

# Copula-based synthetic data generation in medicine

Antoine Faul

Final year PhD student under the supervision of Prof. David Ginsbourger  
and Prof. Ben Spycher

Institute of Mathematical Statistics and Actuarial Sciences  
University of Bern, Switzerland

STRADA workshop, 15th April, 2026

## Synthetic medical data is used to:

- protect privacy of patients.
- perform data augmentation.
- make patient data widely available for research.

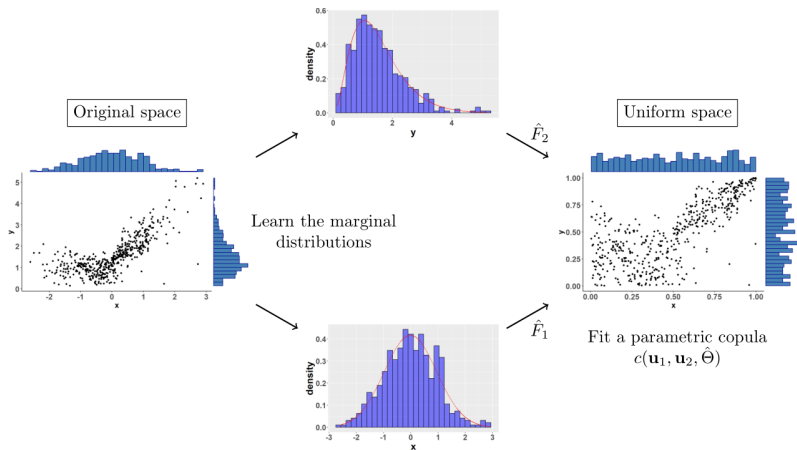
## Synthetic medical data is used to:

- protect privacy of patients.
- perform data augmentation.
- make patient data widely available for research.

## Capture the underlying data distribution by:

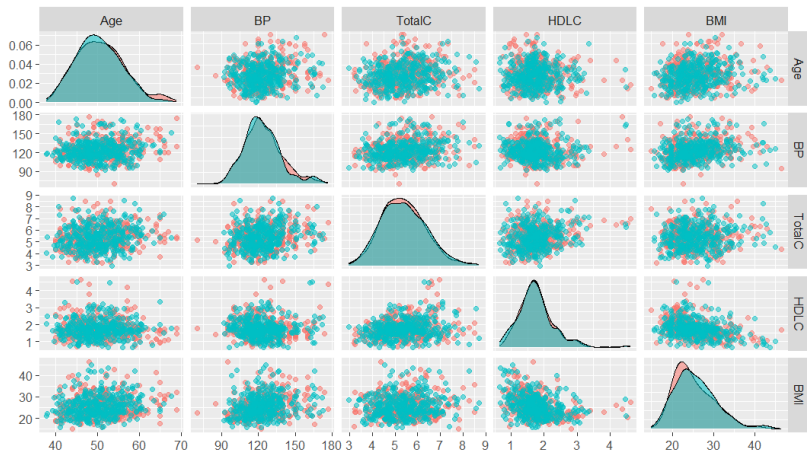
- preserving the main statistical properties of the data.
- not revealing patient information at an individual level.

# Copula-based synthetic data generation



- **Fidelity:** Measures how closely the synthetic dataset's distribution matches that of the real data.
- **Utility:** Assessed by its effectiveness in practical applications, such as downstream regression or classification tasks.
- **Privacy:** Ensures no duplication of individual records, safeguards sensitive information, and protects against potential attacks.

## Fidelity evaluation: Visualization



**Figure 1:** Synthetic data generated with the Gaussian Mixture Copula model (blue) vs original data (red) from the CIMBOLIC dataset

Train on Synthetic Test on Real (TSTR) framework.

| Generator Type | CRPS↓                | wCRPS (1e-2) ↓     | Log-Score ↓          |
|----------------|----------------------|--------------------|----------------------|
| Real           | 0.340 ± 0.031        | 5.03 ± 2.16        | 0.781 ± 0.074        |
| GMCM           | <b>0.345 ± 0.035</b> | <b>5.03 ± 2.41</b> | <b>0.804 ± 0.085</b> |
| Vines          | <b>0.345 ± 0.035</b> | <b>5.04 ± 2.47</b> | <b>0.801 ± 0.089</b> |
| ARF            | 0.348 ± 0.036        | 5.14 ± 2.59        | 0.822 ± 0.083        |
| GC             | 0.358 ± 0.040        | 5.23 ± 2.71        | 0.926 ± 0.169        |
| MARG           | 0.395 ± 0.044        | 5.97 ± 3.16        | 0.937 ± 0.093        |
| CTGAN          | 0.550 ± 0.042        | 8.17 ± 2.60        | 1.26 ± 0.057         |

**Table 1:** Performances of SCORE-2 risk prediction with systematic missing values by using different synthetic dataset in the TSTR framework. All results are reported as the mean ± std, averaged over 100 synthetic dataset and 20 train-test splits.

# THANK YOU !

ANY QUESTIONS ?

| Generator Type ↓ | MMD (e-3) ↓ | JSD (e-3) ↓  | IKL (e-1) ↑ | DCR ↑        | Custom DCR ↑ |
|------------------|-------------|--------------|-------------|--------------|--------------|
| GC               | 31.7 ± 0.6  | 13.90 ± 0.58 | 7.71 ± 0.14 | 0.28 ± 0.006 | 0.11 ± 0.019 |
| CTGAN            | 44.6 ± 1.4  | 31.60 ± 0.59 | 7.98 ± 0.05 | 0.59 ± 0.013 | 0.48 ± 0.018 |
| GMCM             | 3.8 ± 0.4   | 3.34 ± 0.45  | 9.91 ± 0.04 | 0.31 ± 0.007 | 0.14 ± 0.005 |
| MARG             | 7.4 ± 0.6   | 3.39 ± 0.44  | 9.91 ± 0.04 | 0.34 ± 0.009 | 0.33 ± 0.025 |
| ARF              | 3.1 ± 0.4   | 3.93 ± 0.44  | 9.85 ± 0.08 | 0.30 ± 0.008 | 0.11 ± 0.005 |
| Vines            | 3.6 ± 0.3   | 3.34 ± 0.44  | 9.91 ± 0.04 | 0.31 ± 0.008 | 0.14 ± 0.018 |

**Table 2:** Experiment Results Over the Cimboric dataset. All results are reported as the mean ( $\pm$ ) std, averaged over 100 synthetic datasets.