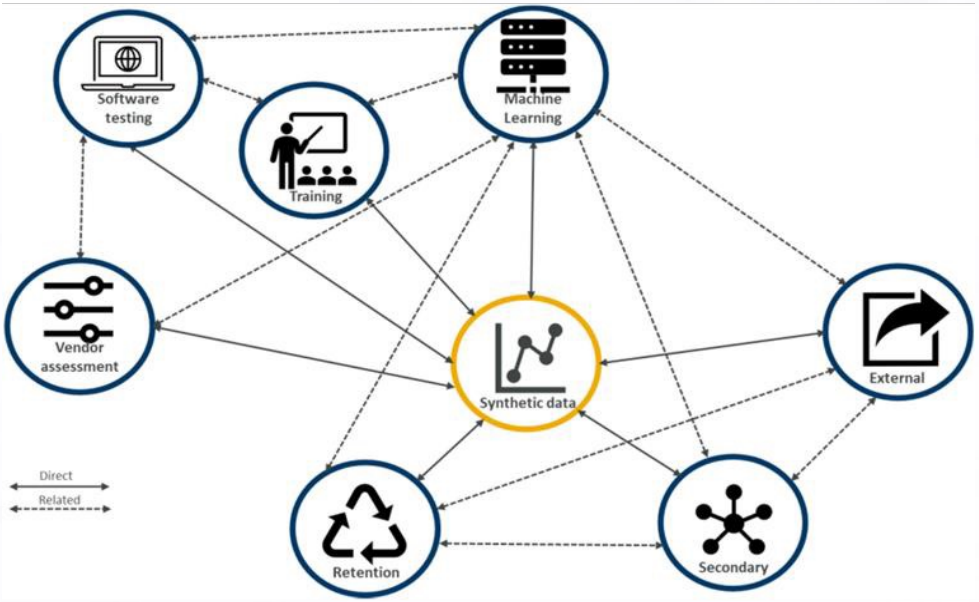
Review

Synthetic data use: exploring use cases to optimise data utility

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

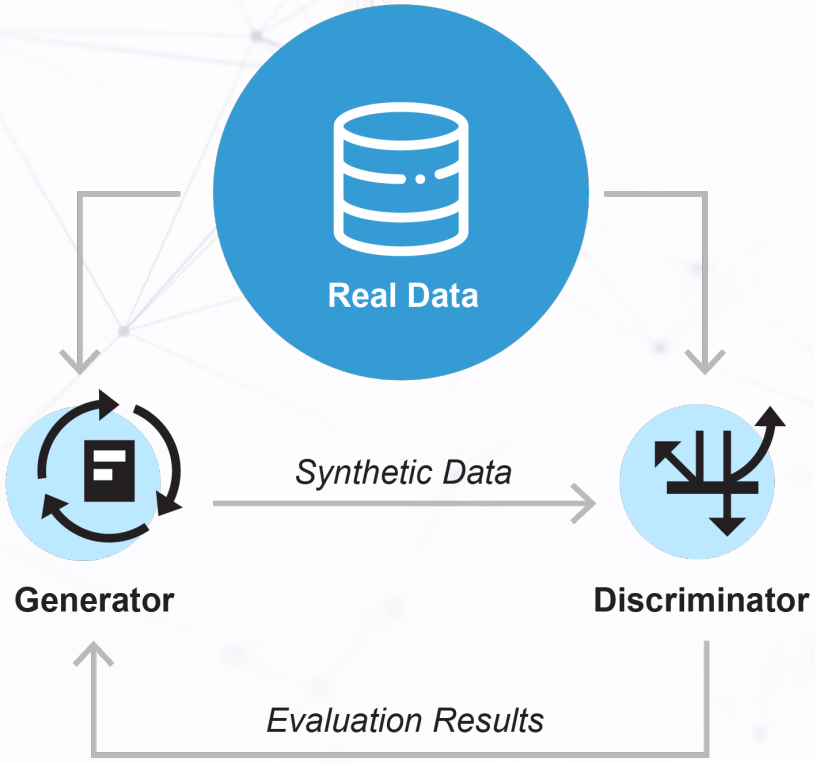
Received: 12 November 2021 / Accepted: 7 December 2021
Published online: 13 December 2021

Can be grouped as:
Privacy use cases
Analytic use cases



Replica Analytics
AN AETION COMPANY

Synthetic Data Generation



Replica Analytics
AN AETION COMPANY

Synthetic Data Generation: Copulas

Copulas are probabilistic models that ‘couple’ together univariate relationships into a multivariate model.

Flexible, compact models that allow complex correlation structures between variables to be modelled

Our work tested 2 types of copulas for data generation: Gaussian and d-vine copulas

Synthetic Data Generation: Copulas

Copulas are fit to the microdata sample using by:

1. Developing a transform to map each variable to a normal distribution using an empirical CDF and Gaussian quantile function
2. Optimizing the correlation between pairs of variables to find the correlation value that minimizes the mutual information between generated data for the pairs of variables

Gaussian Copula

- Models relationships between all pairs of variables

D-vine Copula

- Models relationships based on vine structure

Replica Analytics

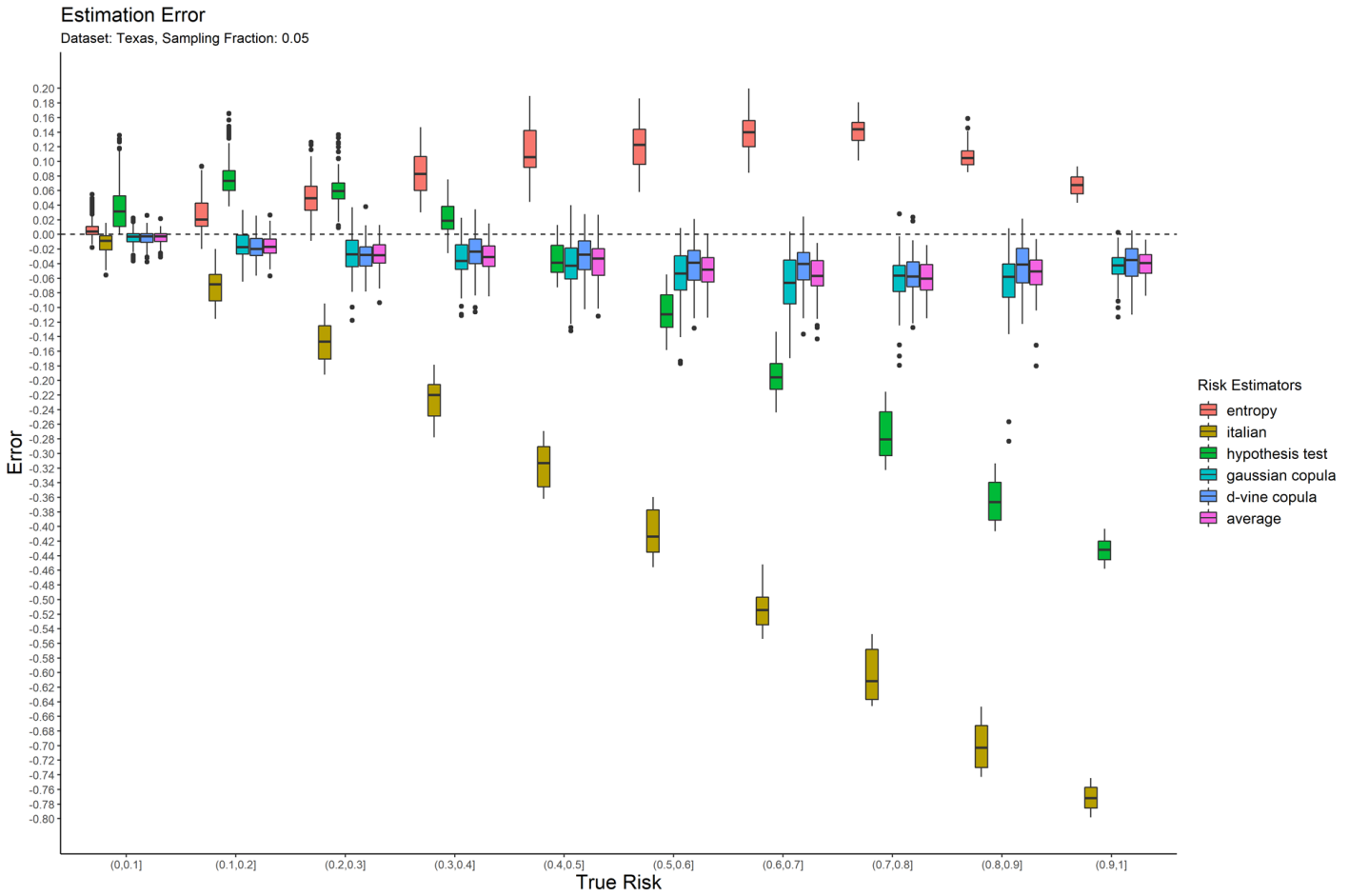
AN AETION COMPANY

Our Assessment Methodology

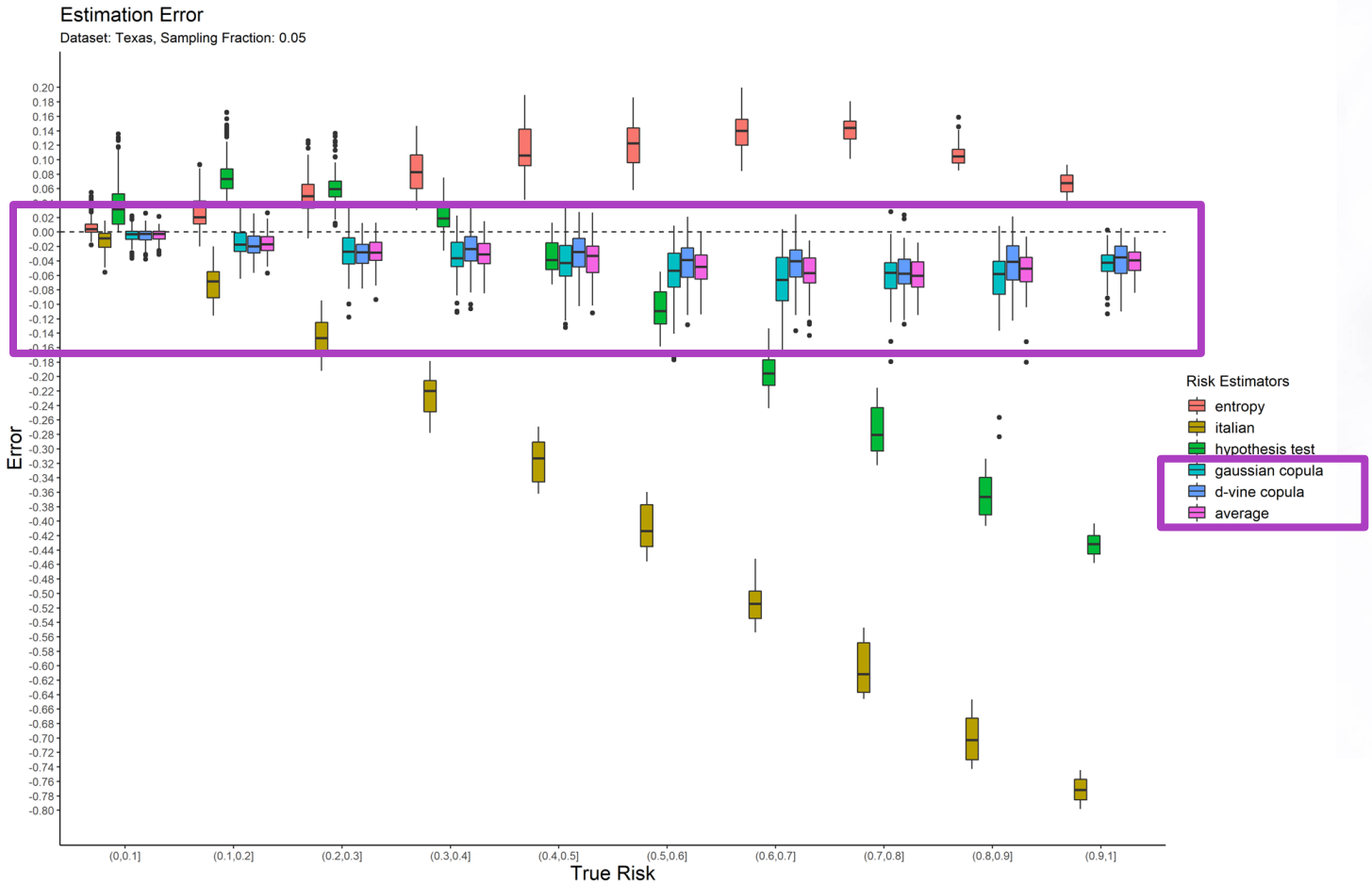
- Assesses 3 variants of our novel estimator: Gaussian copula, d-vine copula, and averaged risk of Gaussian & d-vine copula estimates
- Compared to 3 popular risk estimators: entropy, Bayesian, and hypothesis testing
- 4 different datasets: Texas hospital discharge dataset, Washington hospital discharge dataset, Nexoid COVID survey data, and the UCI adults dataset
- Conducted 1000 iterations per dataset, where each iteration represented a different sampling fraction between 0.01 and 0.99; and a different subset of available quasi-identifiers

Replica Analytics
AN AETION COMPANY

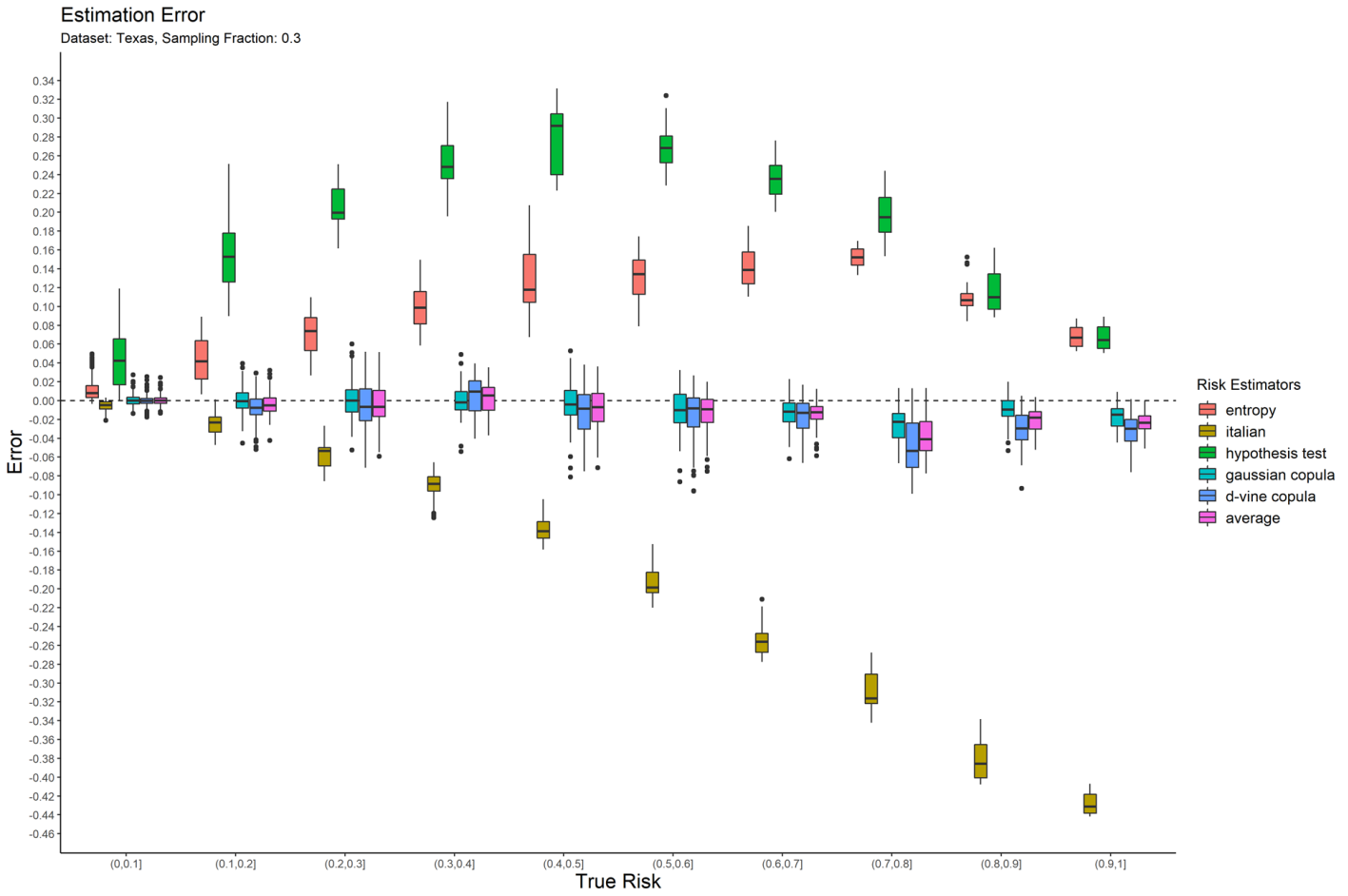
Results: Texas Hospital Discharge



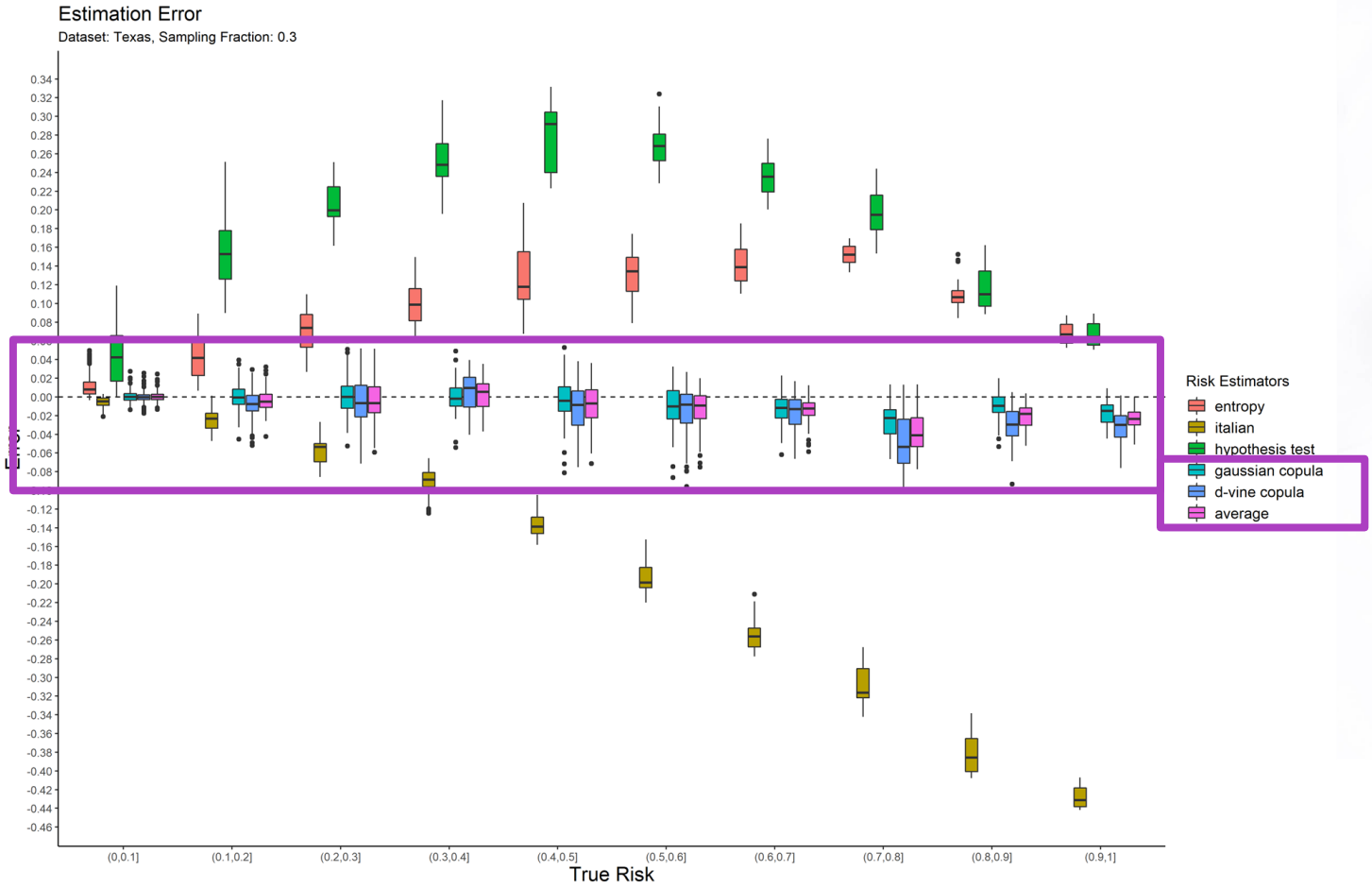
Results: Texas Hospital Discharge



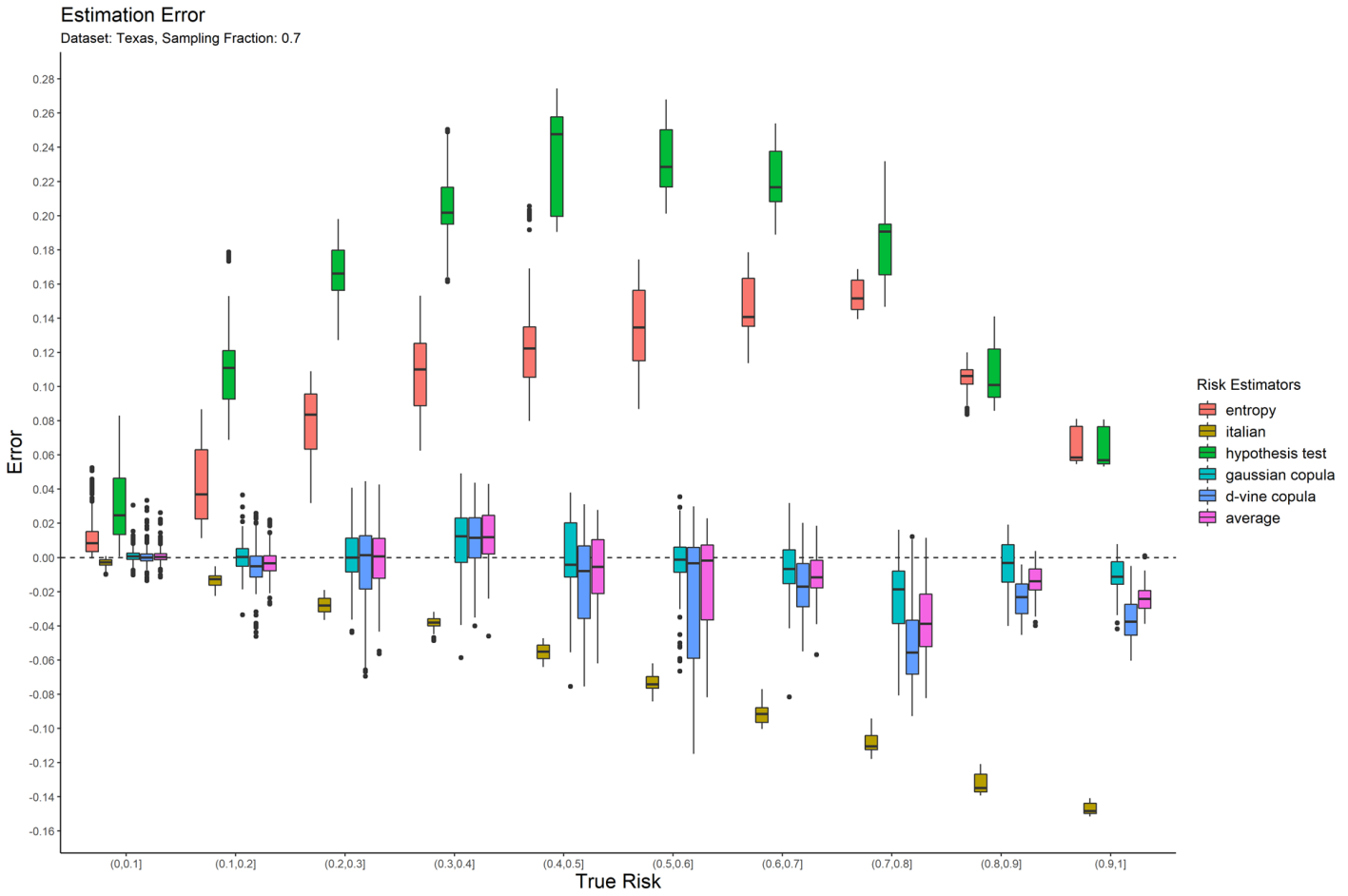
Results: Texas Hospital Discharge



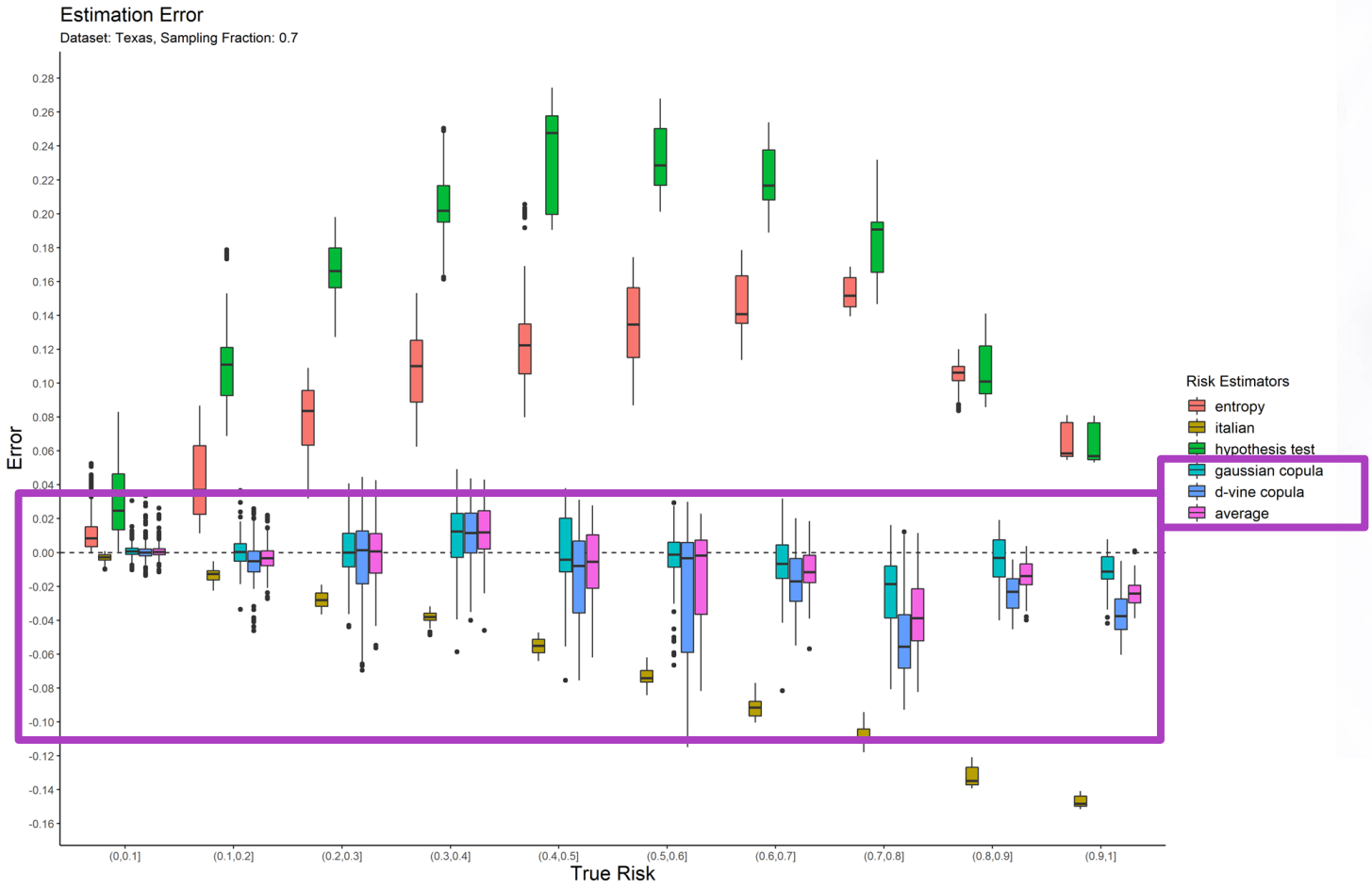
Results: Texas Hospital Discharge



Results: Texas Hospital Discharge



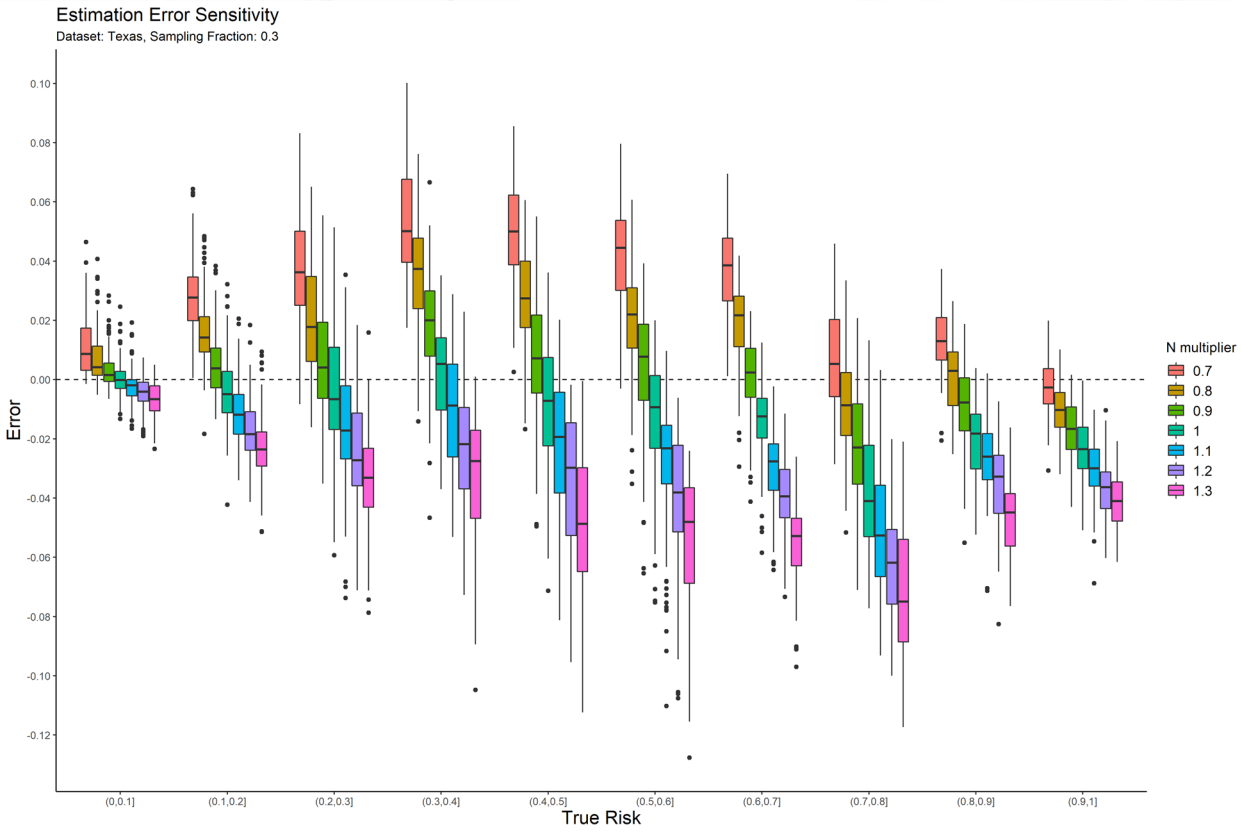
Results: Texas Hospital Discharge



Sensitivity Analysis

Over / under estimation of the true population size results in under / over estimation of the risk, respectively.

+/- 30% of the true population size results in errors within 0.10 of the true value



Replica Analytics

AN AETION COMPANY

Simulation Conclusions

- The entropy method consistently overestimates risk
- The Bayesian (Italian) method consistently underestimates the risk
- The hypothesis estimator overestimates for high sampling fraction and underestimates for lower sampling fractions
- Our risk estimator is highly accurate with the median error in estimated risk less than 0.05
- Our risk estimator is most accurate when the true risk lies between 0 and 0.2; which is where the typical threshold of 0.09 lies and accuracy is most important
- If there is uncertainty about the true population, it is better to synthesize a smaller population as it will produce a more conservative risk estimate

Case Study: Anonymizing flatten.ca COVID data

Goal: assess re-identification risk, apply transformations to mitigate risk so the data can be shared with additional controls

- Online survey of Ontario residents about their experiences with COVID-19
- 18,903 observations in the microdata, with a simulated population of 13,448,494 (the population of Ontario at the time)
- Also performed additional assessments to ensure no quasi-identifier values in the microdata were unique in the population using census data

Replica Analytics
AN AETION COMPANY

Case Study: Anonymizing flatten.ca COVID data

Table shows variables present in the dataset and the transformations applied to mitigate risk

- The sample to population risk after generalization was 0.0723 and the population to sample risk was 0.0009

Variable	Generalizations
Date	Converted to month format
FSA	Forward Sortation Area, which is the first three characters of the postal code
Conditions	Medical conditions diagnosed
age_1	Age categories: <26, 26-44, 45-64, >65
travel_outside_canada	Travel outside Canada in the last 14 days (binary)
Ethnicity	
Sex	
tobacco_usage	
travel_work_school	
covid_results_date	Converted to month format
people_in_household	Removed

Replica Analytics
AN AETION COMPANY

Conclusions

- Our risk estimator produces highly accurate estimates of re-identification risk across a wide range of sampling fractions and true risk values
- We validated our estimator against 3 common risk estimators using 4 different datasets during simulation and a case study
- Risk estimator is integrated into our Replica Synthesis software, making it very easy to use and scaleable
- Our work shows another area for the opportunity of synthetic data generation as part of a privacy assessment workflow

Replica Analytics

AN AETION COMPANY

Questions?

Replica Analytics

AN AETION COMPANY

Thank you!

lmosquera@replica-analytics.com

Replica Analytics

AN AETION COMPANY