DAV/DGVFM
Herbsttagung
2025

Advancing Actuarial Science – Leveraging Synthetic Data for Privacy-Preserving Modelling

DAV/DGVFM Herbsttagung 2025

Noa Zamstein

earnix

Agenda

- Motivation: handling personal information
  - Synthetic data as an enabler
- Why synthetic data matters in actuarial science
- Principles of synthetic data generation
- Utility and privacy considerations
- Summary and conclusions

CALIFORNIA

Three dead, four hospitalized after fiery, high-speed crash in Orange

[redacted] data are obtained from various States' documents, such as the following.

- Police Crash Reports
- Death Certificates
- State Vehicle Registration Files
- Coroner/Medical Examiner Reports
- State Driver Licensing Files
- State Highway Department Data
- Emergency Medical Service Reports
- Vital Statistics and Other State Records

From these documents, the analysts code more than 140 [redacted] data elements. The specific data elements may be modified slightly each year to conform to changing user needs, vehicle characteristics, and highway safety emphasis areas. The data collected within [redacted] do not include any personal identifying information such as names, addresses, or social security numbers. Thus, any data kept in [redacted] data files and made available to the public fully conform to the Privacy Act.

Linking between different data sources can reveal a lot!

Driver data

ST_CASE	STATENAME	AGE	COUNTY	MONTH	HOURNAME	VPICMAKENAME	BODY_TYPNAME	PER_TYPNAME
60609	California	19	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Driver of a Motor Vehicle In-Transport
60609	California	19	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport
60609	California	17	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport
60609	California	14	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport
60609	California	14	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport
60609	California	14	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport
60609	California	26	59	7	2:00am-2:59am	Nissan	4-door sedan, hardtop	Passenger of a Motor Vehicle In-Transport

+

Drugs data

STATE	STATENAME	ST_CASE	VEH_NO	PER_NO	DRUGSPEC	DRUGSPECNAME	DRUGRES	DRUGRESNAME
6	California	60609	1	1	2	Urine	6002	Phencyclidine
6	California	60609	1	1	2	Urine	5060	Delta 9-tetrahydrocannabinol [THC]
6	California	60609	1	2	0	Test Not Given	0	Test Not Given
6	California	60609	1	3	99	Reported as Unknown if Tested	9999	Reported as Unknown if Tested for Drugs
6	California	60609	1	4	0	Test Not Given	0	Test Not Given
6	California	60609	1	5	99	Reported as Unknown if Tested	9999	Reported as Unknown if Tested for Drugs
6	California	60609	1	6	0	Test Not Given	0	Test Not Given
6	California	60609	1	7	99	Reported as Unknown if Tested	9999	Reported as Unknown if Tested for Drugs

+

CNBC

earnix

The problem

There is a huge difficulty in mobilizing personal data due to regulations and privacy concerns

Contrary to intuition, simple de-identification is often insufficient

Common solutions produce data that is either aggregated or violates privacy

earnix

The solution

Create **fictitious** data that prevents identification

Individual level data for modelling

Maintains the **utility** of the original data

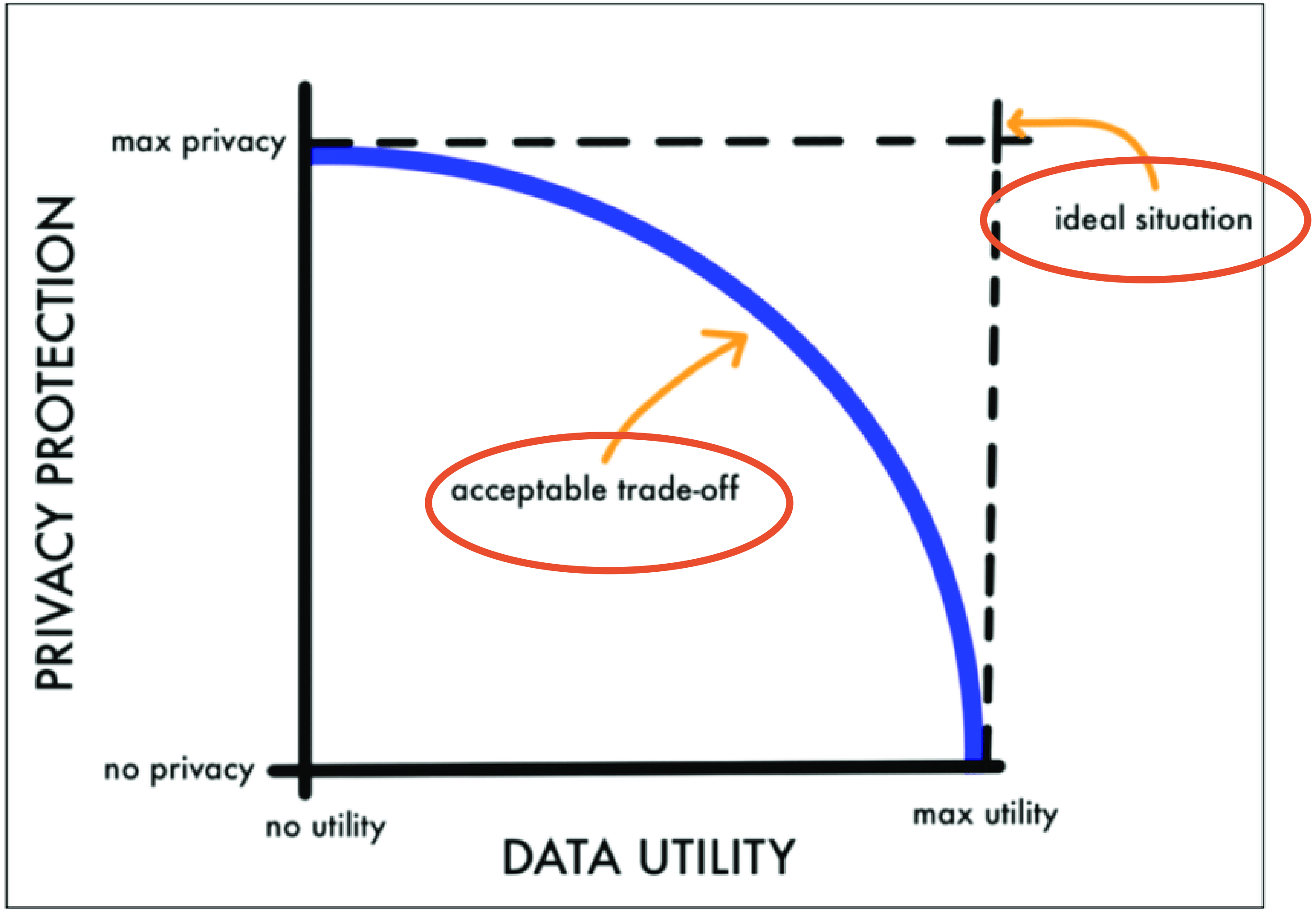
Can be shared, maintaining **privacy**

Before we start, some use cases

- **Erasing old data**
  - We must comply with "The right to be forgotten“
- **Innovative Testing Environments**
  - Simulate data in **external environments**, testing new tools and strategies.

- **Migrating from an on-prem system**
  - Before moving to the cloud show me the benefits on my data
- **Compliance with data minimization principles**
  - Insurers are bound to storing only the necessary amount of customer data
- **Third party data sharing and collaboration**
  - Allow external parties to work with datasets without access to personal data

The tradeoff: utility vs. privacy

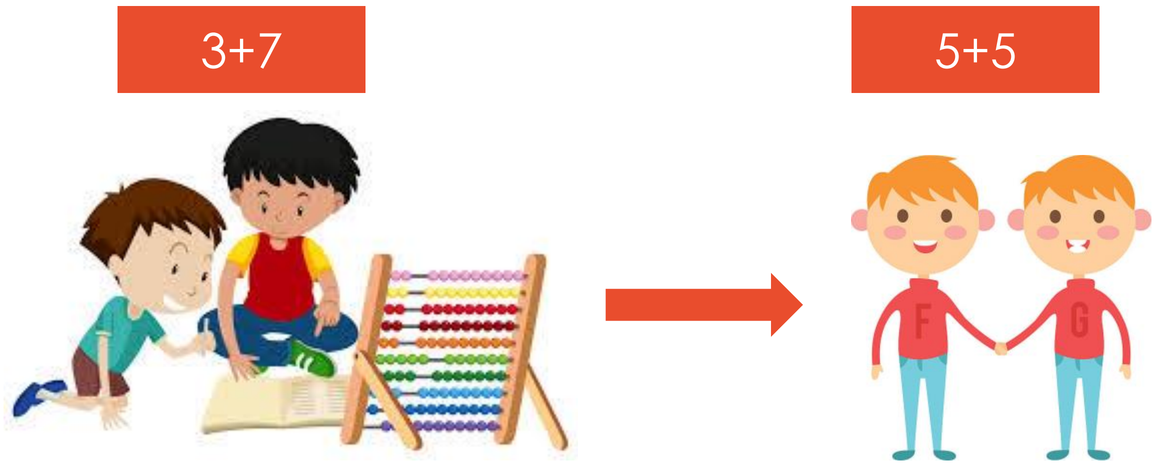


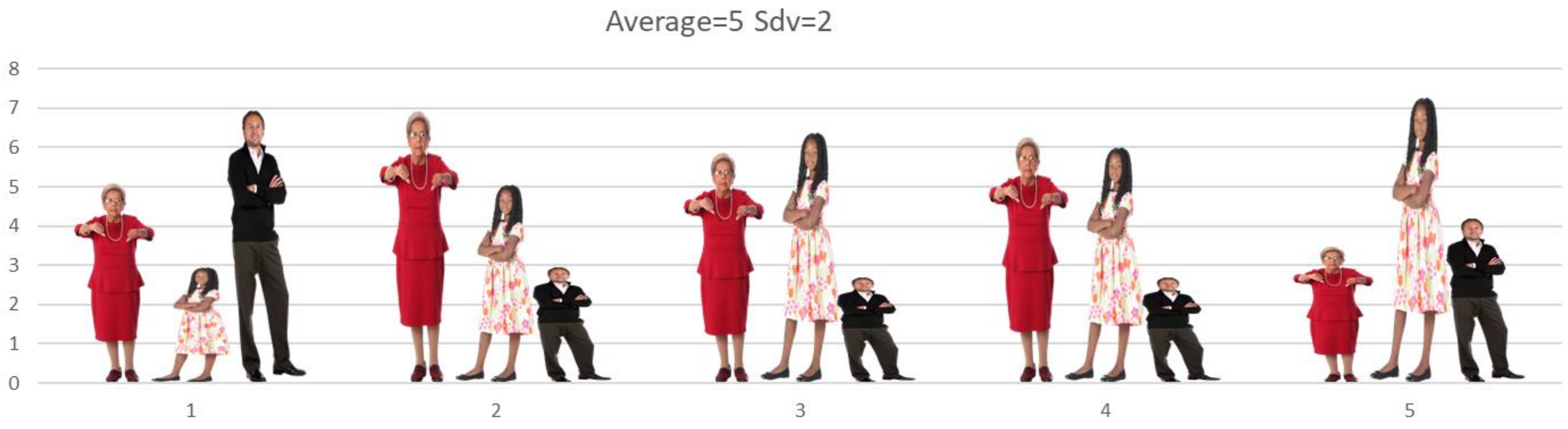
Synthetic data generation principles (1)

The difficult part is to do this accurately in multidimensions

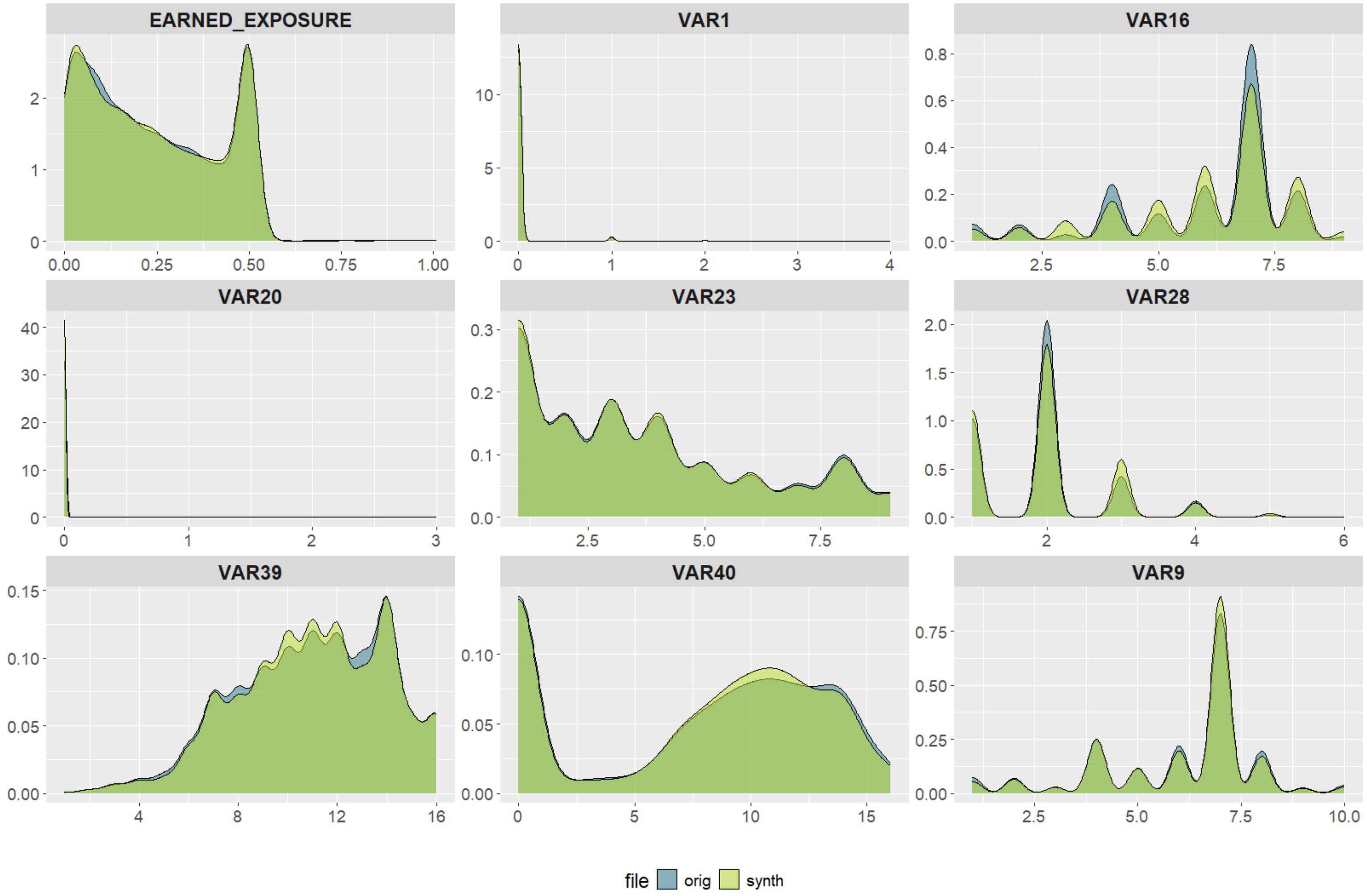
Adapted from: Spot the difference: comparing results of analyses from real patient data and synthetic derivatives (2020). Foraker, R. E. et al, JAMIA open, 3, 557

Synthetic data generation principles (2)





Fidelity of synthetic data (1)



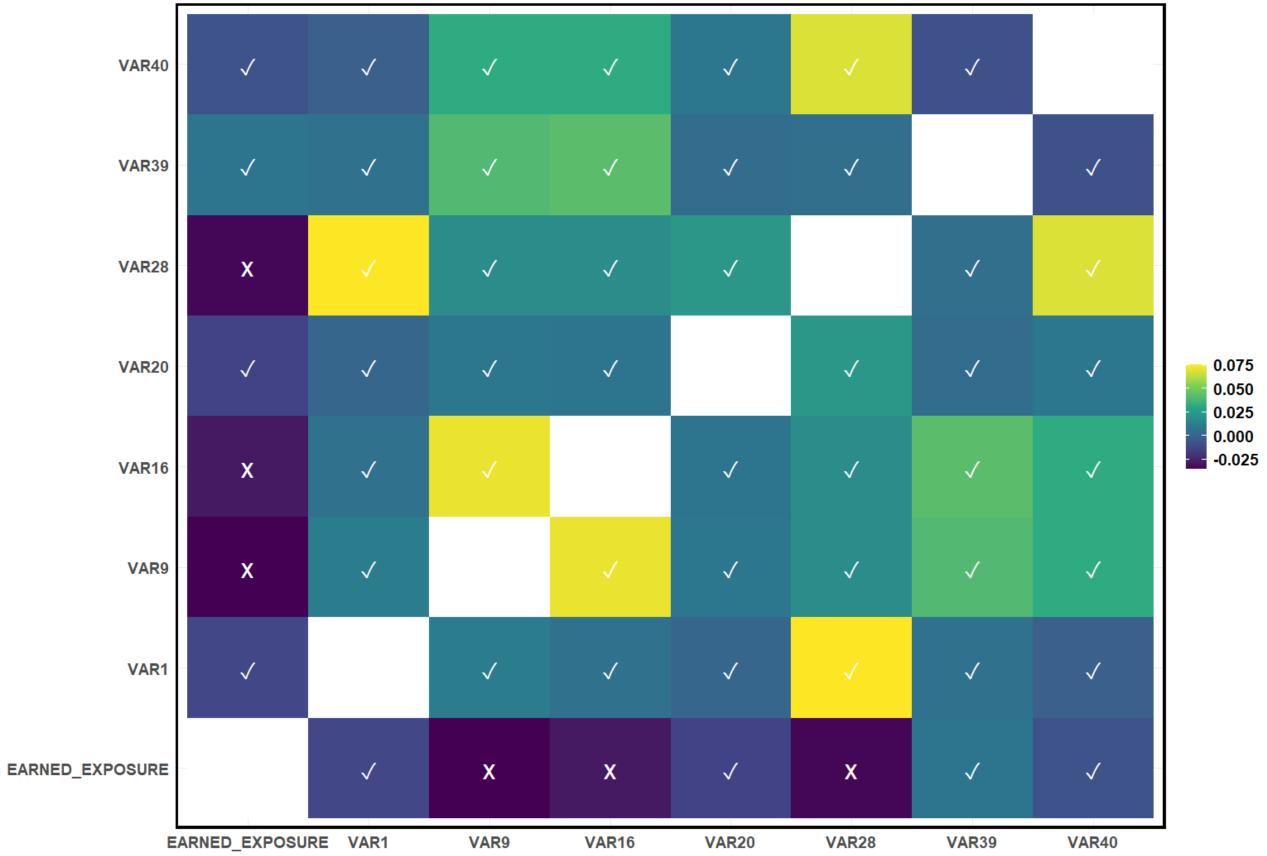
Zamstein, N. (2025). Enhancing Actuarial Ratemaking with Synthetic Data for Privacy Preservation. *CAS E-Forum Quarter 1 (May)*.

Fidelity of synthetic data (2)

Difference in Spearman correlation between original and synthetic data

✓ Statistically insignificant difference

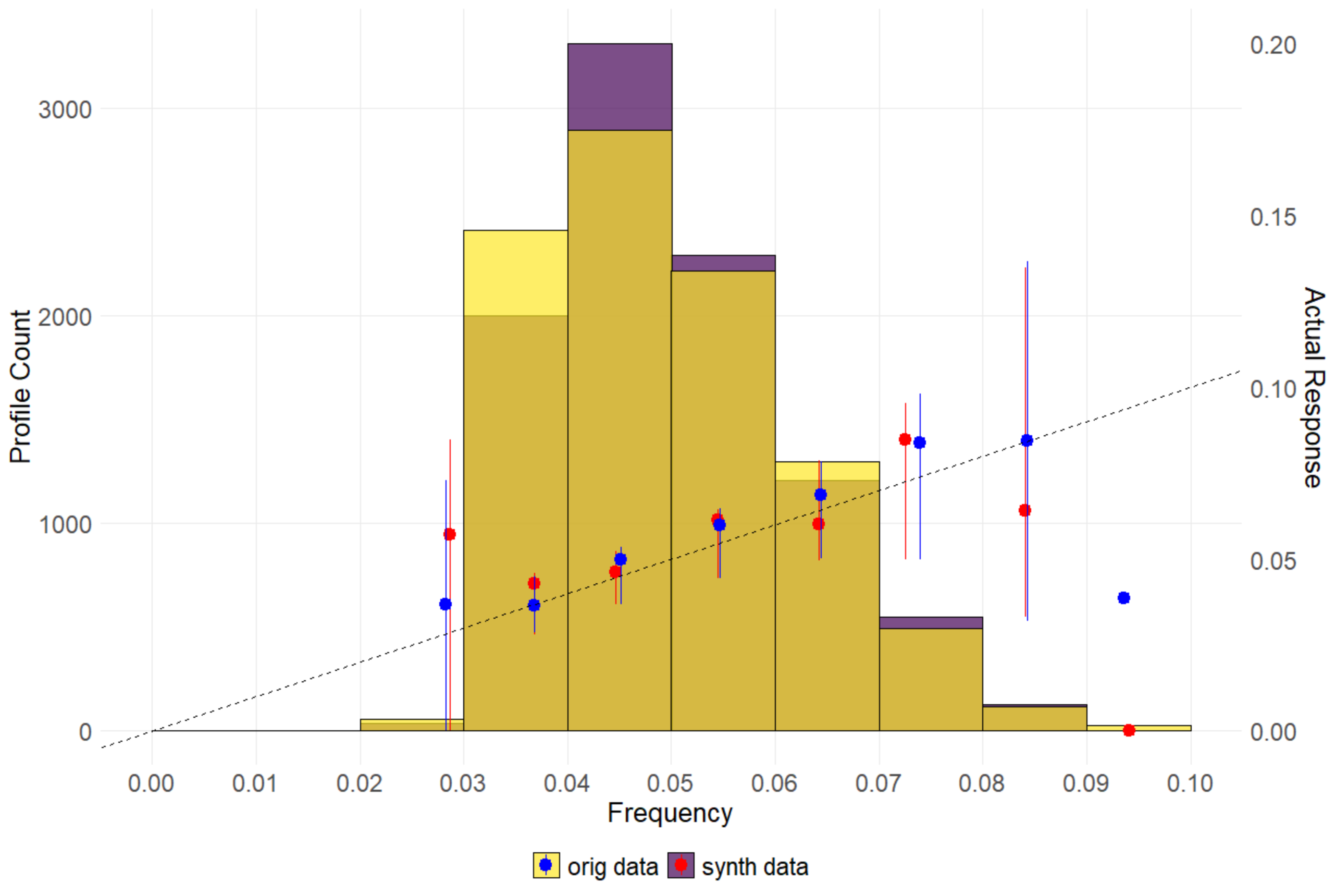
✗ Statistically significant difference



Zamstein, N. (2025). Enhancing Actuarial Ratemaking with Synthetic Data for Privacy Preservation. *CAS E-Forum Quarter 1 (May).*

Fidelity of synthetic data (3)

Frequency models trained on synthetic and original data



Zamstein, N. (2025). Enhancing Actuarial Ratemaking with Synthetic Data for Privacy Preservation. *CAS E-Forum Quarter 1 (May)*.

Privacy of synthetic data (1)

Censoring:

- Generalization to ensure large enough groups of individuals

EARNED_EXPOSURE	ULTIMATE_CLAIM_COUNT	VAR4	VAR27
0.0850	0.0000000	21	BE
0.4959	0.0000000	26	BJ
0.5041	0.0000000	25	BR
0.1945	0.0000000	2	AI
0.0493	0.0000000	27	BF
0.1617	0.0000000	23	AO
0.9149	0.0000000	19	AJ
0.0001	0.0000000	19	AO
0.1644	0.0000000	24	BS
0.3562	0.0000000	29	AK
0.2630	1.0012617	24	BI
0.0548	0.0000000	20	BL
0.4083	0.0000000	20	AD
0.4000	0.0000000	16	BO
0.4082	0.0000000	21	BR
0.3425	0.0000000	21	AJ
0.1617	0.0000000	27	AZ
0.0001	0.0000000	17	BR

EARNED_EXPOSURE	ULTIMATE_CLAIM_COUNT	VAR4	VAR27
0.4822	0.0000000	19	BS
0.4329	0.0000000	19	AI
0.0521	0.0000000	23	BT
0.5041	0.0000000	21	censored
0.3315	0.0000000	18	AJ
0.1699	1.0004295	18	censored
0.0001	0.0000000	16	AJ
0.4959	0.0000000	14	BR
0.3698	0.0000000	26	AJ
0.1370	0.0000000	15	BM
0.1315	0.0000000	19	AJ
0.0658	0.0000000	20	AJ
0.0055	0.0000000	21	censored
0.5041	0.0000000	15	censored
0.1945	0.0000000	27	BR
0.4082	0.0000000	23	censored
0.5014	0.0000000	16	censored
0.4904	0.0000000	20	BF

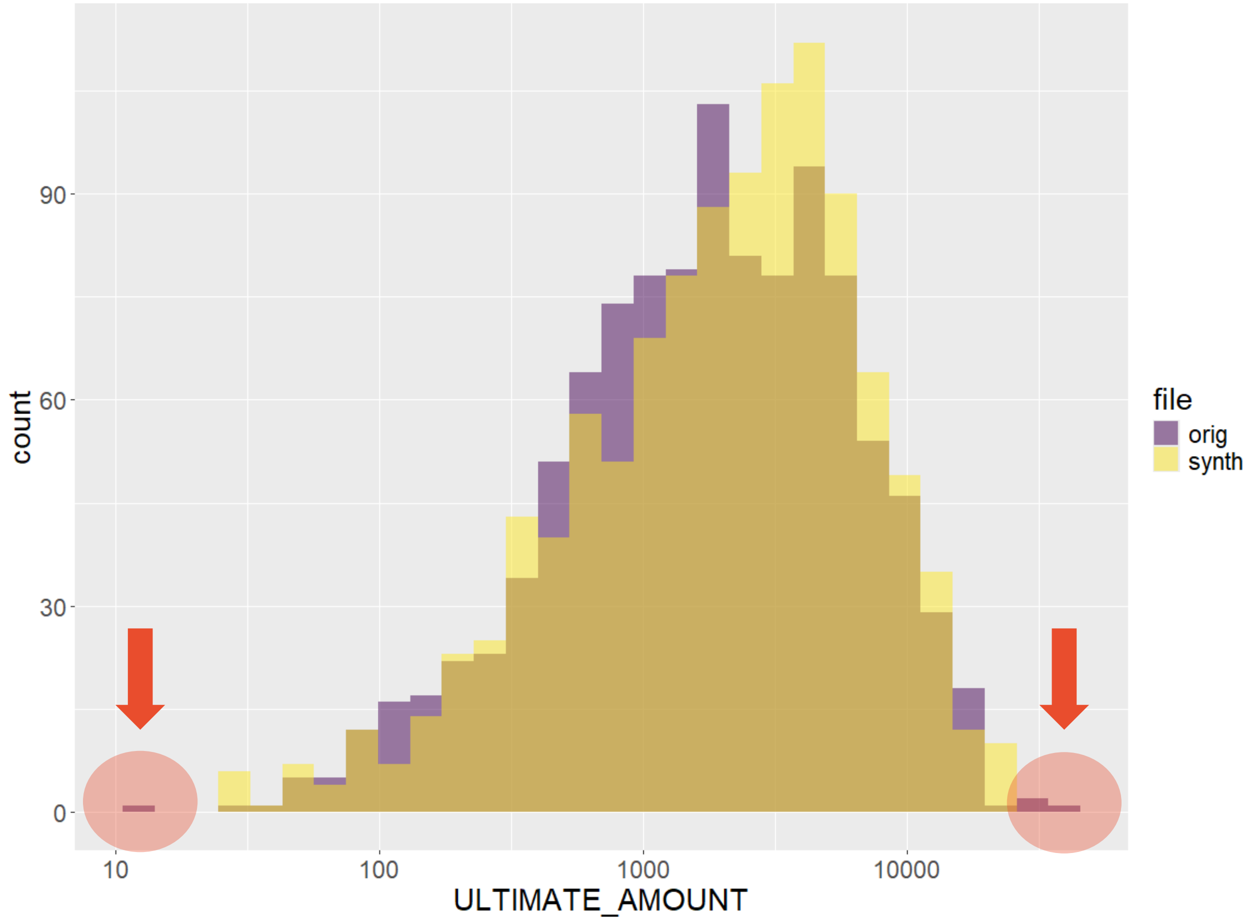
censored

censored

earnix

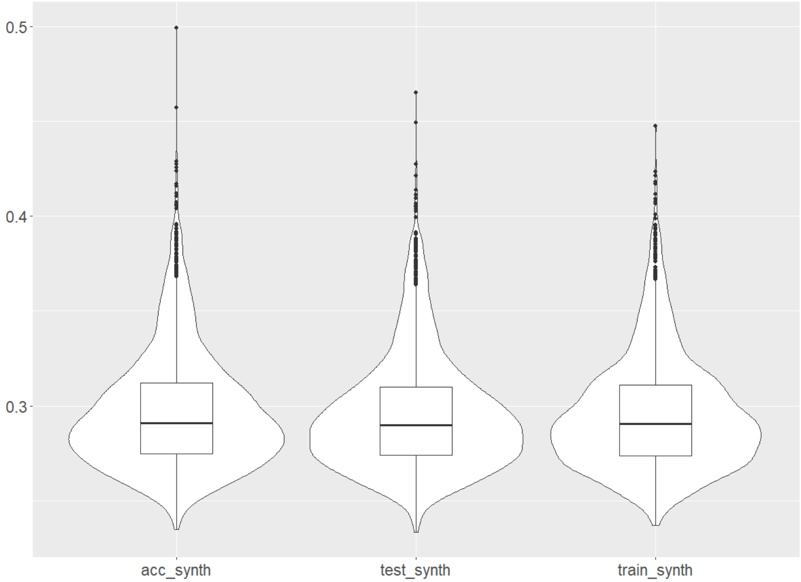
Privacy of synthetic data (2)

Synthetic data should trim **extreme outliers**, such that the distributional parameters are not affected



earnix

Privacy of synthetic data (3)



The distributions of distances are the same, preventing membership inference

Summary and conclusions

- Synthetic data provides a layer of protection against re-identification
- Maintains individual level details for predictive modelling
- Preserves privacy and robust against inference attacks
- Any acceptable solution must balance between the two

Thank You

Noa.Zamstein@earnix.com

This presentation reflects NZ’s personal views and does not necessarily represent the views of Earnix Ltd.

earnix