FACULTY OF MEDICINE AND HEALTH SCIENCES

SYNDARA

The Real Deal Behind the Artificial Appeal: Inferential Utility of Tabular Synthetic Data

Alexander Decruyenaere*, Heidelinde Dehaene*, Paloma Rabaey, Christiaan Polet, Johan Decruyenaere, Stijn Vansteelandt, Thomas Demeester

* Joint first authors

Background

Alongside great opportunities, great precaution should be taken regarding the possible sensitive nature of medical data and related privacy concerns.

Synthetic data are artificial data that mimic the original data in terms of statistical properties. As such, synthetic data might be able to replace the original data in statistical analysis, while **preserving the privacy** of the individual members of the original dataset.

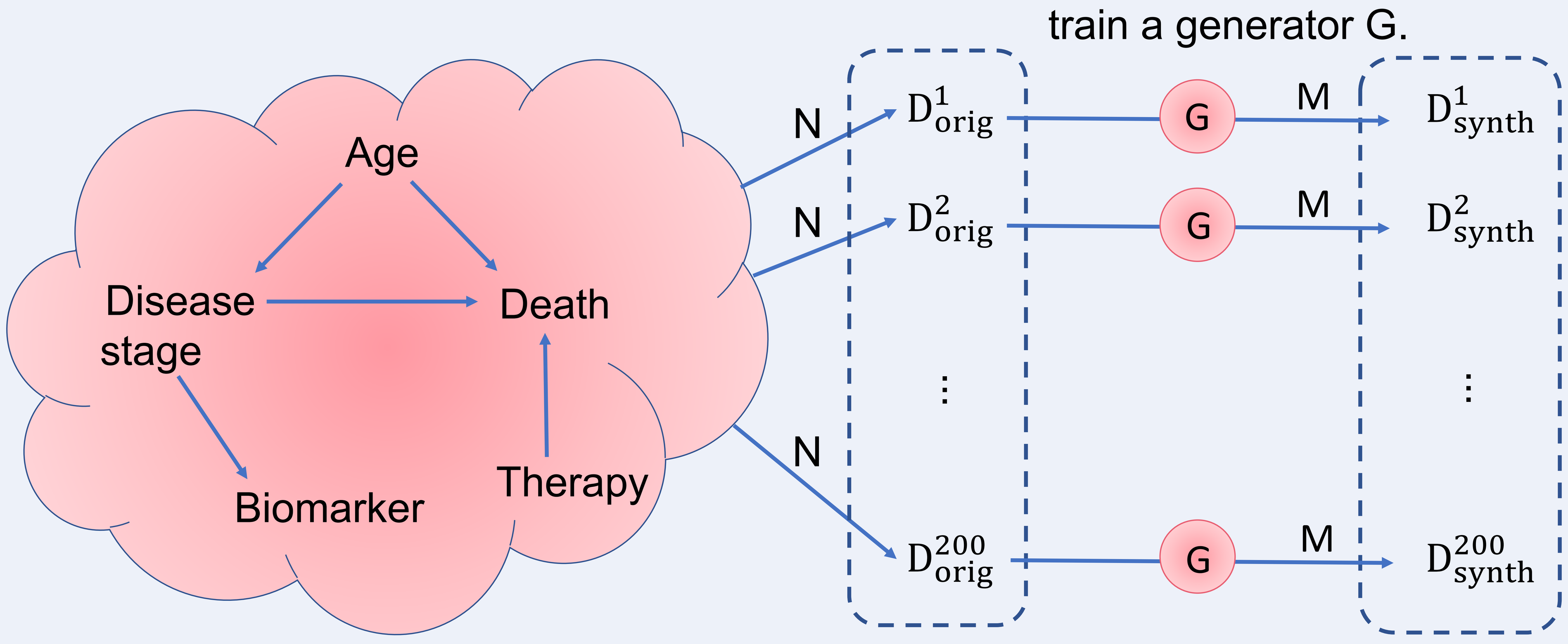
Problem statement

Can a synthetic sample be used to obtain valid estimates for a population parameter and to test hypotheses? We map two possible **pitfalls that may compromise this inferential utility of synthetic data**:

1. **Extra uncertainty** should be acknowledged since the distribution learned by the generative model is an approximation.
2. Statistical inference is typically based on $\sqrt{N}$**-consistency and asymptotic normality**. What is the effect of regularisation bias inherent to deep learning (DL) approaches on the default behaviour of estimators?

Experimental set-up

3. Use each dataset D_{orig} to train a generator G.



4. Generate synthetic dataset D_{synth} with size M using G.

5. Assess bias in point estimation and standard error (SE), and investigate convergence rate of estimators when they are used in D_{synth}.

1. Ground truth data generating process for tabular toy data.

2. Randomly sample original dataset D_{orig} with size N. Repeat 200 times.

Desired properties of an estimator:

Standard error (SE) goes to zero when sample size increases, at rate $1/\sqrt{N}$ and bias at faster rate. True and estimated SE are according.

These properties affect inferential utility, here measured by **type 1 error rate**.

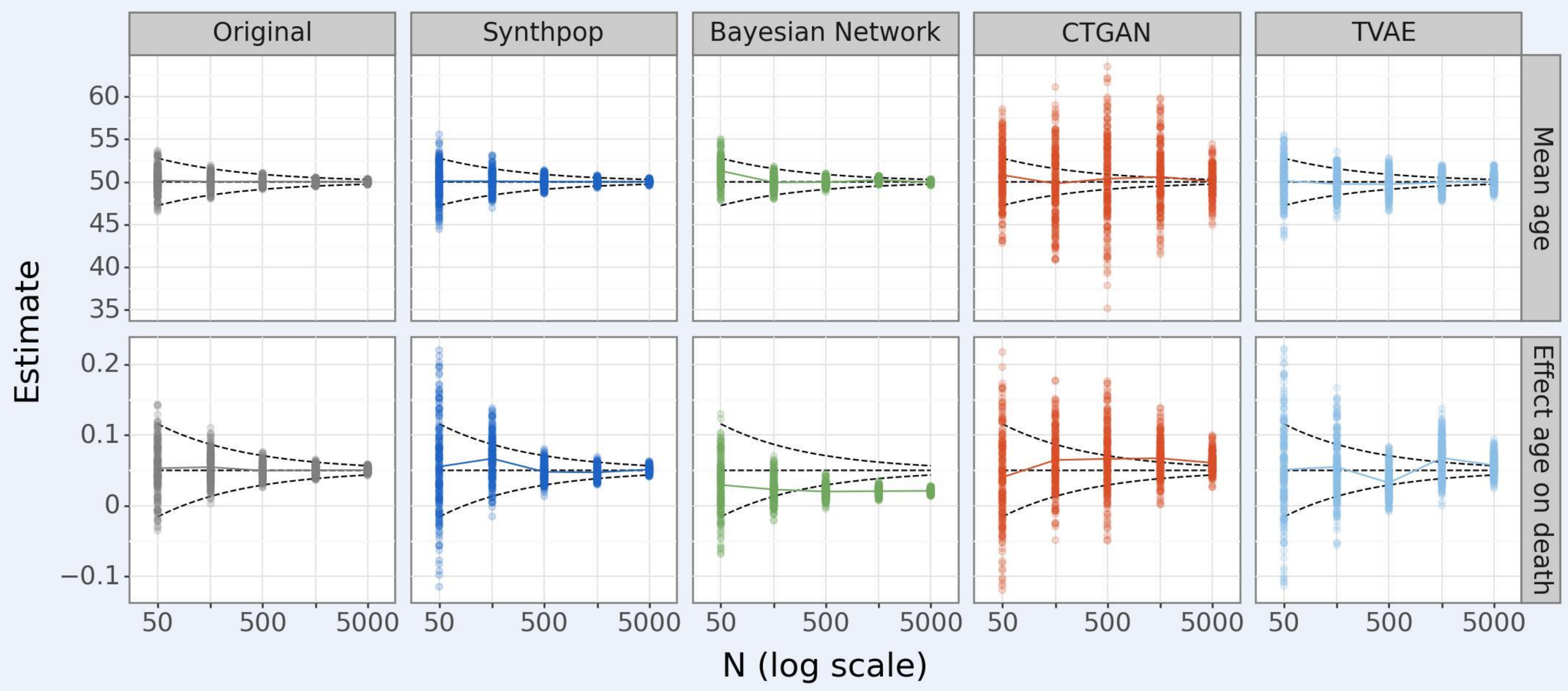
Minimal adaptation estimator SE:

$$\sigma_{\hat{\theta},\text{corrected}} = \sigma_{\hat{\theta},\text{naive}} \sqrt{1 + \frac{M}{N}}$$

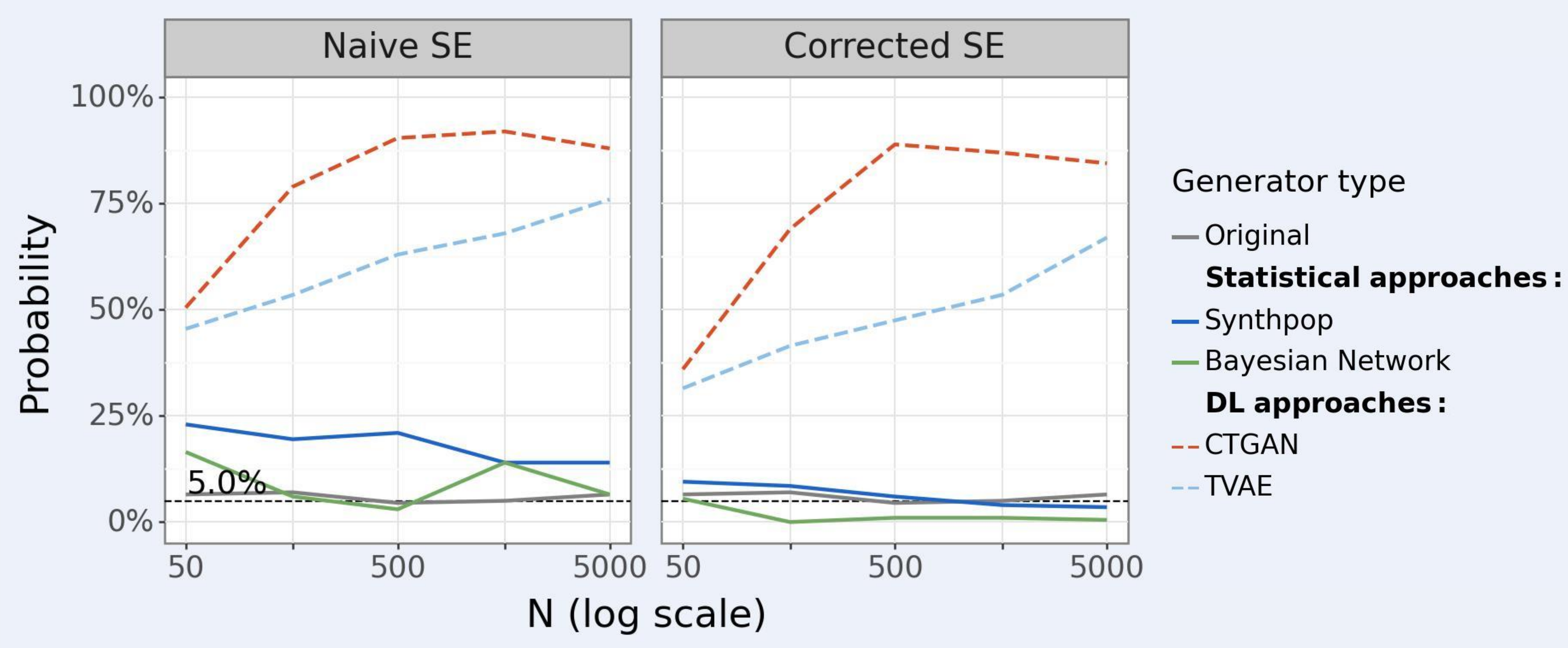
Is this valid for all estimators and generators? What about $\sqrt{N}$-consistency?

Results inferential utility

Finite sample behaviour of estimator



Type 1 error rate for mean age



Main results:

1. Generative model **misspecification** introduces **bias**.
2. True **SEs are larger** for D_{synth} than for D_{orig} and extra variability varies over generative models.
3. Therefore, **naive estimation of SE leads to its underestimation**.
4. **Convergence rate** of the SE of various estimators **differs** between statistical and DL approaches. For the statistical approaches, estimators remain roughly $\sqrt{N}$-consistent. In the **DL approaches**, estimators converge **slower**.
5. **Naive analyses** lead to **inflation** of **type 1 error rate** (compromising inferential utility).
6. **Adaptation** for the SE **controls type 1 error rate only in statistical** and **not in DL** approaches (due to slower-than-$\sqrt{N}$-convergence).

Conclusion:

Before publishing synthetic data, it is essential to develop statistical inference tools for such data.

UZ GENT

UNIVERSITEIT GENT

Alexander.Decruyenaere@ugent.be
Heidelinde.Dehaene@ugent.be
syndara.ai