

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Practical Applications of Synthetic Data Generation

by Lisa Pilgram and Khaled El Emam

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Practical Applications of Synthetic Data Generation

Lisa Pilgram and Khaled El Emam¹

Abstract

Synthetic data generation (SDG) is increasingly applied across sectors for privacy-preserving data sharing, de-biasing and augmentation. Each use case requires a distinct set of evaluation metrics that must account for the stochasticity of the SDG process: membership and attribute disclosure vulnerability are critical for privacy; fidelity and downstream task utility apply more broadly; and fairness and diversity are relevant for de-biasing and augmentation, respectively. Presenting accumulated evidence and through exemplar case studies, it is shown that SDG can perform well across many of these use cases and key learnings from our experiences with synthetic health data are shared.

Key Words: Synthetic data; Privacy; Data sharing; Bias.

1. Introduction

As demand for data has grown, the challenges with data access have also become more acute, especially for data containing personally identifying information. To enable broader access to this kind of data, it is necessary to address privacy concerns. Synthetic data generation (SDG) has gained broader importance as a method for privacy-preserving sharing of data across various sectors (El Emam and Hoptroff 2020; van Breugel et al. 2024). Synthetic data is fake data that mirrors real joint distributions but, if generated properly, does not contain any real personal data.

While there is still some uncertainty around the regulatory environment for synthetic data, national statistical agencies have been considering the use of synthetic data for sharing their data products (United Nations Economic Commission for Europe 2022), and some large data custodians have shared synthetic data publicly such as the CMS Data Entrepreneur's Synthetic Public Use files (U.S. Centers for Medicare & Medicaid Services 2022), cancer data from Public Health England (National Disease Registration Service 2018), synthetic variants of the French public health system claims and hospital dataset (Health Data Hub France 2021), and synthetic microdata from Israel's National Registry of Live Births (Hod and Canetti 2024).

There are three major use cases that synthetic data can then be exploited for: (1) privacy-preserving data sharing, (2) de-biasing, and (3) data augmentation (Jordon et al. 2022). De-biasing refers to the generation of synthetic data to compensate for representation bias (such as racial or gender bias in data). Data augmentation is helpful for analysis tasks when existing datasets have limited sample size. That is often the case in the health care or humanitarian sectors. The privacy-preserving use case of synthetic data is the most prevalent one, as shown for example in the health care literature (Kaabachi et al. 2023).

In this article we give a brief overview of current SDG practices and present insights gained from developing and using synthetic data in its privacy-preserving and de-biasing use cases.

1.1 Synthetic Data Generators

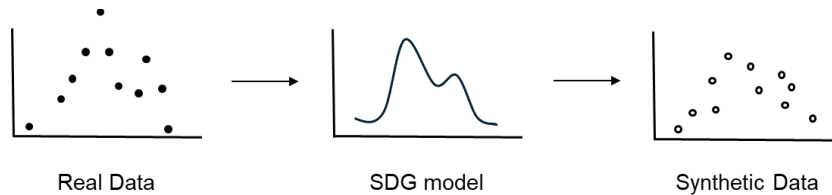
Synthetic data is a general term and refers to a large variety of methods, depending on how it is generated and what it is trying to mimic. SDG in general does not necessarily require real data; it can be based on distributions known a-priori and informed by background knowledge, published summary statistics, and published risk calculators (Templ

¹Faculty of Medicine, University of Ottawa & CHEO Research Institute, Ottawa, Ontario; lpilgram@ehealthinformation.ca and kelemam@ehealthinformatio.ca

et al. 2017; Walonoski et al. 2018; Jeanson et al. 2024; Al-Dhamari et al. 2024). In this paper, however, we are concerned about SDG that involves the training on real individual-level data to learn the joint distribution that is then used for generation.

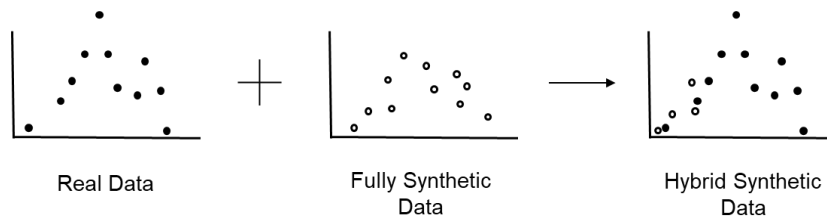
Individual-level data driven methods utilize machine learning or deep learning methods, such as sequential decision trees (Drechsler 2011; Nowok 2015), as well as generative adversarial networks which have been one of the more used types of generative models in research and practice (Hernandez et al. 2022; Ghosheh et al. 2024), and have been applied often for the synthesis of health data. The process of SDG is illustrated in a simplified way in Figure 1.1-1.

Figure 1.1-1
Conceptualization of Synthetic Data Generation



The generated data can be fully synthetic, partially synthetic, and hybrid synthetic (Little 1993; Rubin 1993; Surendra and MohanH 2017). Fully synthetic data refers to an entirely synthetic dataset; partially synthetic to a dataset where certain columns are synthesized and hybrid synthetic data is data that contains both, real and synthetic records. The insights and evidence we present here center around fully synthetic (i.e., the privacy use case) and hybrid data (i.e., the de-biasing use case). The concept of hybrid data is shown in Figure 1.1-2. Unless otherwise stated, we will use the term “synthetic data” to mean fully synthetic structured tabular data only.

Figure 1.1-2
Conceptualization of Hybrid Synthetic Data for De-Biasing



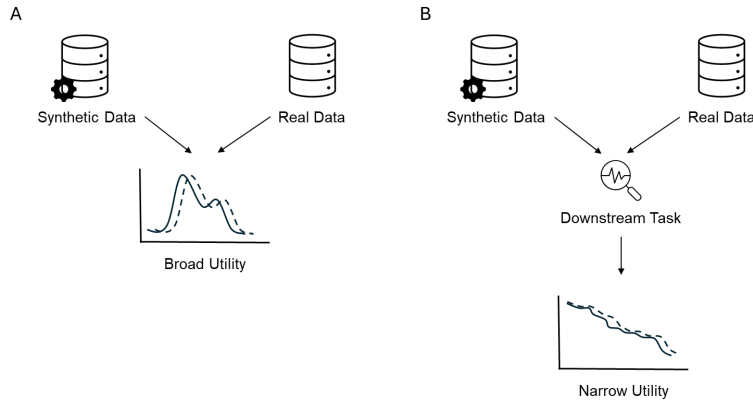
Synthetic data is evaluated using utility and privacy metrics. The use case for the synthetic data ultimately dictates the way it is evaluated. While utility is relevant for all use cases, privacy is not relevant in de-biasing or data augmentation.

1.2 Evaluating the Utility of Synthetic Data

Utility can be defined as how good the synthetic data is for the specific use case. Therefore, the manner in which utility can be evaluated does not necessarily involve the same metrics across all the use cases; there are multiple ways to measure utility (El Emam 2020; El Emam et al. 2022b; Kaabachi et al. 2023; Vallevik et al. 2024; Budu et al. 2024).

In the privacy use case, metrics can be split into broad and narrow utility with broad metrics assessing the resemblance or similarity of the synthetic data to the real data (i.e., fidelity) and narrow utility assessing its functionality (i.e., downstream utility). Fidelity does not require knowledge of how the synthetic data will be used, however, assessing narrow utility is linked to a specific downstream task, for example replicating a survival analysis for cancer patients. This is shown in Figure 1.2-1.

Figure 1.2-1
Broad (A) and Narrow Utility (B) of Synthetic Data



Broad utility can be measured on a univariate or multivariate level. Univariate fidelity metrics, however, give insufficient information as they ignore multivariate relationships. Multivariate metrics, in contrast, can assess joint distributions and dependencies at once which is required when drawing conclusions for the entire dataset. Often, it is desired that broad metrics reflect or are predictive of narrow utility, meaning that it can serve as a proxy for the downstream task. Evidence suggests that propensity score matching, cluster analysis and the multivariate Hellinger distance are particularly effective in capturing the properties of the data that are relevant, for example, in downstream classification tasks (El Emam et al. 2022b). Directly measuring narrow utility is often not a practical approach as actual or representative downstream tasks may not be known *a priori* or may be too computationally demanding in generative model optimization processes.

Yet, narrow utility remains a standard in methodology research. It gives valuable insights into the functionality and effectiveness of synthetic data in practice. In the end, narrow rather than broad utility matters most. When the downstream task involves inference, then the question is whether the same insights and conclusions can be drawn from synthetic data as from real one. To assess this, metrics of replicability are relevant. Table 1.2-1 shows the four metrics that have been used to give a comprehensive picture of replicability (El Emam et al. 2024).

Table 1.2-1
Definitions of Replicability for Inferential Tasks

Metric	Explanation
Decision agreement	Whether estimates have the same direction and statistical significance as the real data.*
Estimate agreement	Whether the synthetic estimates are within the 95% CI of the real data.
Standardized difference	Whether the difference in the estimates is consistent with the null hypothesis of no difference.
CI overlap	The proportion of the overlap between the synthetic and real 95% CI.

*not applicable to descriptive downstream tasks

When the downstream task is prediction rather than inference, then the approach “train on synthetic, test on real” (TSTR) provides a measure of narrow utility (Hyland et al. 2017). TSTR uses a real holdout dataset to assess the prediction performance of say, a machine learning (ML) model trained on synthetic data. A low TSTR performance indicates that the model is not generalizing well to unseen real data. Consequently, the interpretation of TSTR does not necessarily require the comparison to the performance when the model is trained on real data (TRTR). Such a comparison, however, helps to understand the underlying reasons for a low TSTR performance. If the TRTR performance is high, then the SDG failed in generating high utility synthetic data. If TRTR is low, then it is not about SDG but about the real training data that is not suitable for the downstream task. And if the TSTR performance is

higher than TRTR then that may indicate that the SDG process is improving the diversity of the dataset and its ability to generalize to unseen data.

The utility evaluations mentioned above are mainly concerned with synthetic data under the privacy use case where the goal is to have a proxy for the real data. Data augmentation and de-biasing require further evaluations.

In the case of data augmentation, the goal is to generate more diverse records that can then be added to the real data. Diversity metrics can be applied to assess the extent to which SDG increases diversity (Sajjadi et al. 2018; Alaa et al. 2022). The expectation is that the more diverse dataset will be used to train models with better prognostic performance on unseen data (i.e., lower generalization error). These can be considered as broad metrics for the augmentation use case. Narrow metrics for augmentation can be defined for prognostic tasks in a manner that is similar to TSTR.

In the case of de-biasing, fairness needs to be assessed (Dwork et al. 2012; Hardt et al. 2016; Yan et al. 2020; Juwara et al. 2024; Vallevik et al. 2024). Bias can be measured, for example, by statistical parity difference (SPD). SPD compares the probability of the outcome between two groups that might be affected by bias. Equal opportunity (EOD) difference compares their true-positive rate, and average odds differences (AOD) extends the EOD by the false positive rate (Dwork et al. 2012; Hardt et al. 2016; Yan et al. 2020). Using these as standalone metrics, however, can be misleading as the ground truth disparity is assumed to be zero. In healthcare research, for example, if the outcome is a diagnosis, then a certain gender disparity is expected for a large number of diseases. This means that the data analyst must have some knowledge about the ground truth disparity to be able to interpret the bias that is present in their data, and consequently the bias that is present in the de-biased data after SDG.

1.3 Privacy of Synthetic Data

If during SDG, the generative model is well trained in the sense that it does not overfit or memorize, then there is no one-to-one mapping between the synthetic data and the real data. Yet, if the generative model overfits or memorizes (some of) the real data, then privacy disclosure may occur in the resulting synthetic data. This makes a privacy assessment mandatory in any privacy use case, but not in data augmentation or de-biasing use cases.

Most of the privacy literature relates to three privacy disclosure concepts: identity disclosure, membership disclosure and attribute disclosure (Kaabachi et al. 2023; Boudewijn et al. 2023; Vallevik et al. 2024; Budu et al. 2024). Identity disclosure is when an individual's identity can be assigned to a record; membership disclosure is when an individual's membership in the dataset can be inferred; and attribute disclosure is when an individual's sensitive information can be inferred from the attributes of a dataset.

Given that throughout SDG the link to a real record is not preserved, identity disclosure should be protected *by design*, and it is challenging to adapt metrics of identity disclosure to synthetic data to account for this inherent process. Consequently, membership and attribute disclosure are seen as the key vulnerabilities to be evaluated in synthetic data. They are both inferences. Correct inferences, however, are not privacy invasive *per se* but can occur independent of any individual's data disclosed. For example, a model can be trained on the synthetic data that predicts survival for cancer patients. If an individual's record is not in the training data, then the survival prediction is independent of that individual and can be seen as general knowledge rather than as compromising the privacy of that individual.

The assessment of membership disclosure vulnerability in synthetic data generally tries to mimic an adversarial attack and requires thoughtful consideration of the adversarial assumptions. For example, if the assumption is that the adversary draws random targets from the same population as the training data was sampled from, then the mimicked sample of targets used in the evaluation must align with the proportion of members and non-members in the population (El Emam et al. 2022a).

Attribute disclosure is a prediction task where a model is trained to predict the sensitive information about a target based on the synthetic data. For example, if the sensitive information is categorical, then a classification model is trained on the synthetic data. Its prediction performance is evaluated for the SDG training data to decide whether the prediction is meaningful for individuals who could potentially experience privacy violation due to being part of the SDG training data. An Area Under the Receiver Operating Characteristics Curve (AUROC) of 0.2, for example, would not provide an adversary with correct inferred information. Generally, an AUROC lower than or equal to 0.7 or 0.6 is

considered as poor accuracy (Hond et al. 2022). Once meaningful prediction performance is confirmed, it should then be compared against disclosure-independent inference (i.e., knowledge generation) to assess the vulnerability of the synthetic dataset. Such a baseline can be established using a holdout dataset (Giomi et al. 2023; Francis and Wagner 2024).

This privacy assessment continues to be required even when differentially private SDG is applied. This is because the privacy budget in differential privacy drives vulnerability but the extent remains unclear except for budgets close to 0 (Dwork et al. 2019). In practice, privacy budgets are typically larger. For example, the US Census Bureau implemented differential privacy with a relatively high budget of 18.19 (Abowd et al. 2022).

1.4 The Stochasticity of SDG

Thinking of SDG as a form of multiple imputation, the same considerations in terms of its randomness apply. Consequently, multiple synthetic datasets and corresponding combining rules can increase robustness of results. This is recognized by various privacy metrics for synthetic data (Elliot 2015; Taub et al. 2018; Stadler et al. 2020), and was comprehensively evaluated for inferential replicability where a plateau was reached at 10 iterations (El Emam et al. 2024). This is particularly relevant in methodological research where a robust estimate of the privacy and utility of an SDG model is necessary when, for example, benchmarking multiple models. For a data-release scenario, it is more relevant to measure the privacy for the specific synthetic dataset(s).

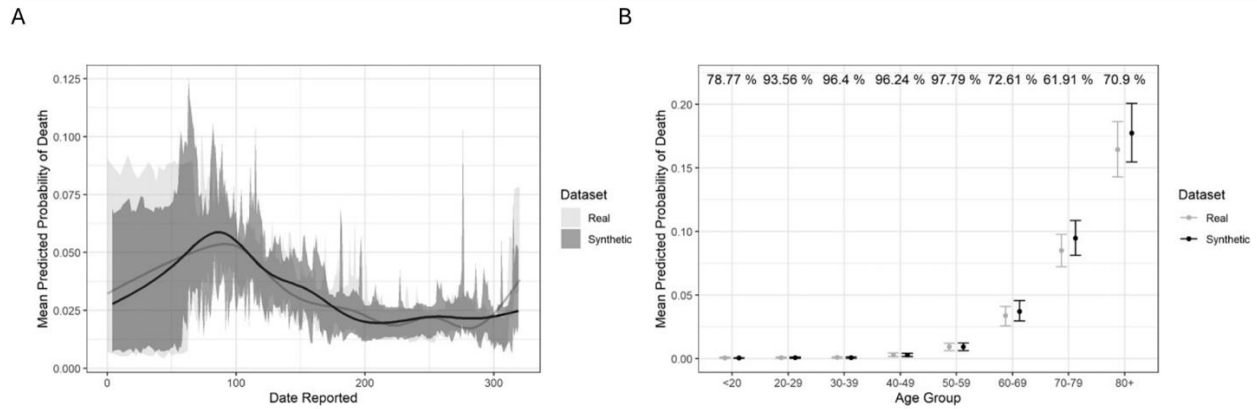
2. Synthetic Data Generation to Preserve Privacy

In this section we present two representative examples where synthetic data was generated for privacy use cases.

2.1 Synthetic COVID-19 Cases

The COVID-19 pandemic highlighted the importance of data sharing among researchers while at the same time revealing significant barriers associated with getting access to data. Barriers to data access are primarily driven by privacy concerns. SDG can serve as an enabler to share data in such a global health emergency. In (El Emam et al. 2021), the potential of SDG to enable privacy-preserving sharing of COVID-19 data while maintaining utility was evaluated. The study used 90,514 COVID-19 cases from Ontario. Synthetic data was generated using sequential classification and regression trees. Membership and attribute disclosure vulnerability were assessed, and downstream utility was defined as mortality prediction. The results of a binary classification model using gradient boosted decision trees demonstrated that the prediction performance did only slightly differ from the real dataset (AUROC 0.945 using real data vs. AUROC 0.940 using synthetic data). The model built on synthetic data had the same high importance predictors (i.e., age and date) as the real one (see Figure 2.1-1). At the same time privacy vulnerability was reported to be low: synthetic data successfully served as a privacy-preserving proxy for the real data.

Figure 2.1-1
Probability of Death by Date (A) and Probability by Death by Age (B) Derived from Synthetic Data. From (El Emam et al. 2021).



2.2 Synthetic Cancer Clinical Trial Data

In biopharma, sharing data from clinical trials is meant to ensure transparency and reproducibility and is required, for example, under Health Canada transparency policies as part of market authorization. Data sharing also offers a way to conduct research for secondary purposes with valuable insights on, for example, drug safety while easing the burden on patients. SDG is therefore a tool with high potential benefit for this industry. In (El Kababji et al. 2023), synthetic datasets based on eight breast cancer clinical trials were assessed for their ability to replicate the analysis of the real data. Three different SDG models were compared: sequential tree-based synthesis, GANs and variational autoencoders. Utility among the SDG models varied depending on the clinical trial, but sequential tree-based synthesis reliably resulted in, for example, high estimate and decision agreement. Privacy vulnerability was low across all synthetic datasets. This study revealed that the choice of SDG model matters and will not be consistent across different datasets. It also contributed to the growing evidence that synthetic data can satisfy both, privacy and utility.

3. Synthetic Data Generation to Mitigate Bias

The idea of mitigating bias via SDG is to generate additional records from the under-represented group to augment their representation. This can be achieved through conditional generation or rejection sampling after the fact. In biomedical research, representation bias is a common and serious concern. It results in a dataset that does not adequately reflect the distribution of the patient population and ultimately to analyses that lack external validity. In (Juwara et al. 2024), the authors used SDG as a de-biasing method was tested on four health care datasets. As a first step, the different types of bias were introduced and then the de-biased dataset evaluated against the initial dataset (i.e., ground truth) as described above. The analysis concluded that through the augmentation with SDG, prediction performance could be maintained for minority groups in low to medium bias situations, but not consistently in high bias situations.

4. Main Observations and Limitations

Evidence is accumulating that SDG is able to capture relevant and quite complex patterns in the real data and can thereby serve as a good proxy for real data in a privacy use case, or as a supplement in the de-biasing use case. We can summarize our learnings, mostly with tabular healthcare data, as follows:

- (1) Training a generative model depends on the size (number of observations and variables) of the source dataset. If the number of observations is small then there is a risk of overfitting, but that risk is diluted if the number of variables is small. Overfitting can be detrimental to privacy but can be beneficial for utility (e.g., fidelity and replicability). There are no sample size calculators for generative models, but good results on both privacy and utility have been obtained on small datasets (e.g., small clinical trials).
- (2) It is not necessary to know how the synthetic data will be analyzed to train the generative models. There is evidence that generative models capture many of the relevant patterns in the source data for typical analysis tasks.
- (3) Preprocessing of the real data is crucial and can have a non-trivial impact on the performance of the generative models. This means that the implementation of a generative model will matter and not all implementations are equally effective, even of the same type of generative model.
- (4) There is not one superior type of generative model across all datasets but rather there is dataset-specific superiority.
- (5) When the analytic workload is estimating a parameter and its confidence interval, then using the combining rules to calculate the value from 10 synthetic data is necessary to avoid invalid estimates and inferences.
- (6) Metrics for evaluating different types of utility (fidelity and replicability) and privacy (membership and attribute disclosure) have been developed by the community. Consensus benchmarks for what is acceptable are still under development.
- (7) Synthetic data can help improve the fairness and accuracy of prognostic and inferential models when representation bias in observational data is low to medium. It is more challenging to mitigate the impacts of representation bias consistently in real world datasets when bias is high.

References

- Abowd, J.M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P., (2022), "The 2020 Census Disclosure Avoidance System TopDown Algorithm", *Harvard Data Science Review*. Paper available at <https://doi.org/10.1162/99608f92.529e3cb9>.
- Alaa, A., Breugel, B.V., Saveliev, E.S., Schaar, and M. van der (2022), "How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models", *Proceedings of the 39th International Conference on Machine Learning*, PMLR, pp 290–306.
- Al-Dhamari, I., Abu Attieh, H., and Prasser, F. (2024), "Synthetic datasets for open software development in rare disease research", *Orphanet J Rare Dis* 19:265. Paper available at <https://doi.org/10.1186/s13023-024-03254-2>.
- Boudewijn, A., Ferraris, A.F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., and Chauvenet, C.R. (2023), "Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions", arXiv:2311.17453, <https://doi.org/10.48550/arXiv.2311.17453>.
- Budu, E., Etminani, K., Soliman, A., and Rögnavaldsson, T. (2024), "Evaluation of synthetic electronic health records: A systematic review and experimental assessment", *Neurocomputing* 128253. Paper available at <https://doi.org/10.1016/j.neucom.2024.128253>.
- Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Springer-Verlag, New York.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012), "Fairness through awareness", *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Association for Computing Machinery, New York, NY, USA, pp. 214–226.

- Dwork, C., Kohli, N., and Mulligan, D. (2019), "Differential Privacy in Practice: Expose your Epsilons!", *Journal of Privacy and Confidentiality* 9. Paper available at <https://doi.org/10.29012/jpc.689>.
- El Emam, K. (2020), "Seven Ways to Evaluate the Utility of Synthetic Data", *IEEE Security Privacy* 18:56–59. Paper available at <https://doi.org/10.1109/MSEC.2020.2992821>.
- El Emam, K., and Hoptroff, R. (2020), *Practical Synthetic Data Generation*, O'Reilly, Sebastopol, CA.
- El Emam, K., Mosquera, L., and Fang, X. (2022a), "Validating A Membership Disclosure Metric For Synthetic Health Data", *JAMIA Open* 5:ooac083.
- El Emam, K., Mosquera, L., Fang, X., and El-Hussuna, A. (2022b), "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study", *JMIR Med Inform* 10:e35734. Paper available at <https://doi.org/10.2196/35734>.
- El Emam, K., Mosquera, L., Fang, X., and El-Hussuna, A. (2024), "An evaluation of the replicability of analyses using synthetic health data", *Sci Rep* 14:6978. Paper available at <https://doi.org/10.1038/s41598-024-57207-7>.
- El Emam, K., Mosquera, L., Jonker, E., and Sood, H. (2021). "Evaluating the utility of synthetic COVID-19 case data", *JAMIA Open* 4:ooab012. Paper available at <https://doi.org/10.1093/jamiaopen/ooab012>.
- El Kababji, S., Mitsakakis, N., Fang, X., Beltran-Bless, A., Pond, G., Vandermeer, L., Radhakrishnan, D., Mosquera, L., Paterson, A., Shepherd, L., Chen, B., Barlow, W.E., Gralow, J., Savard, M.-F., Clemons, M. and El Emam, K. et al (2023), "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets", *JCO Clin Cancer Inform* e2300116. Paper available at <https://doi.org/10.1200/CCI.23.00116>.
- Elliot, M. (2015), Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team, Manchester University.
- Francis, P., Wagner, D. (2024), "Towards more accurate and useful data anonymity vulnerability measures", arXiv:2403.06595 [cs.CR].
- Ghosheh, G.O., Li, J., and Zhu, T. (2024), "A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records", *ACM Comput Surv* 56:147:1-147:34. Paper available at <https://doi.org/10.1145/3636424>.
- Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B. (2023), "A Unified Framework for Quantifying Privacy Risk in Synthetic Data", *Proceedings on Privacy Enhancing Technologies*.
- Hardt, M., Price, E., and Srebro, N. (2016), "Equality of Opportunity in Supervised Learning", arXiv:1610.02413 [cs.LG].
- Health Data Hub France (2021), "SNDS synthétiques", *Système national des données de santé*. https://documentation-snds.health-data-hub.fr/formation_snds/donnees_synthetiques/. Accessed: 2022-01-20.
- Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., and Rankin, D. (2022), "Synthetic data generation for tabular health records: A systematic review", *Neurocomputing*, 493:28–45. DOI:10.1016/j.neucom.2022.04.053.

- Hod, S., and Canetti, R. (2024), "Differentially Private Release of Israel's National Registry of Live Births", arXiv.org. Paper available at <https://arxiv.org/abs/2405.00267v1>. Accessed: 2024-09-28.
- Hond, A.A.H. de, Steyerberg, E.W., and Calster, B. van (2022), "Interpreting area under the receiver operating characteristic curve", *The Lancet Digital Health* 4:e853–e855. Paper available at [https://doi.org/10.1016/S2589-7500\(22\)00188-1](https://doi.org/10.1016/S2589-7500(22)00188-1).
- Hyland, S.L., Esteban, C., and Rättsch, G. (2017), "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs", arXiv:170602633 [cs, stat]
- Jeanson, F., Farkouh, M.E., Godoy, L.C., Minha, S., Tzuman, O., and Marcus, G. (2024), "Medical calculators derived synthetic cohorts: a novel method for generating synthetic patient data", *Sci Rep* 14:11437. Paper available at <https://doi.org/10.1038/s41598-024-61721-z>.
- Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., and Weller, A. (2022), "Synthetic Data -- what, why and how?", arXiv:2205.03257 [cs.LG].
- Juwara, L., El-Hussuna, A., and El Emam, K. (2024), "An evaluation of synthetic data augmentation for mitigating covariate bias in health data", *Patterns*, Paper available at <https://doi.org/10.1016/j.patter.2024.100946>.
- Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., and Raisaro, J.-L. (2023), "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility", *Metrics*, 2023.11.28.23299124.
- Little, R. (1993), "Statistical Analysis of Masked Data", *Journal of Official Statistics*, 9:407–426.
- National Disease Registration Service (2018), The Simulacrum, In: The Simulacrum. <https://simulacrum.healthdatainsight.org.uk/>. Accessed 2021-11-27.
- Nowok, B. (2015), *Utility of synthetic microdata generated using tree-based methods*, Helsinki.
- Rubin, D. (1993), "Discussion: Statistical Disclosure Limitation", *Journal of Official Statistics*, 9:462–468.
- Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018), "Assessing Generative Models via Precision and Recall", arXiv:180600035 [cs, stat].
- Stadler, T., Oprisanu, B., and Troncoso, C. (2020), "Synthetic Data -- A Privacy Mirage", arXiv:201107018 [cs].
- Surendra, H., and Mohan, H. S., (2017), "A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing", *International Journal of Scientific & Technology Research*
- Taub, J., Elliot, M., Pampaka, M., and Smith, D. (2018), "Differential Correct Attribution Probability for Synthetic Data: An Exploration", In: Domingo-Ferrer J, Montes F (eds) *Privacy in Statistical Databases*, Springer International Publishing, Cham, pp 122–137.
- Templ, M., Meindl, B., Kowarik, A., and Dupriez, O. (2017), "Simulation of Synthetic Complex Data: The R Package simPop", *Journal of Statistical Software*, 79:1–38. Paper available at <https://doi.org/10.18637/jss.v079.i10>.

- United Nations Economic Commission for Europe (2022), "Synthetic Data for Official Statistics - A Starter Guide", <https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide>. Accessed: 2023-01-24.
- U.S. Centers for Medicare & Medicaid Services (2022), CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF), https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF. Accessed: 2022-07-17.
- Vallevik, V.B., Babic, A., Marshall, S.E., Elvatun, S., Brøgger, H.M.B., Alagaratnam, S., Edwin, B., Veeraragavan, N.R., Befring, A.K., and Nygård, J.F. (2024), "Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare", *International Journal of Medical Informatics*, 185:105413. Paper available at <https://doi.org/10.1016/j.ijmedinf.2024.105413>.
- van Breugel, B., Liu, T., Oglic, D., and van der Schaar, M. (2024), "Synthetic data in biomedicine via generative artificial intelligence", *Nature Reviews Bioengineering*, 1–14.
- Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and McLachlan, S. (2018), "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record", *J Am Med Inform Assoc*, 25:230–238. Paper available at <https://doi.org/10.1093/jamia/ocx079>.
- Yan, S., Kao, and H., Ferrara, E. (2020), "Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes", *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Association for Computing Machinery, New York, NY, USA, pp. 1715–1724.