# Testing exchangeability between Real and Synthetic data

Helena Löfström, Lars Carlsson and Ernst Ahlberg

COPA 2024, Milano

# Outline

Mölnlycke®

- Introduction
- Methodology
- Results
- Conclusion
- Future work

Mölnlycke

# Why synthetic data?

**Data sources**

Sensitive data

**tomtA**

1. **Learns** patterns and relationships of your data
2. **Generates** a safe precises equivalent of your sensitive data by applying computational differential privacy and AI
3. **Measure and manage** the privacy and utility of the generated data

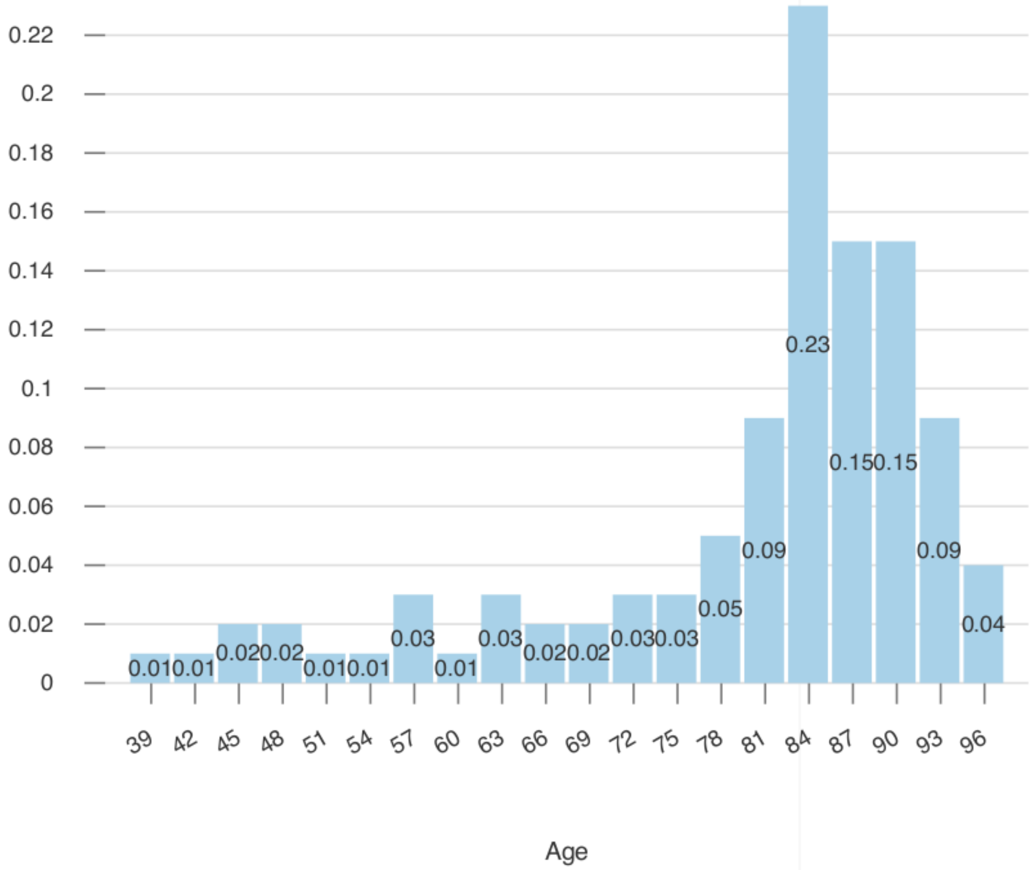
Generated safe and precise data

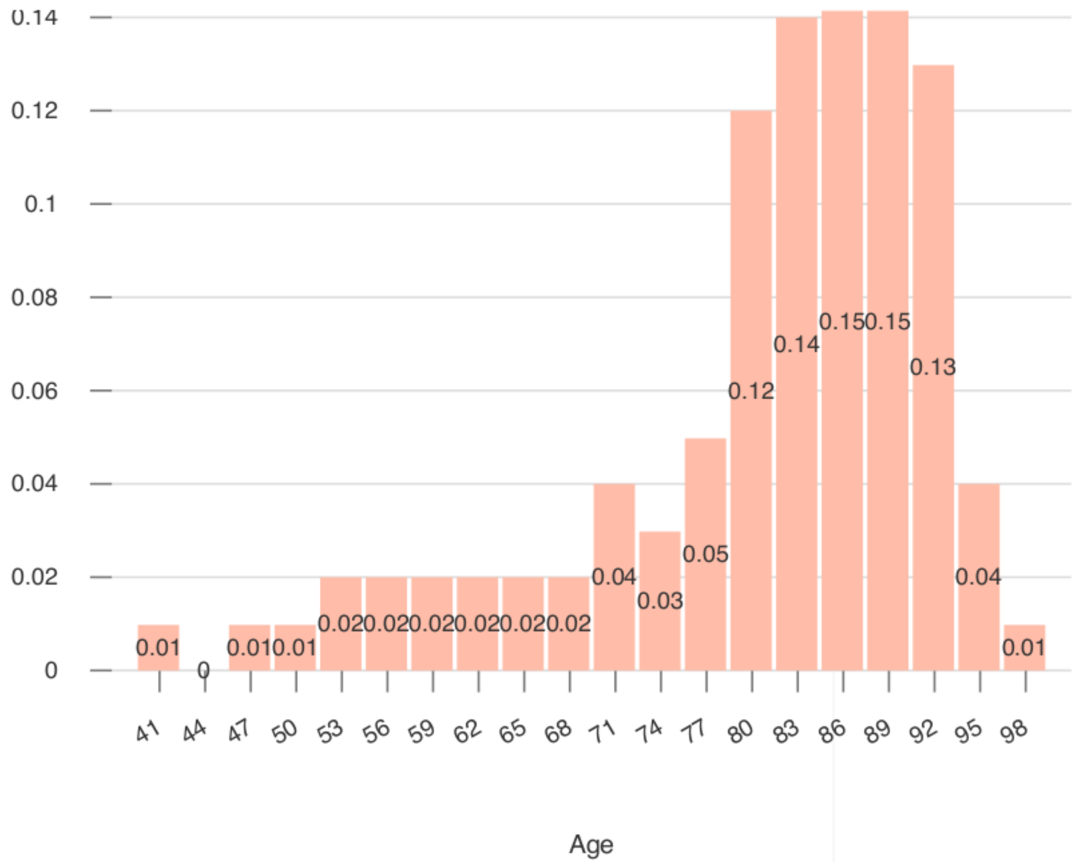
**Data sources**

Source tomta.ai

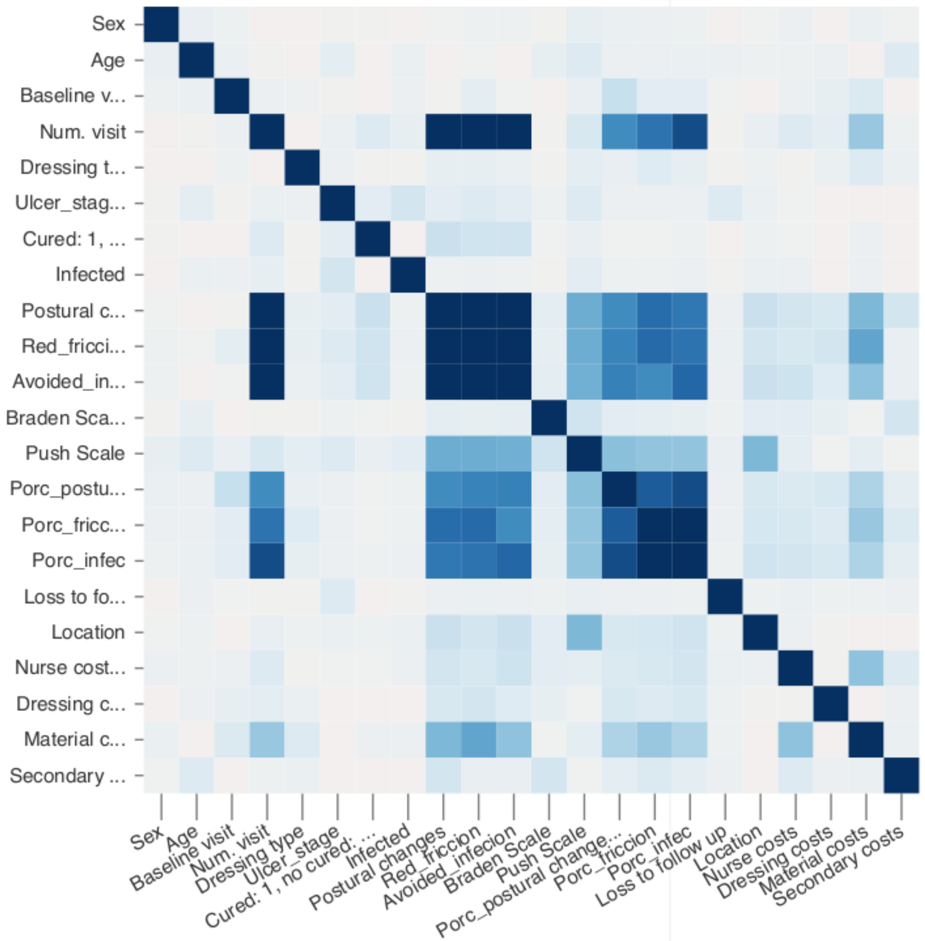
# The Challenge – can synthetic data be taken as real data?

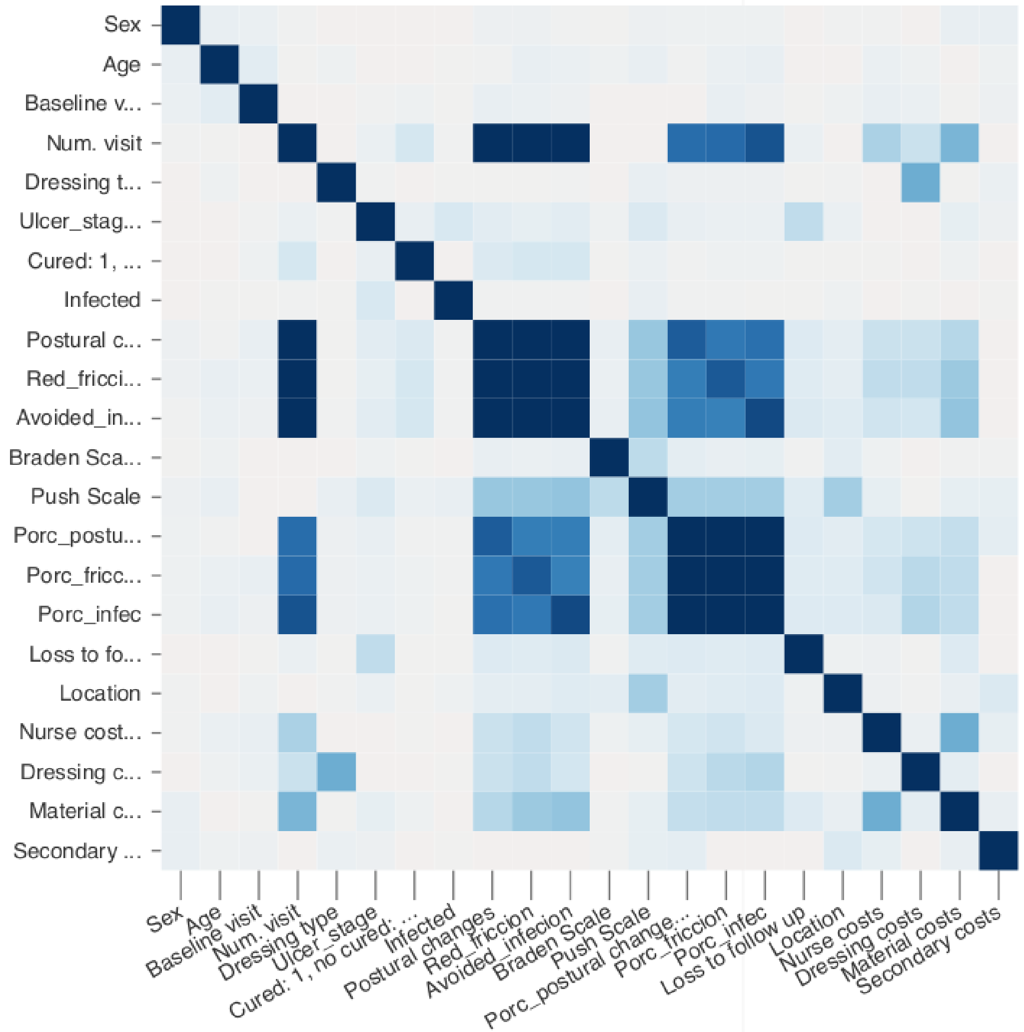
# Comparing distributions



# Comparing correlations



# Does the exchangeability assumption hold?

# Methodology - test

1. Data generation and preprocessing
2. Application of the conformal transducer
3. Application of the martingale test for exchangeability
4. Evaluation of the datasets

**Algorithm 1:** Simple Jumper $(p_1, p_2, \dots, p_n) \mapsto (S_1, S_2, \dots, S_n)$

```
C_{-1} := C_0 := C_1 := 1/3
C := 1
J = 0.01
for k = 1, 2, ..., n do
    for ε ∈ {−1, 0, 1} do
        C_ε := (1 − J)C_ε + (J/3)C
    end
    for ε ∈ {−1, 0, 1} do
        C_ε := C_ε(1 + ε(p_k − 0.5))
    end
    S_k := C := C_{-1} + C_0 + C_1
end
```

# Methodology - data and evaluation

- Real on Real (R-R): The conformal transducer is trained and tested on the real dataset to verify that our test works.
- Synthetic on Synthetic (S-S): The conformal transducer is trained and tested on the synthetic dataset, similar to the R-R case.
- Real on Synthetic (R-S): The conformal transducer, trained and calibrated on the real dataset, generates p-values for the synthetic dataset.
- Synthetic on Real (S-R): The conformal transducer, trained and calibrated on the synthetic dataset, generates p-values for the real dataset.

| Dataset Name | Instances | Features | Numeric Features | Source |
|---|---|---|---|---|
| adult-sdv | 32561 | 15 | 6 | (Patki et al., 2016) |
| credit-g | 1000 | 21 | 7 | (Vanschoren et al., 2013) |
| spambase | 4601 | 58 | 58 | (Vanschoren et al., 2013) |
| qsar-biodeg | 1055 | 42 | 42 | (Vanschoren et al., 2013) |
| adult | 48842 | 15 | 6 | (Vanschoren et al., 2013) |
| RWI | 88588 | 7 | 7 | (Vanschoren et al., 2013) |

Table 1: Additional information on the datasets used in this study.

Mölnlycke

# Results

| Dataset | Mean | Std | Min | Max |
|---|---|---|---|---|
| adult-sdv | -63.61 | 6.58 | -75.52 | -45.10 |
| credit-g | -1.40 | 1.87 | -3.49 | 5.05 |
| spambase | -8.48 | 2.56 | -12.51 | -1.85 |
| qsar-biodeg | -2.01 | 1.84 | -3.87 | 7.06 |
| adult | -96.63 | 7.09 | -109.32 | -79.45 |
| RWI | -174.65 | 7.36 | -192.76 | -159.12 |

Table 2: Some metrics on the distribution of the martingale values for the 50 runs in the R-R case.

| Dataset | Mean | Std | Min | Max |
|---|---|---|---|---|
| adult-sdv | -63.34 | 4.93 | -73.59 | -48.57 |
| credit-g | -1.81 | 1.82 | -3.49 | 6.04 |
| spambase | -8.73 | 2.49 | -12.45 | -1.68 |
| qsar-biodeg | -1.45 | 3.04 | -3.73 | 16.17 |
| adult | -96.58 | 5.53 | -106.71 | -82.01 |
| RWI | -174.72 | 10.01 | -195.17 | -156.26 |

Table 3: Some metrics on the distribution of the martingale values for the 50 runs in the S-S case.

# Results cont'd

| Dataset | Mean | Std | Min | Max |
|---|---|---|---|---|
| adult-sdv | 76.50 | 35.83 | 8.83 | 169.04 |
| credit-g | 57.72 | 8.62 | 36.61 | 72.18 |
| spambase | 519.12 | 26.78 | 440.35 | 579.82 |
| qsar-biodeg | 127.23 | 6.70 | 111.12 | 140.12 |
| adult | -8.35 | 26.97 | -62.07 | 60.97 |
| RWI | 709.78 | 0.00 | 709.78 | 709.78 |

Table 4: Some metrics on the distribution of the martingale values for the 50 runs in the R-S case.

| Dataset | Mean | Std | Min | Max |
|---|---|---|---|---|
| adult-sdv | -65.53 | 5.26 | -73.60 | -52.25 |
| credit-g | 2.93 | 6.28 | -3.25 | 26.78 |
| spambase | 29.82 | 14.30 | 1.04 | 57.76 |
| qsar-biodeg | 13.87 | 13.68 | -3.46 | 57.60 |
| adult | -79.51 | 15.20 | -99.85 | -37.54 |
| RWI | 708.55 | 5.21 | 684.25 | 709.78 |

Table 5: Some metrics on the distribution of the martingale values for the 50 runs in the S-R case.

# Conclusion

The synthetic data does not appear to come from the same distribution as the real data.

# Future work/considerations

- Extend current scope to regression
- Is it possible to initiate adversarial attacks using synthetic data generation?
- How can we be sure that the underlying relationship remains?