FACULTY OF MEDICINE AND HEALTH SCIENCES

SYNDARA

Debiasing Synthetic Data Generated by Deep Generative Models

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

* Joint first authors

Background

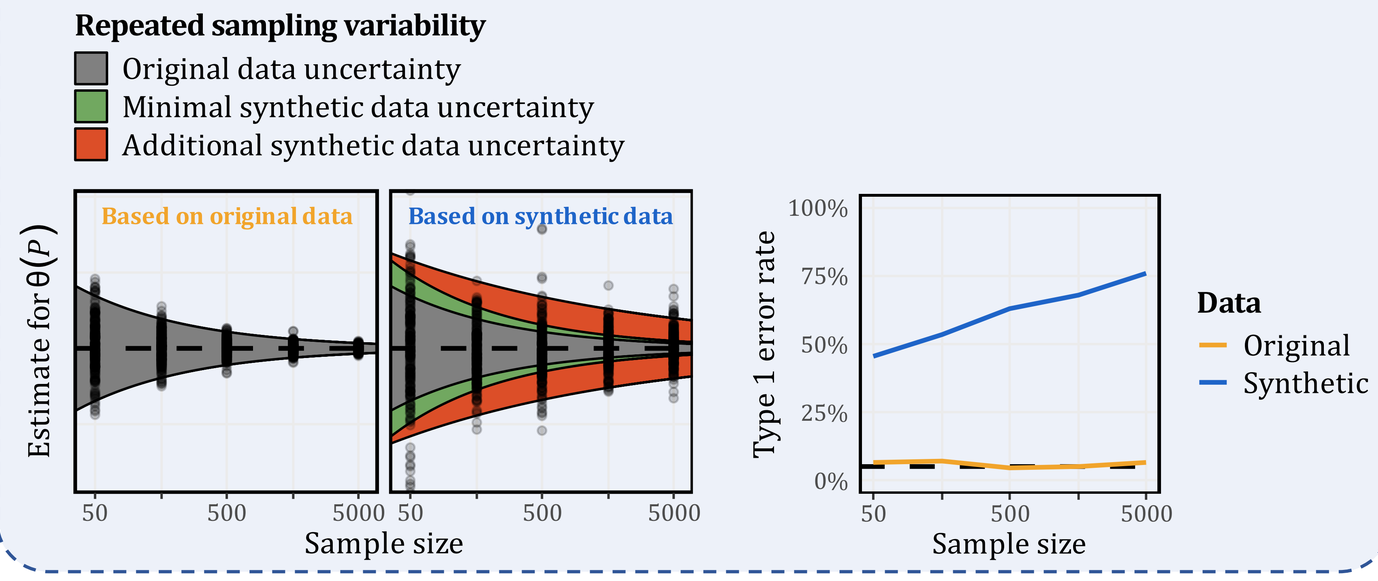
Alongside great opportunities, great precaution should be taken regarding the possible sensitive nature of medical data and related privacy concerns. **Synthetic data** are artificial data that mimic the original data in terms of statistical properties. As such, synthetic data might be able to replace the original data in statistical analysis, while **preserving the privacy** of the individual members of the original dataset.

Problem statement

The use of **deep generative models (DGMs)** for **synthetic data generation** induces bias and imprecision into synthetic data analyses, **inflating the type 1 error rate**. This **compromises their inferential utility** as opposed to original data analysis, even for simple parameters like the population mean [3].

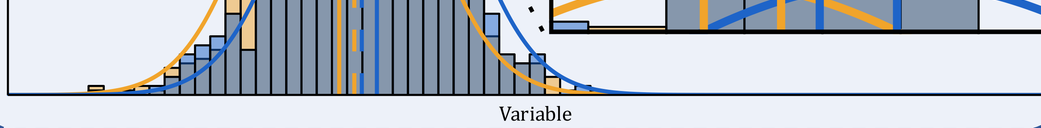
Prior approaches only consider the extra uncertainty arising from a parametric data generation strategy. This is however **insufficient when DGMs are used** to generate synthetic data, as they overlook the effects of regularization bias [3].

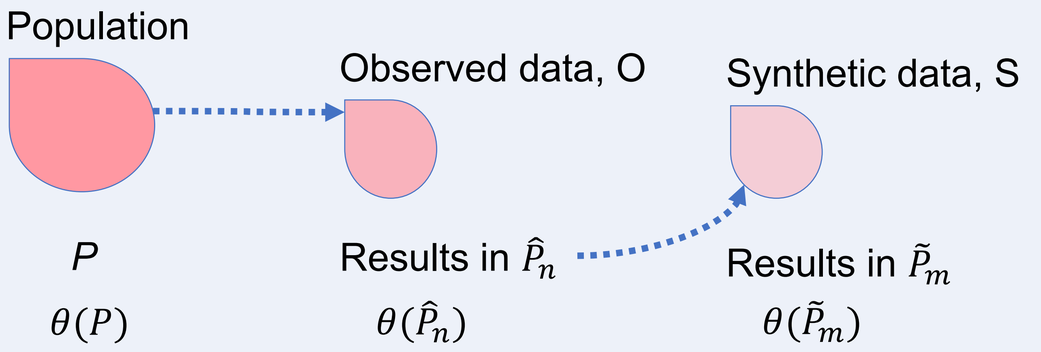
Data-adaptive methods cannot succeed to estimate all features of the data-generating distribution well and are designed to optimize the prediction error instead of the error in the estimator [1,5,2,4]. This leads to **excess variability** and **slow convergence**, which are not addressed in previous methods.



Notation & Set-Up

DAG used to generate original population:





Methodology

We use 2 von Mises expansions to study the **difference** between $\theta(\tilde{P}_m)$ and $\theta(P)$. We show that this reduces to:

$$\theta(\tilde{P}_m) - \theta(P) = \frac{1}{m}\sum_{i=1}^{m} \emptyset(S_i, \hat{P}_n) - \frac{1}{m}\sum_{i=1}^{m} \emptyset(S_i, \tilde{P}_m) + o_p(m^{-1/2})$$
$$+ \frac{1}{n}\sum_{i=1}^{n} \emptyset(O_i, P) - \frac{1}{n}\sum_{i=1}^{n} \emptyset(O_i, \hat{P}_n) + o_p(n^{-1/2})$$

where $\emptyset(., P)$ is the efficient influence curve (EIC) or the functional derivative of $\theta(P)$. We identify 2 problematic **bias** terms:

$$-\frac{1}{m}\sum_{i=1}^{m} \emptyset(S_i, \tilde{P}_m)$$

Origin: the use of data-adaptive estimates.

Solution to make this zero: analyse synthetic data with **debiased estimators**, derived from the EIC [5].

$$-\frac{1}{n}\sum_{i=1}^{n} \emptyset(O_i, \hat{P}_n)$$

Origin: the use of a DGM to obtain $\hat{P}_n$.

Solution to make this zero: **shift the variable of interest** in the synthetic data. Can be done for all pathwise differentiable parameters, but the exact implementation depends on the EIC.

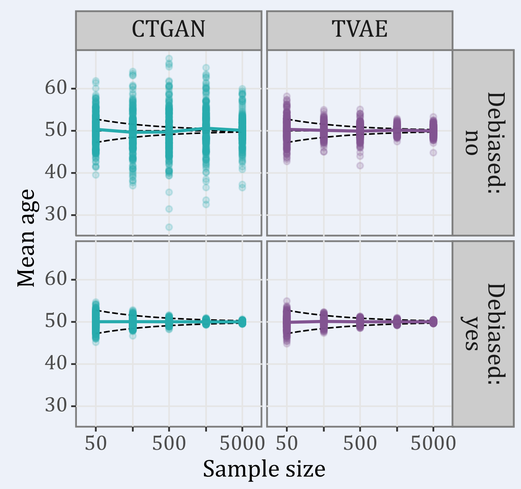
Example for the population mean with $\emptyset(O, P) = O - \theta(P)$

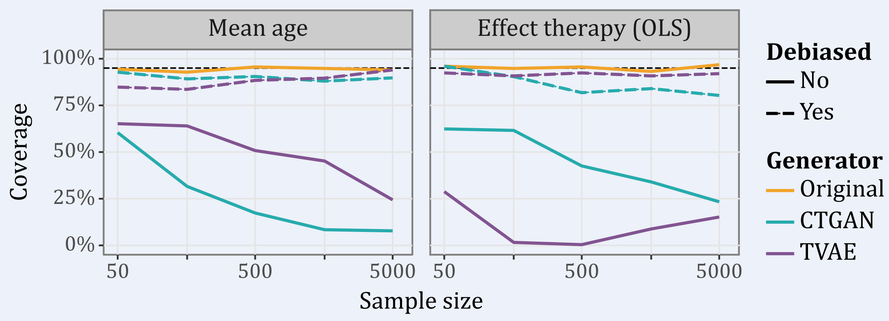
$$\theta(\tilde{P}_m) = \frac{1}{n}\sum_{i=1}^{m} S_i$$

Add $\bar{O} - \theta(\hat{P}_n)$ to S_i where $\theta(\hat{P}_n)$ is approximated based on the DGM.

Results

1. SE of estimators converge at approximately root-*n* rates.
2. Results in **empirical coverage levels** for the 95% CI that in most cases **approximate the nominal level**.





References

[1] Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Baltimore Johns Hopkins University Press.

[2] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, (21(1):C1-C68.

[3] Decruyenaere, A., Dehaene, H., Rabaey, P., Polet, C., Decruyenaere, J., Vansteelandt, S., and Demeester, T. (2024). The real deal behind the artificial appeal: Inferential utility of tabular synthetic data. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

[4] Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *American Statistician*, 76(3):292–304.

[5] van der Laan, M. J. and Rose, S. (2011). *Targeted Learning.* Springer Series in Statistics. Springer New York, New York, NY.

UZ GENT

UNIVERSITEIT GENT

Alexander.Decruyenaere@ugent.be

Heidelinde.Dehaene@ugent.be