Synthetic Data: Can We Trust Statistical Estimators?

Alexander Decruyenaere[°*†], **Heidelinde Dehaene**[°*†], **Paloma Rabaey**[§†], **Christiaan Polet**[*†], **Johan Decruyenaere**[*†], **Stijn Vansteelandt**[†], **Thomas Demeester**[§†]

[*]Ghent University Hospital – SYNDARA [§]imec [†]Ghent University

{firstname.lastname}@ugent.be

Abstract

The increasing interest in data sharing makes synthetic data appealing. However, the analysis of synthetic data raises a unique set of methodological challenges. In this work, we highlight the importance of inferential utility and provide empirical evidence against naive inference from synthetic data (that handles these as if they were really observed). We argue that the rate of false-positive findings (type 1 error) will be unacceptably high, even when the estimates are unbiased. One of the reasons is the underestimation of the true standard error, which may even progressively increase with larger sample sizes due to slower convergence. This is especially problematic for deep generative models. Before publishing synthetic data, it is essential to develop statistical inference tools for such data.

1 Introduction

In recent years, the use of synthetic data as a privacy-preserving substitute for real sensitive data has been gaining attention (Drechsler, 2011; Raab et al., 2016; Raghunathan, 2021). Synthetic data are artificial data that mimic the original data in terms of statistical properties, without revealing any individual records (Chen et al., 2021). As such, synthetic data might be able to replace the original data in statistical analysis, while preserving the privacy of the individuals of the original dataset and thereby fulfilling the regulatory privacy constraints. This is especially appealing for medical research (Jordon et al., 2022; Zhang et al., 2020). Synthetic data can be generated using statistical techniques or deep learning (DL) approaches such as GANs and VAEs (Hernandez et al., 2022).

Statistical inference is typically based on $\sqrt{N}$-consistency and asymptotic normality (Boos and Stefanski, 2013). In this work, we empirically provide evidence against naive inference from synthetic data, whereby synthetic data are treated as if they were actually observed. First, the extra uncertainty coming from the fact that the data were generated by a predictive model themselves should be acknowledged. Otherwise, this will yield estimators with standard errors (SEs) that are too optimistically small (Raab et al., 2016). Second, the regularisation bias inherent to data-adaptive (DL) techniques (i.e., their bias-variance trade-off is optimised with respect to the prediction error instead of the error in the estimator) may diminish too slowly as the sample size grows larger, causing excess variability that is difficult to systematically account for (Brain and Webb, 1999; Hines et al., 2022). This may deliver estimators that are biased and/or whose SE converges slower than $1/\sqrt{N}$.

The remainder of this paper is structured as follows. In Section 2, we introduce a framework that allows us to empirically assess the behaviour of statistical estimators when it is (incorrectly) assumed that synthetic data can be treated like original data. We consider a setting with low-dimensional tabular data that is commonly featured in applied research. Section 3 discusses the statistical properties of the estimators and addresses how to correct for some of the added variability inherent to synthetic data in the model-based SE, as originally proposed by Raab et al. (2016). Section 4 presents the results of our simulation study, draws attention to the difference in inferential utility between synthetic data generated through statistical vs. DL techniques, and elaborates on why the corrected SE fails with DL techniques. Finally, Section 5 provides some concluding remarks and hints at future work.

[°]Joint first authors.

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

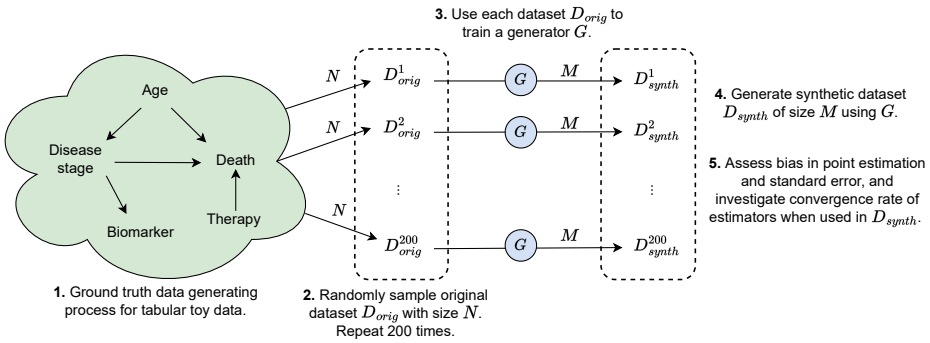

Figure 1: Visual representation of our experimental setup.

2 Experimental Setup

Figure 1 shows the steps that are taken in order to assess the finite sample performance of various statistical estimators in the context of synthetic data. The code to reproduce our simulation study is available on `https://github.com/syndara-lab/inferential-utility-workshop`.

Step 1: Ground truth data distribution We simulate low-dimensional tabular toy data sampled from an arbitrary ground truth population. A generic setting with five variables is considered: *age* (continuous), *disease stage* (ordinal with four categories), *biomarker* (continuous), *therapy* (binary), and *death* (binary). The dependency structure of these variables is encoded by a directed acyclic graph (DAG), shown in Figure 1. For details on the distributions of the variables, we refer to Appendix A.1.

Step 2: Sample original data D_{orig} In a Monte Carlo simulation, we randomly draw N samples from the population, which form an original dataset D^i_{orig}. To study the behaviour of statistical estimators with increasing sample sizes, we vary N log-uniformly between 50 and 5000 ($N \in \{50, 160, 500, 1600, 5000\}$). For each value of N, 200 Monte Carlo simulation runs are used.

Step 3: Train generative model G Each dataset D^i_{orig} of size N is used to train several types of generative models. The considered state-of-the-art generative models fall into two categories: statistical and DL approaches (Hernandez et al., 2022). The statistical approaches we investigate include `Synthpop` (Nowok et al., 2016) (R package which facilitates fitting a series of conditional predictive models, the type of which is inferred from the data, the order provided by the user) and Bayesian network with user-specified DAG (`BN`) (Ankan and Panda, 2015). For the DL approaches, we consider two popular deep generative models adapted for tabular synthetic data, namely `CTGAN` and `TVAE` (Xu et al., 2019). Additional information on these methods can be found in Appendix A.2.

Step 4: Generate synthetic data D_{synth} Using the trained generative model G, we generate a single synthetic dataset D^i_{synth} of size M. In our study, we set $M = N$ to retain the sample size of the original dataset, to test whether the user would get similar results on both. For a particular choice of N, this leaves us with 200 synthetic datasets for each of the four generator types.

Step 5: Assess statistical properties of estimators based on D_{synth} The last step concerns the main contribution of this work and is explained in-depth in the next section.

3 Evaluating Statistical Properties based on Synthetic Data

We assess the behaviour of some standard statistical estimators when it is (incorrectly) assumed that D_{synth} can be treated like D_{orig}. These estimators are estimated on each synthetic dataset D^i_{synth}, resulting in an estimate of the population parameter in question. For each generator type, we then have 200 estimates per sample size N, for a variety of sample sizes N, allowing us to evaluate their **empirical bias** (i.e., the average difference between the estimates and the population parameter) and **empirical SE** (i.e., the standard deviation of the estimates). Standard statistical analysis assumes that the bias converges faster than the SE with the latter diminishing at a rate of $1/\sqrt{N}$. The corrected SE discussed later in this section, though taking into account the added variability of the synthetic data generating process, still relies on the same assumption, thus remaining invalid in the context of deviating convergence rates.

As a result of the excess variability induced by the generation process that may, in addition, decrease slower than $\sqrt{N}$, we foresee atypical behaviour of the estimators (which will be confirmed by our simulation results in Section 4). This impacts the **type 1 error rate** (the probability to find a significant effect when in truth there is none) of statistical analyses performed on synthetic data.

Motivated by commonly used estimators in applied research, our analysis in Section 4 initially focuses on two statistical estimators: sample mean of *age* ("*mean age*") and a coefficient ("*effect age on death*") of a multiple logistic regression model with *death* as outcome and *age*, *stage* and *therapy* as covariates. We report results for the other coefficients of this regression model (effect of *therapy* and *stage* on *death*) in Appendix A.3.

As will be shown empirically in Section 4, the model-based SE of the considered estimators systematically underestimates the empirical SE by ignoring the extra variability caused by the synthetic data generation process. To partially account for this added variability, we will use the following SE estimator proposed by Raab et al. (2016) and adapted to our setting:

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

where $\sigma_{\hat{\theta},\text{ naive}}$ is the model-based SE in the synthetic data, M the sample size of the synthetic data D_{synth}, and N the sample size of the original data D_{orig}. Note that this is a minimal correction since it only applies to $\sqrt{N}$-consistent estimators, hence the added variability resulting from the regularisation bias in data-adaptive techniques (i.e., DL approaches), is not accounted for.

4 Simulation study

Bias and SE The estimates are depicted in Figure 2 for the two estimators and for all four generator types. This figure allows a qualitative and empirical assessment of two key properties of statistical estimators: bias and SE. The funnel indicates the behaviour of an unbiased and $\sqrt{N}$-consistent estimator based on original data. Table 1 lists the same information numerically, summarising the relative bias ($\text{RE}_{\hat{\theta}}$) and the relative underestimation of the empirical SE by the naive model-based SE ($\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$). Figure A1 and Table A1 in the appendix show these metrics for all estimators.

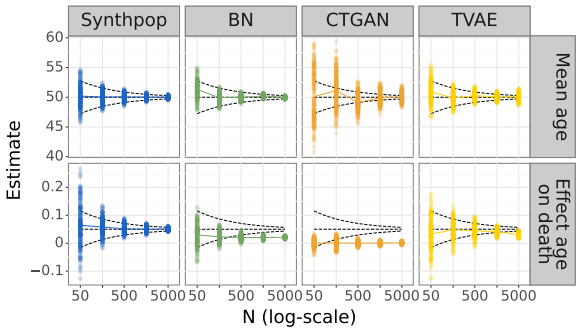

Figure 2: Simulation study results for the two estimators. Each dot is an estimate per Monte Carlo run (200 in total per N). The population parameter is represented by the horizontal dashed line. The funnel indicates the behaviour of an unbiased and $\sqrt{N}$-consistent estimator based on observed data.

First, generative model misspecification will introduce bias. This is, for example, noticeable in the *effect age on death* coefficient generated by `BN`, which is incorrectly estimated due to discretisation of the continuous variable age during the generation of $death$. Interestingly, the DL approaches (`CTGAN` and `TVAE`), despite being more flexible and also tuned to prevent overfitting, fail to adequately fit the joint distribution in our simulation study, leading to bias for several estimators (e.g., biased null effect for *effect age on death* by `CTGAN`).

Second, the empirical SEs are larger for D_{synth} than for D_{orig} and may vary over generative models. Larger empirical SEs reflect the additional (predictive) uncertainty in the generation of synthetic data, which seems most pronounced with DL approaches. This uncertainty is discarded in the naive use of model-based SEs, leading to underestimation of the empirical SE (as is evident from Table 1).

Convergence rate Assuming a power law N^{-a} in convergence rate for the empirical SE, we estimated the exponent a from five logarithmically spaced sample sizes N between 50 and 5000, shown for the two estimators in Table 2 and all estimators in Table A2 and Figure A2 in the appendix. As expected, the empirical SE of estimators based on original data indeed converges at a rate of $1/\sqrt{N}$ (i.e., $a_{SE} = 0.5$). Fully parametric generative models are also expected to yield estimators of which the SE decreases at a rate of $1/\sqrt{N}$. This is largely confirmed for the statistical approaches we

Table 1: Relative error (RE) for the two estimators, averaged over 200 Monte Carlo runs. $\text{RE}_{\hat{\theta}}$ is the relative bias of the estimates $\hat{\theta}$. $\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$ is the RE between the naive model-based SE ($\hat{\sigma}_{\hat{\theta},naive}$) and the empirical SE. Positive and negative values indicate a relative over- and underestimation.

	Synthpop				**BN**			
	$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$		$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$	
Estimator	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$
Mean age	0.20	0.01	-36.21	-31.23	2.59	-0.10	-6.17	-2.32
Effect age on death	30.55	0.83	-38.11	-28.42	-41.76	-57.81	-9.98	-3.13
	CTGAN				**TVAE**			
	$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$		$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$	
Estimator	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$
Mean age	0.21	0.13	-51.19	-79.15	1.81	0.13	-48.92	-75.99
Effect age on death	-100.83	-98.41	24.45	-29.73	-29.65	-37.95	-18.03	-53.96

Table 2: Estimated exponent a [95% CI] for the power law convergence rate N^{-a} for empirical SE. Deviations from 0.50 indicate that the estimator is not $\sqrt{N}$-consistent.

	Mean age	**Effect age on death**
Original	0.50 [0.49; 0.50]	0.49 [0.47; 0.52]
Synthpop	0.45 [0.43; 0.47]	0.43 [0.36; 0.49]
BN	0.50 [0.49; 0.50]	0.50 [0.47; 0.52]
CTGAN	0.29 [0.25; 0.32]	0.53 [0.47; 0.58]
TVAE	0.39 [0.33; 0.46]	0.40 [0.36; 0.44]

Table 3: Type 1 error rates (%) of the one-sample t-test at $\alpha = 5\%$ for the population mean of age with naive model-based SEs and corrected SEs. The null hypothesis states that the population mean of age is equal to the ground truth.

	Naive SE		**Corrected SE**	
	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$
Synthpop	20.0	15.0	9.0	6.5
BN	16.5	6.5	5.5	0.5
CTGAN	39.5	78.0	15.0	68.0
TVAE	34.5	62.5	19.0	49.0

studied, especially `BN`, though `Synthpop` converges a bit slower (i.e., $a_{SE} < 0.5$), as non-parametric components are built into its generation process. By contrast, the SEs produced by the more data-adaptive DL approaches converge much slower (i.e. $a_{SE} << 0.5$), except for the logistic regression coefficients that are estimated as (biased) null effects on synthetic data generated by `CTGAN`. The slower-than-$\sqrt{N}$-convergence leads to a progressively increasing underestimation of the empirical SE by the model-based SE (which assumes $\sqrt{N}$-convergence) as the sample size grows larger, as seen in Table 1.

The convergence rate of the bias (or equivalently, convergence rate of the estimator) was estimated analogously and is presented in Table A2 in the appendix. As opposed to default behaviour, the bias may converge slower than the SE ($a_{bias} \leq a_{SE}$) for some estimators and generators.

Inferential utility The null hypothesis significance testing framework typically uses an estimate and its associated (un)certainty (reflected by the SE) as a test statistic to reject a null hypothesis. Due to (1) slower convergence of (the bias of) the estimator compared to the SE and/or (2) underestimation of the empirical SE by the naive model-based SE, we indeed observe an inflation of the type 1 error rate beyond the nominal level of $\alpha = 5\%$, as shown in Table 3 and Figure A3 in the appendix: the more the empirical SE is underestimated by the naive model-based SE (which is especially the case for the DL approaches with increasing sample size), the larger the inflation. Use of corrected SEs (as introduced in Section 3) will control the type 1 error rate at approximately 5%, but only for the (parametric) statistical approaches. By contrast, the corrected SE does not account sufficiently for the predictive uncertainty of the DL approaches due to (slower-than-$\sqrt{N}$-convergence of) their regularisation bias, resulting in an actual type 1 error rate still higher than the nominal level.

5 Conclusion

Naive inference from synthetic data leads to a flurry of false-positive findings, even when the estimates are unbiased. One of the reasons is the underestimation of the true standard error, which may even progressively increase with larger sample sizes due to slower convergence. This is especially problematic for deep generative models. Before publishing synthetic data to enable data sharing and accelerate research, it is essential to develop statistical inference tools for such data.

Acknowledgements

This research was funded by a grant received from the Fund for Innovation and Clinical Research of Ghent University Hospital. Paloma Rabaey's research is funded by the Research Foundation Flanders (FWO-Vlaanderen) with grant number 1170122N. This research also received funding from the Flemish government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*, 2015.

Dennis D Boos and Leonard A Stefanski. *Essential statistical inference: theory and methods*, volume 591. Springer, 2013.

Damien Brain and Geoffrey I Webb. On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales*, pages 117–128, 1999.

Richard Chen, Ming Lu, Tiffany Chen, Drew Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5:1–5, 06 2021.

Jörg Drechsler. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media, 2011.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76:1–48, 01 2022.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how?, 2022.

Beata Nowok, Gillian M Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26, 2016.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016.

Gillian M Raab, Beata Nowok, and Chris Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97, 2016.

Trivellore E Raghunathan. Synthetic data. *Annual review of statistics and its application*, 8:129–140, 2021.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *Advances in neural information processing systems*, 32, 2019.

Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108, 2020.

A Details experimental setup

A.1 Ground truth data distribution

It was opted to work with low-dimensional tabular data given its frequent use in applied medical research. Different traditional and commonly used regression models were taken into account when choosing the nature of the variables. We wish to have a mix of variables that are continuous, binary, ordinal, normally distributed or skewed. To obtain these requirements and reflect a generic clinical setting, the data generating process consists of the following five variables: *age* (continuous), *disease stage* (ordinal with four categories or stages), *biomarker* (continuous), *therapy* (binary), and *death* (binary). It is assumed that a patient is observed at a given point in time, which is further referred to as the baseline time point. At this time, patient data about *age*, *disease stage*, *biomarker*, and the random assignment of *therapy* is gathered. The binary outcome variable *death* is evaluated at a later time point. This design is a simplification of reality since we assume that there are no missing data and we do not consider the data as longitudinal.

Age follows a normal distribution with mean 50 and standard deviation 10. *Disease stage* was generated according to a proportional odds cumulative logit model where an increase in *age* causes an increase in the odds of having a *disease stage* higher than a given stage k, $\nu_{age} = -0.05$. The variable *biomarker* is a quantification of the *disease stage* and was also based on a Generalized Linear Model (GLM), where *biomarker* follows a gamma distribution and its mean changes in function of *disease stage*. It was constructed in such a way that a higher *disease stage* results in higher values for the *biomarker* ($\gamma_0 = 4$, $\gamma_{stage} = \{0, -1, -2, -3\}$ for stage I-IV, respectively). *Therapy* is considered to be 1:1 randomly assigned and is therefore sampled from a Bernouilli distribution with a probability of 0.50. The last variable, *death*, is generated by using a binomial logistic regression model in which the odds of *death* increase with an increasing *age* ($\beta_{age} = 0.05$), a higher *disease stage* ($\beta_{stage} = \{0, 0.50, 1.00, 1.50\}$ for stage I-IV, respectively), and absence of *therapy* ($\beta_{therapy} = -0.50$).

A.2 Generative models

The hyperparameters of CTGAN and TVAE were tuned using the Tree-structured Parzen Estimator algorithm with the 5-fold cross-validated average inverse of the Kullback–Leibler divergence (IKLD) between the original and the synthetic dataset as objective score, averaged over five seed initializations. We use the `Python` package `SDV` (Patki et al., 2016) to train our CTGAN and TVAE models. Numerical features are preprocessed using a cluster-based normalizer.

Our *synthpop* statistical approach relies on the `synthpop` package for `R`, which provides a routine to generate synthetic data (Nowok et al., 2016). This framework encompasses both parametric and non-parametric methods to sequentially fit a series of conditional joint distributions, based on the observed data. We restrict ourselves to the default parametric methods and provide information of the dependency structure of our data via specification of a directed acyclic graph (DAG). As a result, this defines the order of the sequence and which variables need to be included as predictors in the conditional models.

Our Bayesian Networks were implemented using the `Python` package `pgmpy` (Ankan and Panda, 2015). The ground-truth DAG was provided upfront, while the conditional probability distributions (CPDs) are estimated by Maximum Likelihood Estimation (MLE) and synthetic data are generated via forward sampling.

A.3 Full results simulation study

Figure A1 shows the simulation results for all estimators (both mean of *age* and all logistic regression coefficients for the variable *death*). Table A1 summarises the relative bias ($\text{RE}_{\hat{\theta}}$) and the relative underestimation of the empirical SE by the naive model-based SE ($\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$) for all estimators. Table A2 show the rate at which the empirical SE and bias decrease in function of N on the log-scale for the same set of estimators. Figure A2 visually compares the rate of convergence of the SE across all generators.

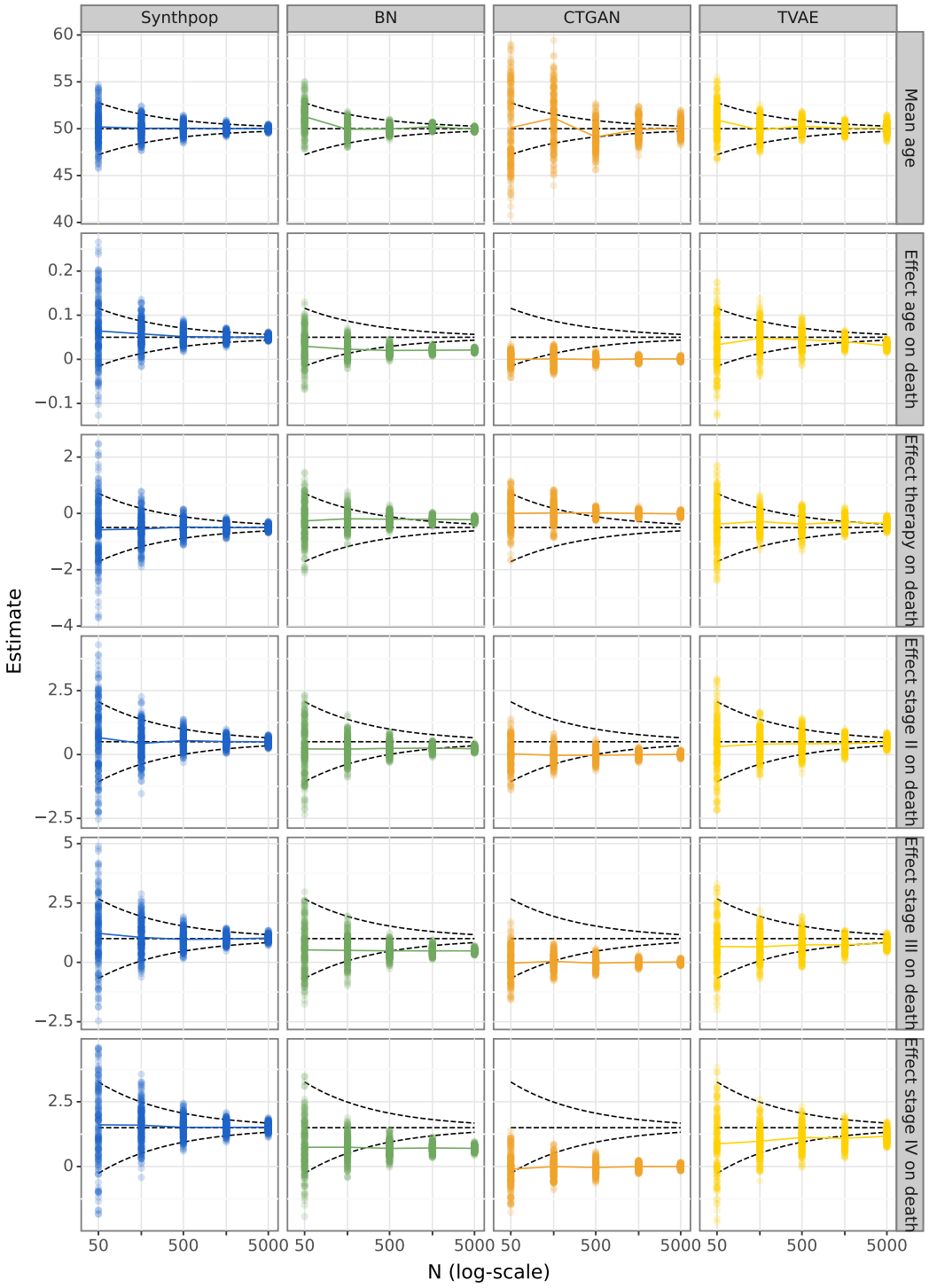

Figure A1: Simulation study results for all estimators. Each dot is an estimate per Monte Carlo run (200 dots in total per value of N). The population parameter is represented by the horizontal dashed line. The funnel indicates the behaviour of an unbiased and $\sqrt{N}$-consistent estimator based on observed data.

Table A1: Relative error (RE) for all estimators, averaged over 200 Monte Carlo runs. $\text{RE}_{\hat{\theta}}$ is the relative bias of the estimates $\hat{\theta}$. $\text{RE}_{\hat{\sigma}_{\hat{\theta}}}$ is the RE between the naive model-based SE ($\hat{\sigma}_{\hat{\theta},naive}$) and the empirical SE. Positive and negative values indicate a relative over- and underestimation.

	Synthpop				**BN**			
	$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\theta}}$		$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\theta}}$	
Estimator	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$
Mean								
Mean age	0.20	0.01	-36.21	-31.23	2.59	-0.10	-6.17	-2.32
Logistic regression								
Effect age on death	30.55	0.83	-38.11	-28.42	-41.76	-57.81	-9.98	-3.13
Effect therapy on death	23.94	-0.26	-34.81	-30.13	-50.16	-55.24	-11.63	-15.83
Effect stage II on death	31.98	-0.04	-30.26	-30.46	-49.14	-54.56	-12.22	-6.64
Effect stage III on death	23.83	2.31	-31.76	-32.47	-42.65	-51.54	-6.69	0.39
Effect stage IV on death	7.29	1.15	-25.42	-36.90	-45.20	-53.25	-9.68	-16.36
	CTGAN				**TVAE**			
	$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\theta}}$		$\text{RE}_{\hat{\theta}}$		$\text{RE}_{\hat{\sigma}_{\theta}}$	
Estimator	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$	$N = 50$	$N = 5000$
Mean								
Mean age	0.21	0.13	-51.19	-79.15	1.81	0.13	-48.92	-75.99
Logistic regression								
Effect age on death	-100.83	-98.41	24.45	-29.73	-29.65	-37.95	-18.03	-53.96
Effect therapy on death	-99.92	-96.48	10.41	-17.28	-17.43	-28.77	-21.86	-57.59
Effect stage II on death	-104.41	-99.37	38.07	6.12	-37.74	-5.12	-19.34	-51.34
Effect stage III on death	-104.58	-98.90	30.65	12.75	-34.81	-19.28	-10.76	-53.25
Effect stage IV on death	-104.90	-99.79	25.92	10.90	-41.73	-21.86	-9.56	-63.30

Table A2: Estimated exponent a [95% CI] for the power law convergence rate N^{-a} for empirical SE and bias of all estimators. Standard statistical analysis assumes that the bias converges faster than the SE with the latter diminishing at a rate of $1/\sqrt{N}$ ($a_{SE} = 0.50$). Bold values indicate instances where the bias converges slower than the SE for the same generator and estimator.

		Generator			
	Original	Synthpop	BN	CTGAN	TVAE
Estimator, SE					
Mean					
Mean age	0.50 [0.49; 0.50]	0.45 [0.43; 0.47]	0.50 [0.49; 0.50]	0.29 [0.25; 0.32]	0.39 [0.33; 0.46]
Logistic regression					
Effect age on death	0.49 [0.47; 0.52]	0.43 [0.36; 0.49]	**0.50** [0.47; 0.52]	**0.53** [0.47; 0.58]	0.40 [0.36; 0.44]
Effect therapy on death	0.50 [0.48; 0.51]	0.44 [0.39; 0.49]	**0.49** [0.47; 0.51]	**0.51** [0.47; 0.56]	0.41 [0.38; 0.43]
Effect stage II on death	0.49 [0.47; 0.50]	0.44 [0.40; 0.48]	**0.49** [0.46; 0.51]	**0.52** [0.50; 0.55]	0.41 [0.40; 0.43]
Effect stage III on death	0.49 [0.48; 0.51]	0.44 [0.39; 0.49]	**0.49** [0.46; 0.51]	**0.53** [0.51; 0.55]	0.42 [0.40; 0.43]
Effect stage IV on death	0.49 [0.48; 0.51]	0.44 [0.41; 0.47]	**0.49** [0.47; 0.50]	**0.54** [0.51; 0.56]	0.41 [0.38; 0.43]
Estimator, bias					
Mean					
Mean age	0.93 [0.81; 1.04]	0.93 [0.79; 1.07]	0.68 [0.44; 0.93]	0.59 [0.33; 0.85]	0.65 [0.53; 0.78]
Logistic regression					
Effect age on death	0.88 [0.66; 1.10]	0.81 [0.65; 0.97]	**0.32** [0.18; 0.47]	**0.24** [0.15; 0.32]	0.51 [0.20; 0.82]
Effect therapy on death	0.91 [0.82; 0.99]	0.91 [0.87; 0.96]	**0.41** [0.26; 0.57]	**0.33** [0.21; 0.44]	0.51 [0.29; 0.73]
Effect stage II on death	0.99 [0.74; 1.24]	1.03 [0.85; 1.22]	**0.46** [0.29; 0.63]	**0.36** [0.23; 0.49]	0.66 [0.54; 0.79]
Effect stage III on death	0.90 [0.68; 1.13]	0.84 [0.64; 1.04]	**0.37** [0.23; 0.52]	**0.27** [0.17; 0.37]	0.47 [0.34; 0.60]
Effect stage IV on death	0.84 [0.75; 0.92]	0.80 [0.67; 0.93]	**0.32** [0.19; 0.45]	**0.22** [0.14; 0.29]	0.42 [0.31; 0.52]

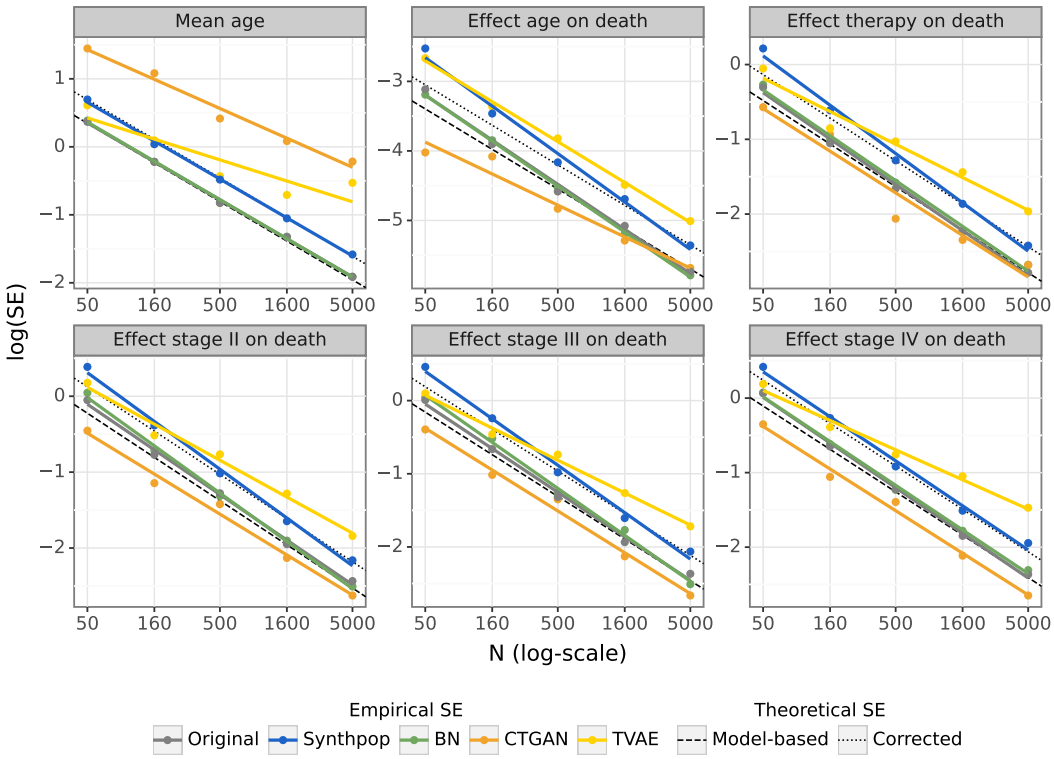

Figure A2: Convergence rate of the empirical standard error (SE). If the SE is of the form SE $= cN^{-a}$, where c is a constant, then $\log(SE) = c + (-a)\log N$. Therefore slope a represents the convergence rate and the vertical offset c indicates the asymptotic variance. The dashed line indicates the behaviour of the SE of an unbiased and $\sqrt{N}$-consistent estimator based on observed data, whereas the dotted line indicates the assumed behaviour of the SE of the same estimator based on synthetic data, following the correction proposed by Raab et al. (2016). Note that the asymptotic variance of the logistic regression coefficients ("effect [predictor] on death") estimated on synthetic data generated by `CTGAN` is smaller than on the observed data, as `CTGAN` delivers a biased null effect due to generative model misspecification.

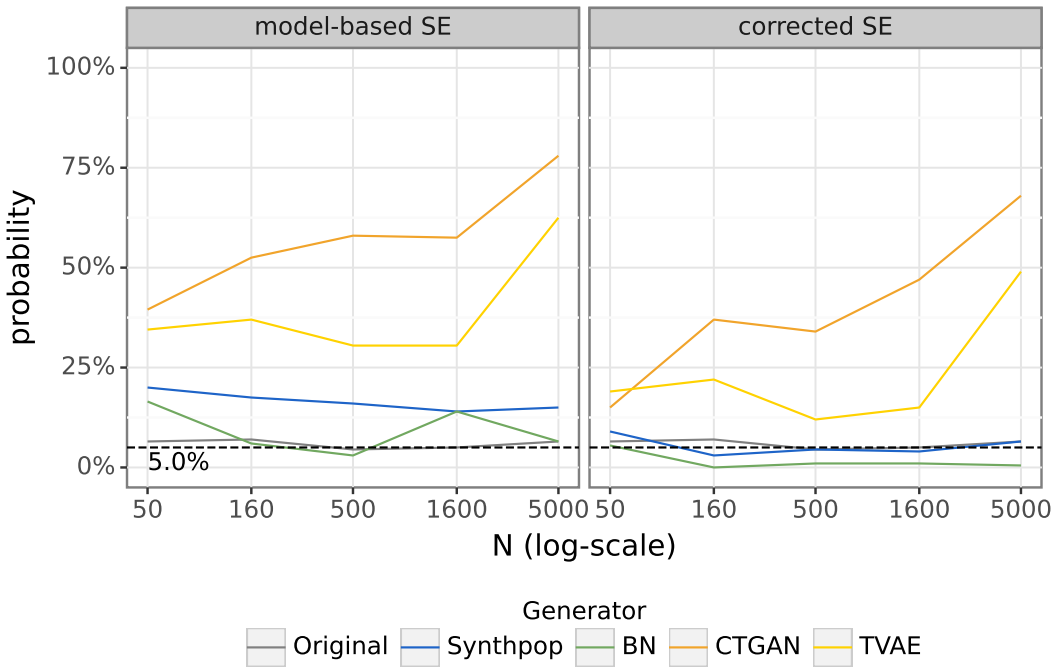

Figure A3: Type 1 error rate of the one-sample t-test at $\alpha = 5\%$ for the population mean of age with naive model-based and corrected SEs. The null hypothesis states that the population mean of age is equal to the ground truth.