

The Real Deal Behind the Artificial Appeal

Inferential Utility of Tabular Synthetic Data

Decruyenaere A*, Dehaene H*, Rabaey P, et al.

poster #406

UAI conference, July 18th, 2024

Naive statistical inference from synthetic data cannot be trusted

Original data



Age	Gender	Allergy	NumOfMedication	HDL	Colestero.	LDL	Glucose	Urea	ICD
27	M	FALSE	13	83.61	154	80.71	190.69	4.60	A18
49	M	FALSE	5	68.10	212	160.85	220.64	2.70	A18
47	M	FALSE	1	88.38	228	159.39	220.24	1.50	A02
62	F	FALSE	1	62.77	286	230.15	236.20	4.50	A18
43	F	FALSE	4	83.44	161	86.49	193.69	4.90	A18
59	M	FALSE	1	95.50	200	116.73	206.71	6.15	A18
60	F	FALSE	14	90.84	250	168.12	222.56	4.44	A18
64	F	FALSE	6	72.08	176	107.29	203.05	5.08	A18
59	M	FALSE	4	61.40	177	126.37	210.16	1.37	A02
58	M	FALSE	13	84.43	181	101.86	200.80	1.72	A04
51	F	TRUE	14	53.62	226	185.53	226.84	5.43	A18
56	M	FALSE	3	97.95	197	115.41	206.22	5.81	A18

Naive statistical inference from synthetic data cannot be trusted

Original data



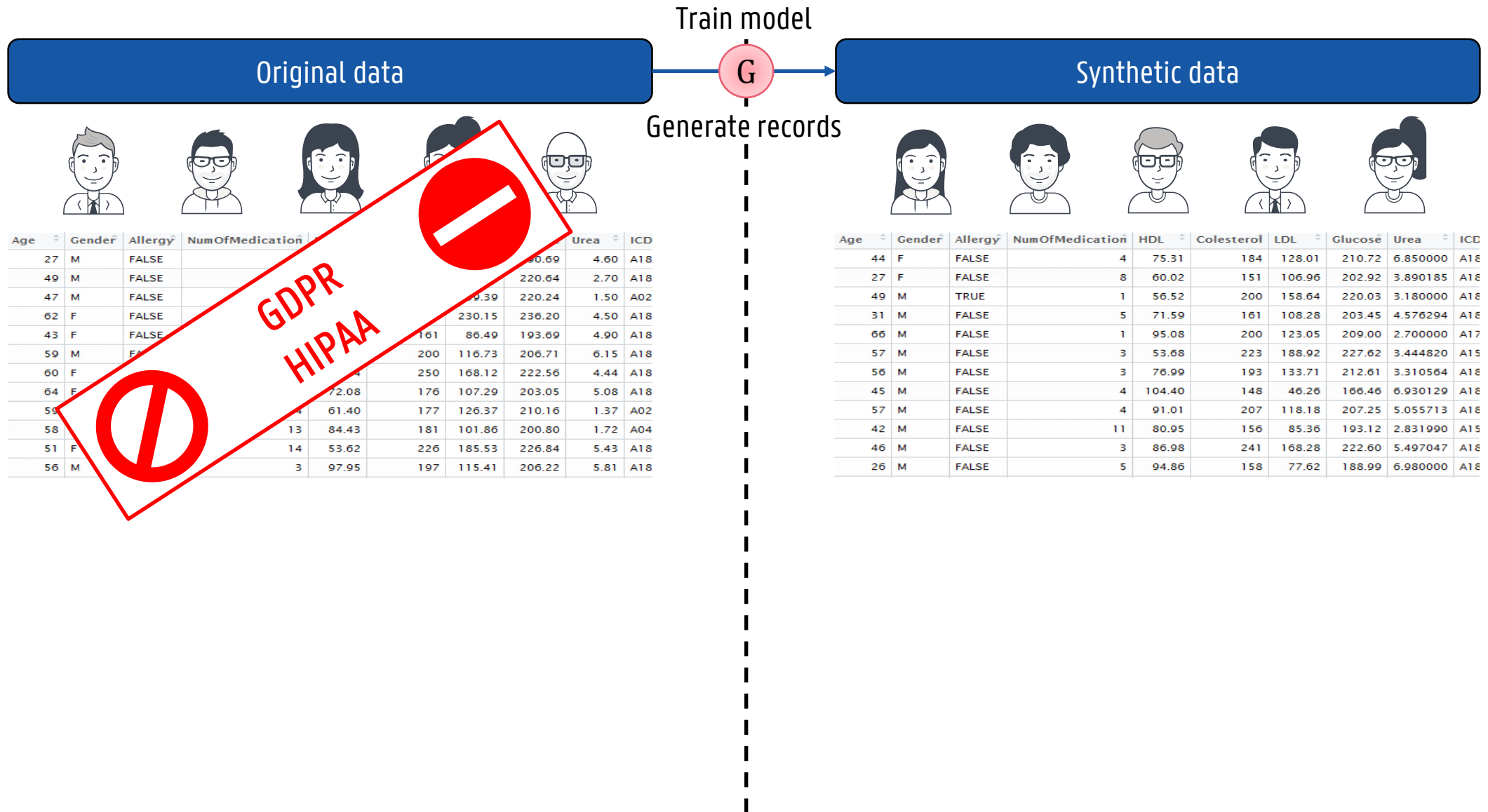
Age	Gender	Allergy	NumOfMedication	Urea	ICD
27	M	FALSE		30.69	4.60 A18
49	M	FALSE		220.64	2.70 A18
47	M	FALSE		9.39	220.24 1.50 A02
62	F	FALSE		230.15	236.20 4.50 A18
43	F	FALSE	161	86.49	193.69 4.90 A18
59	M	FALSE	200	116.73	206.71 6.15 A18
60	F	FALSE	250	168.12	222.56 4.44 A18
64	F	FALSE	72.08	176	107.29 203.05 5.08 A18
59	F	FALSE	61.40	177	126.37 210.16 1.37 A02
58	F	FALSE	13 84.43	181	101.86 200.80 1.72 A04
51	F	FALSE	14 53.62	226	185.53 226.84 5.43 A18
56	M	FALSE	3 97.95	197	115.41 206.22 5.81 A18



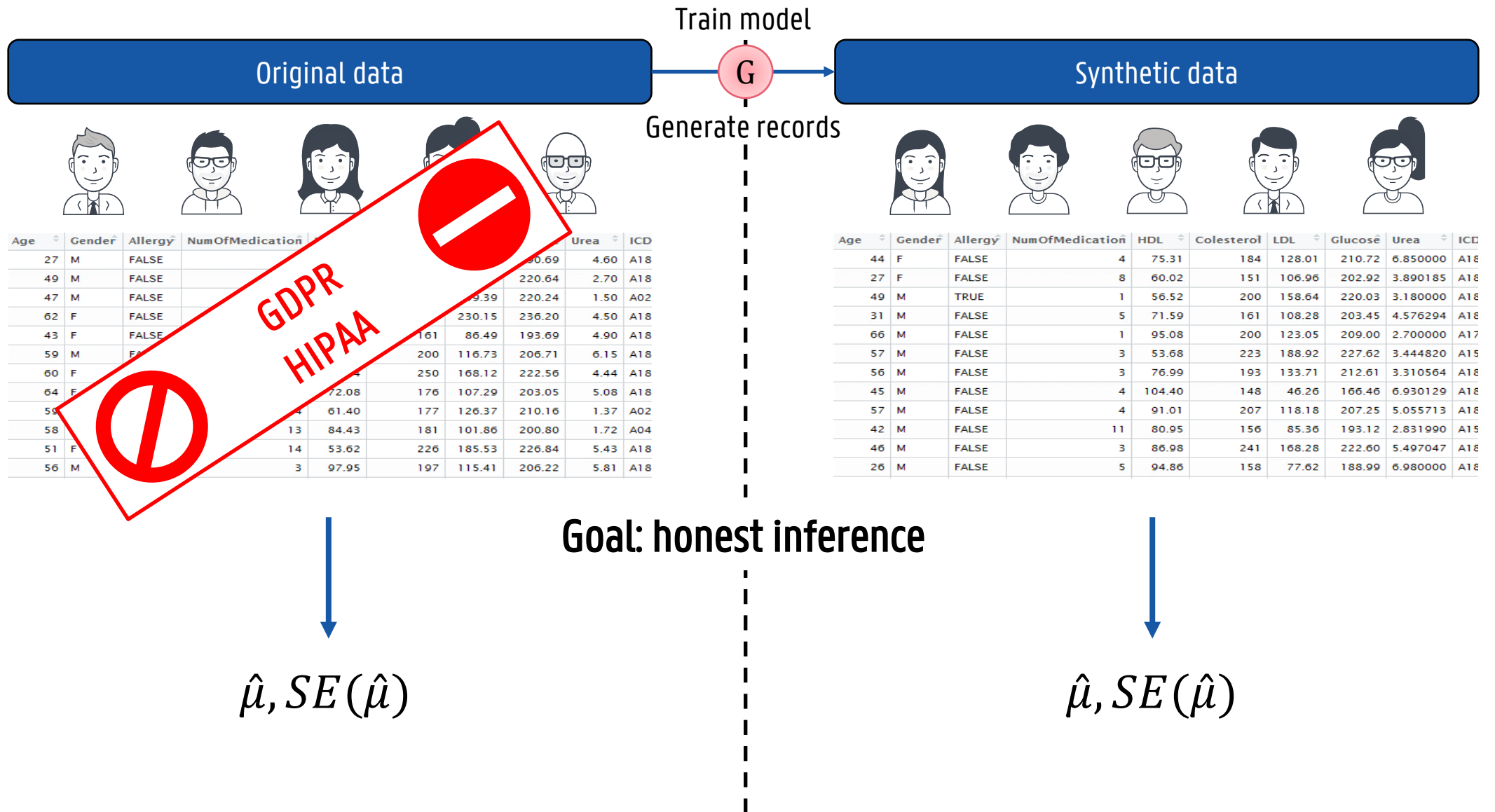
GDPR
HIPAA



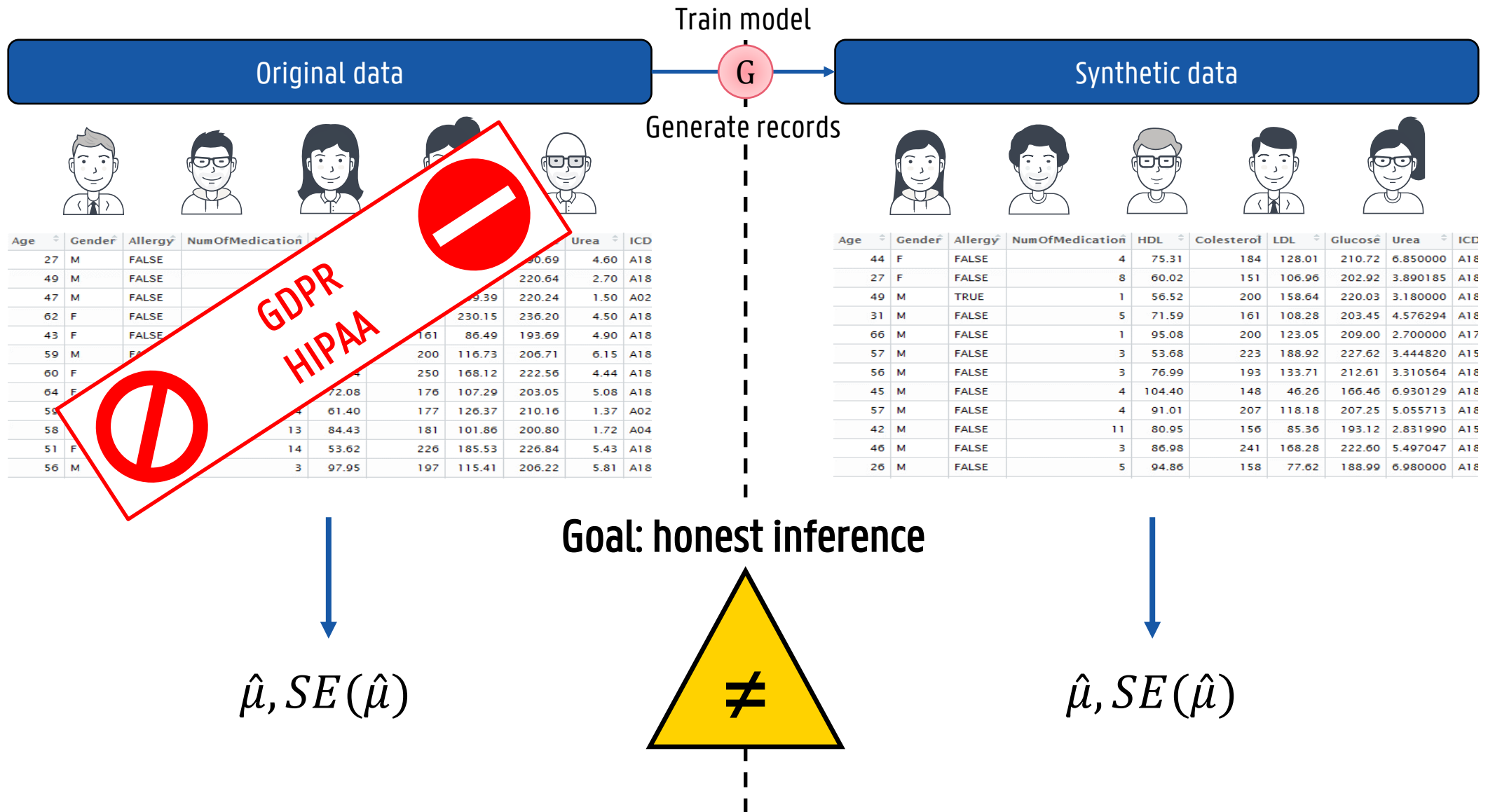
Naive statistical inference from synthetic data cannot be trusted



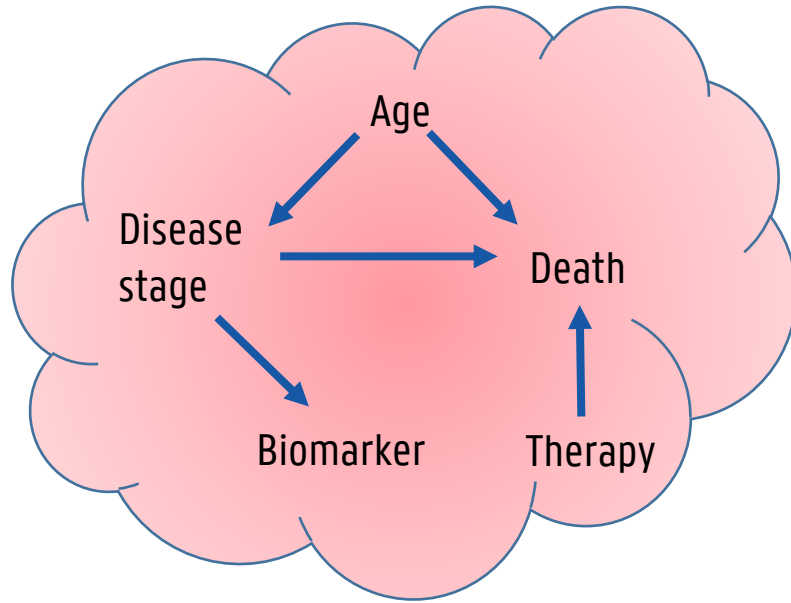
Naive statistical inference from synthetic data cannot be trusted



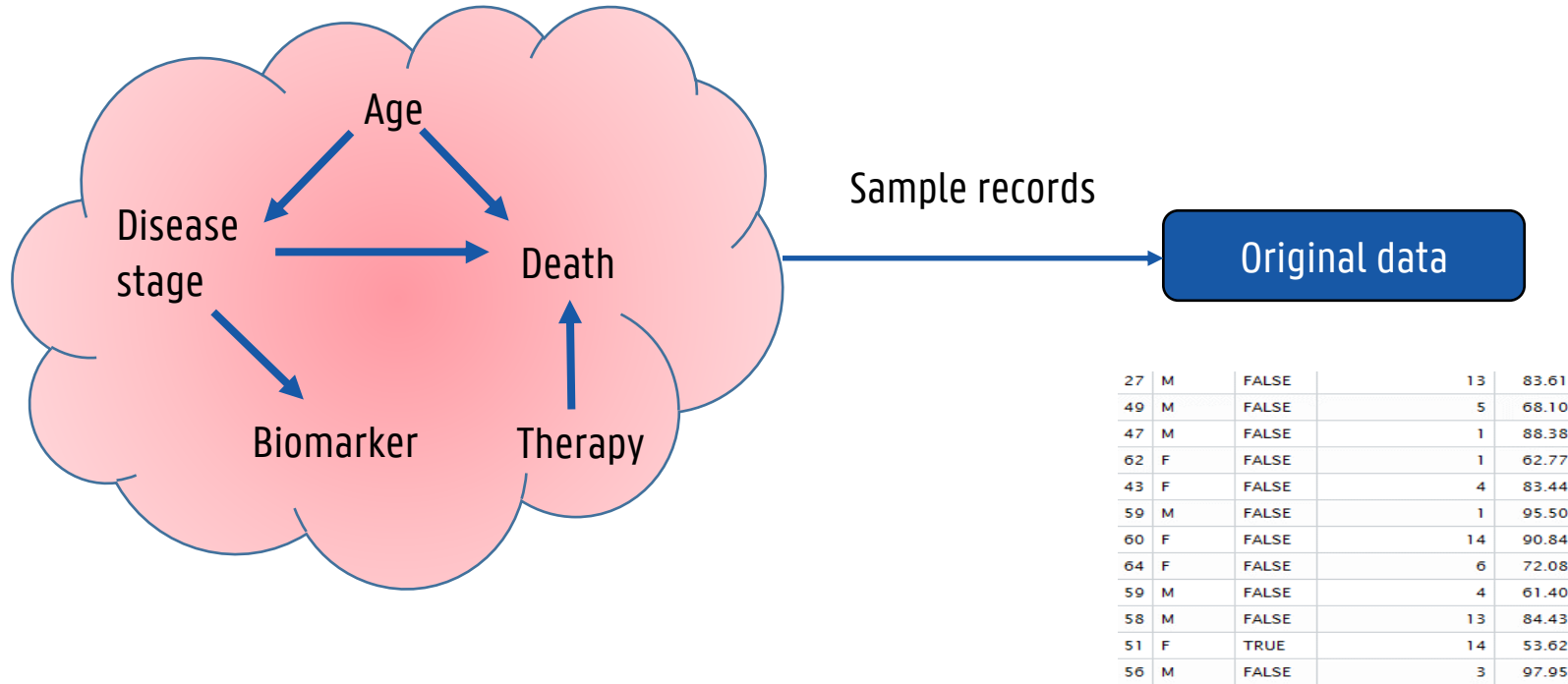
Naive statistical inference from synthetic data cannot be trusted



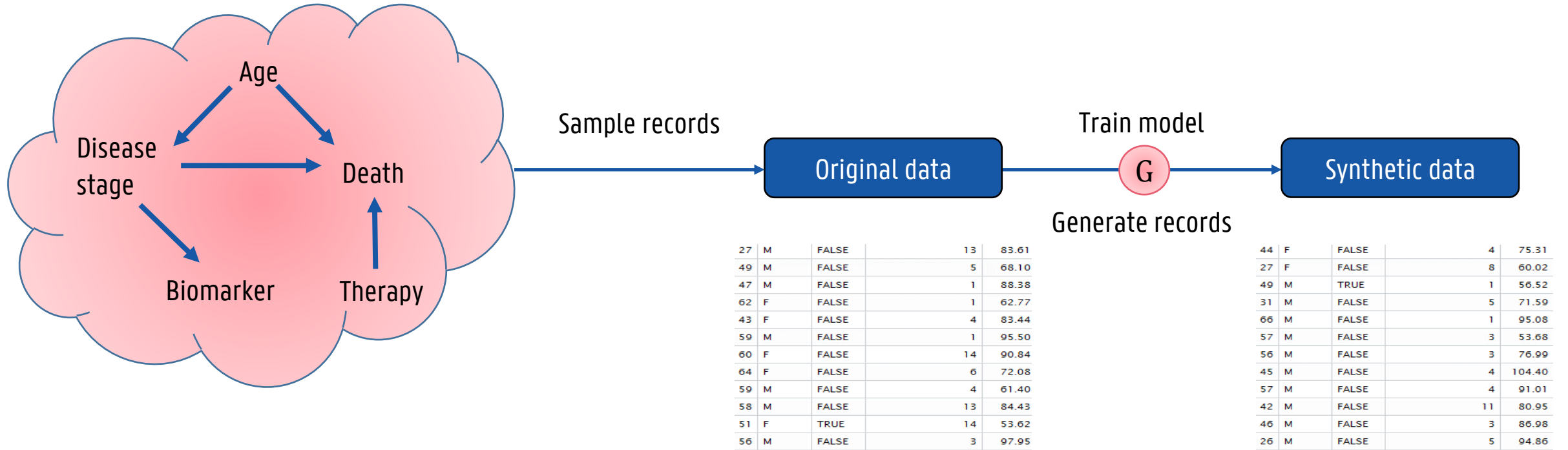
We investigated the behaviour of estimators in synthetic data



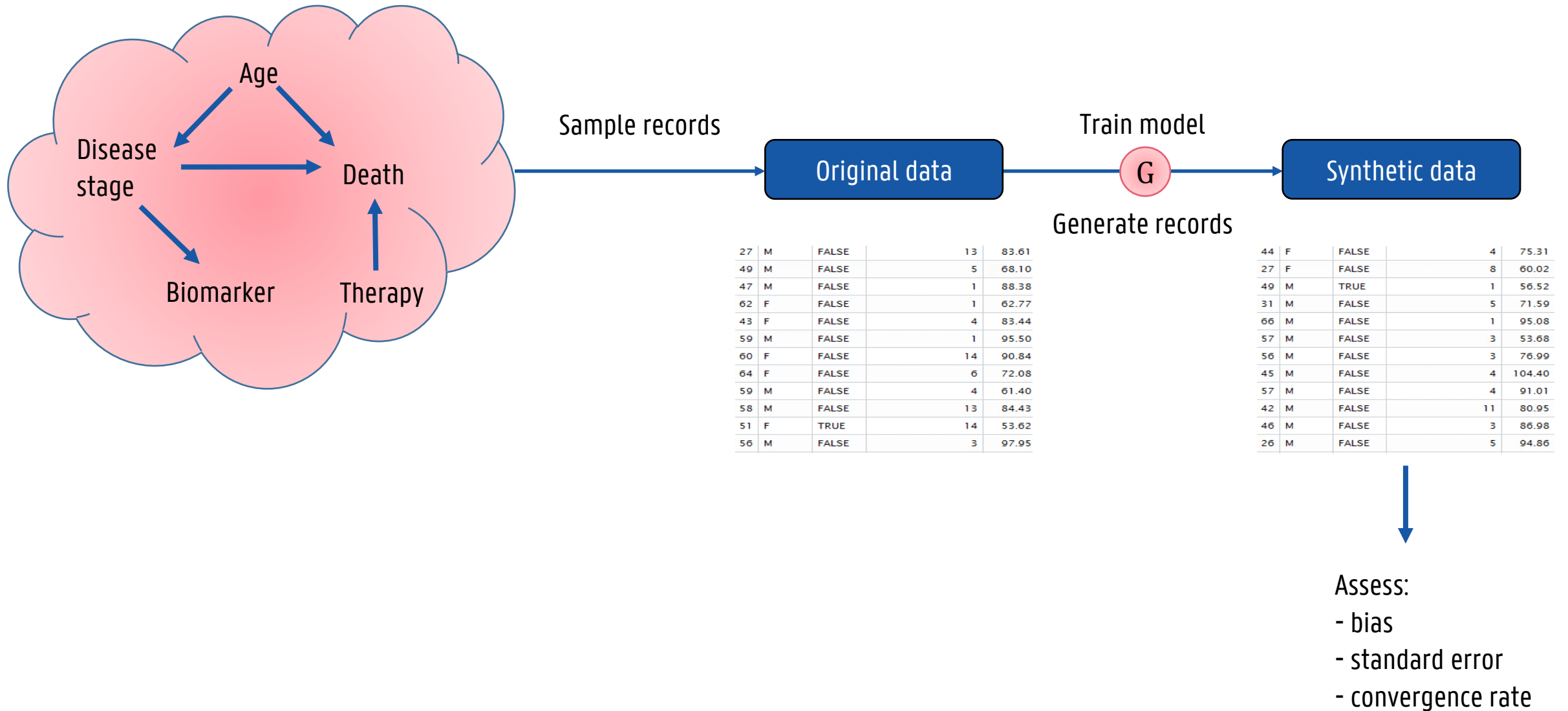
We investigated the behaviour of estimators in synthetic data



We investigated the behaviour of estimators in synthetic data

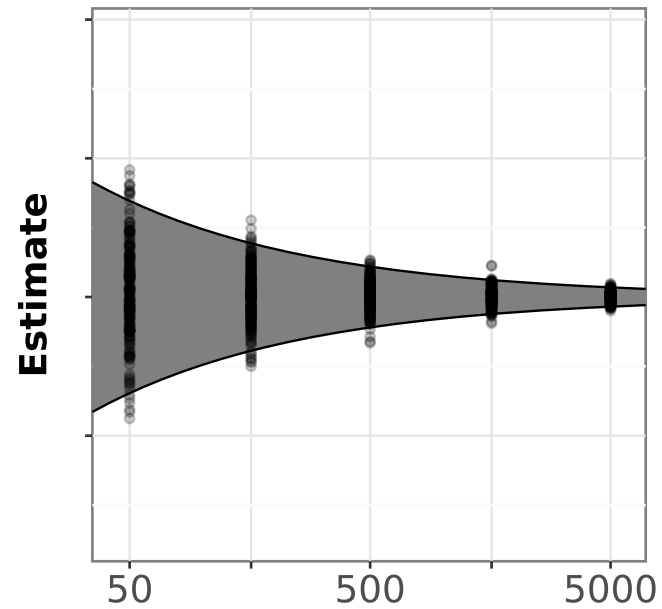


We investigated the behaviour of estimators in synthetic data



Original standard errors are **insufficient** for synthetic data

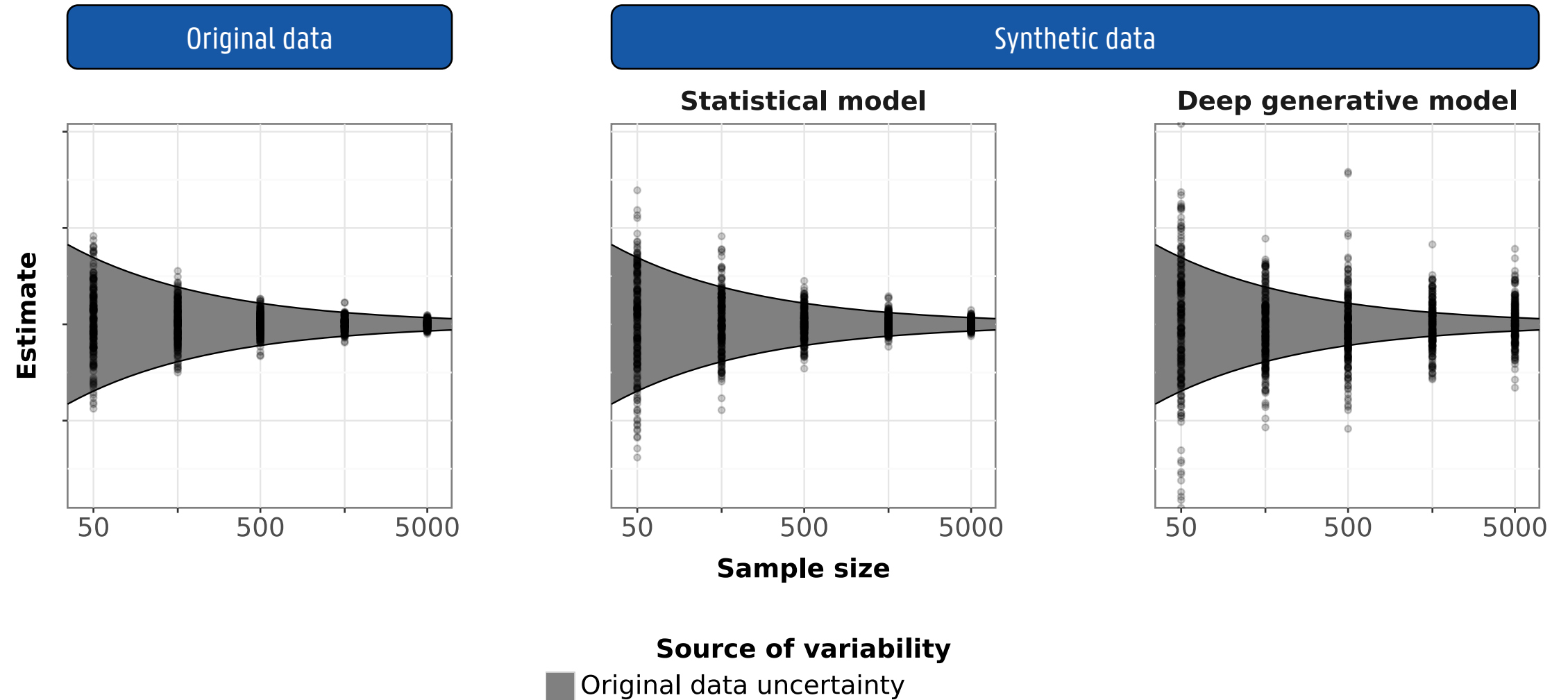
Original data



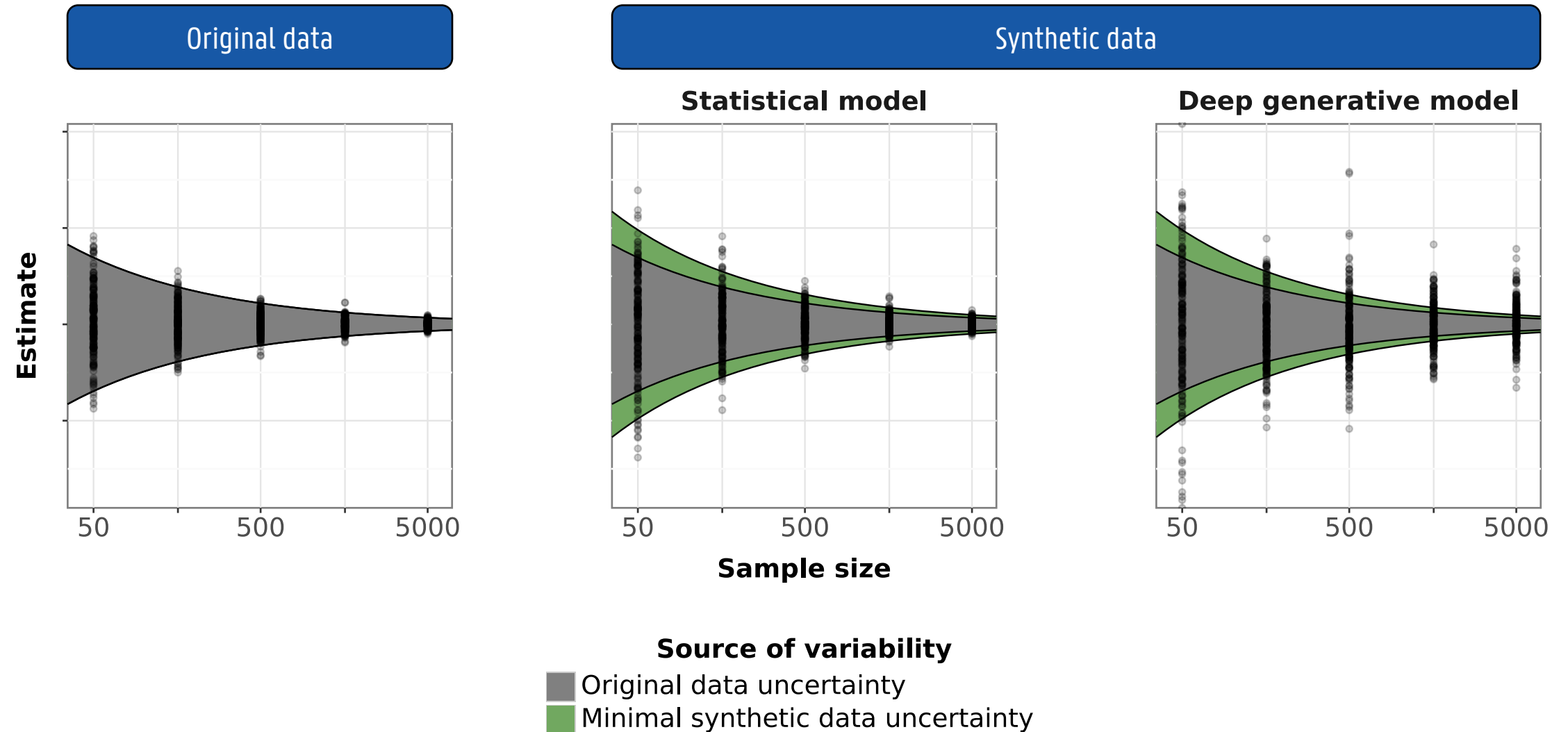
Source of variability

■ Original data uncertainty

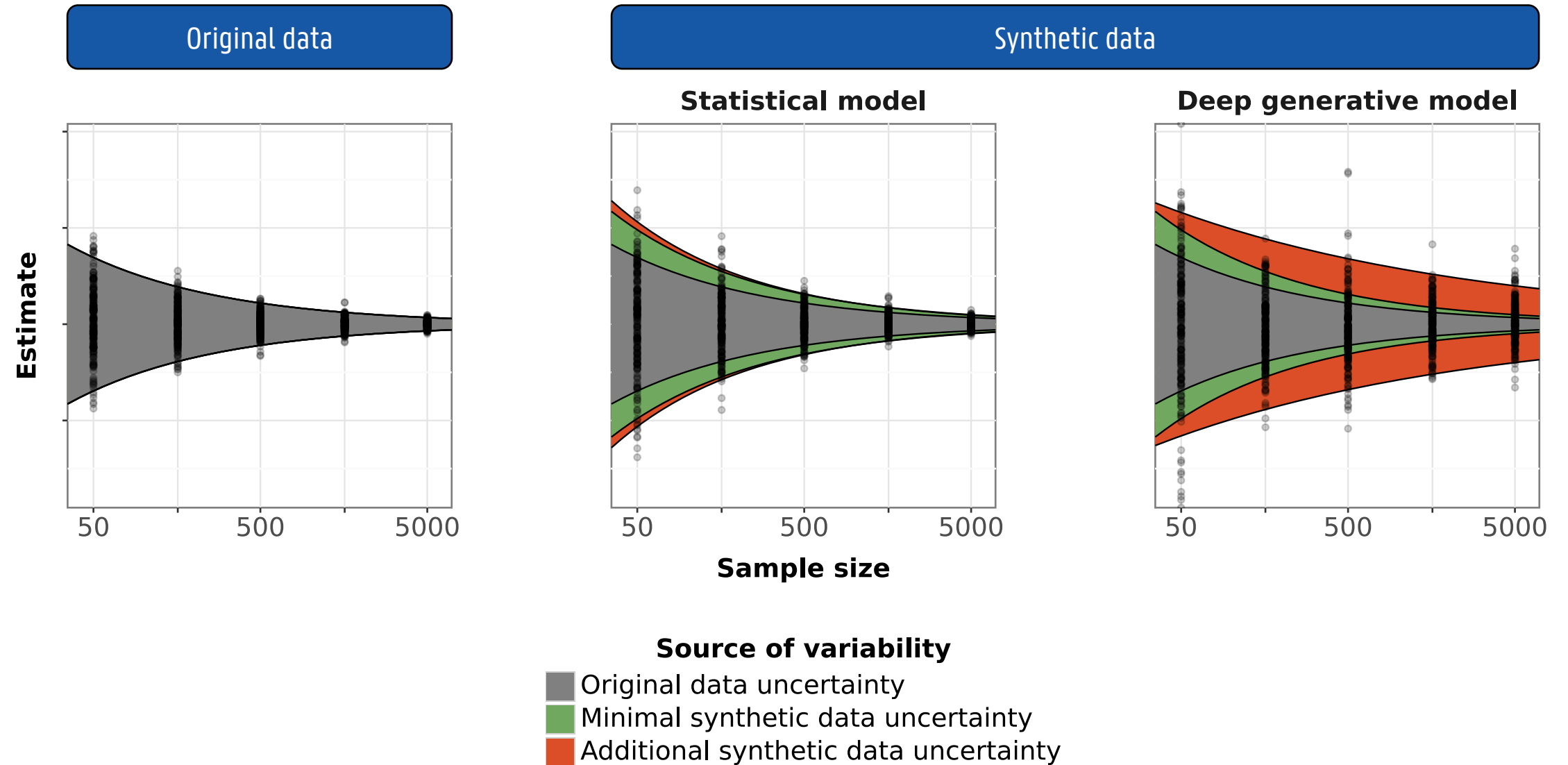
Original standard errors are **insufficient** for synthetic data



Corrected standard errors are **sufficient** only for statistical models



Corrected standard errors are **insufficient** for deep generative models



Find out more!



Heidelinde




Paloma




Alexander

Visit our poster (#406)



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 Rabary, 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.

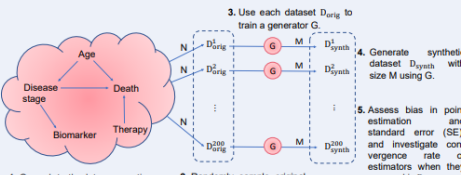
Synthetic data are artificial data that mimic the original data in terms of statistical properties. As such, synthetic data might be able to replace the original data in statistical analysis, while **preserving the privacy** of the individual members of the original dataset.

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**.

- Extra uncertainty** should be acknowledged since the distribution learned by the generative model is an approximation.
- 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



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**.

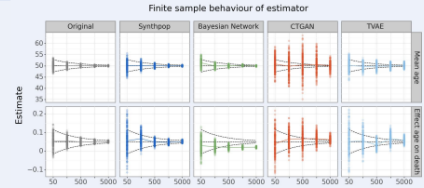
Minimal adaptation estimator SE:

$$\sigma_{\beta, corrected} = \sigma_{\beta, naive} \sqrt{1 + \frac{M}{N}}$$

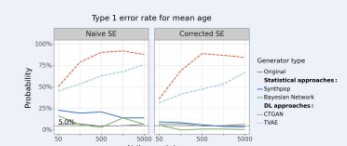
Is this valid for all estimators and generators? What about \sqrt{N} -consistency?

Results inferential utility

Finite sample behaviour of estimator





Type 1 error rate for mean age



Main results:


- Generative model **misspecification** introduces **bias**.
- True **SEs are larger** for D_{synth} than for D_{orig} and extra variability varies over generative models.
- Therefore, **naive estimation of SE leads to its underestimation**.
- 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**.
- Naive analyses** lead to inflation of **type 1 error rate** (compromising inferential utility).
- 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.

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

synudara.ai



SYNthetic DATA for Research Acceleration



SYNDARA

