Evaluation Methods for Synthetic Data in Pursuit of Open Data

Bing Hu[1] **Mohammad Ahmed Basri**[2] **Abu Yousuf Md Abdullah**[3] **Shu-Feng Tsao**[4] **Zahid Butt**[4] **Helen Chen**[1,4,5]
[1]Cheriton School of Computer Science, [2]Department of Systems Design Engineering, [3]School of Planning
[4]School of Public Health Sciences, [5]Department of Statistics and Actuarial Science
University of Waterloo
{bingxu.hu, mabasri, aymabdullah, s7tsao, zahid.butt, helen.chen}@uwaterloo.ca

Abstract

Real data containing sensitive or personal data often requires lengthy approval processes and stringent restrictions for access. Synthetic data that resembles the real data and is generated from the real data following findable, accessible, interoperable, and reusable (FAIR) standards is a promising approach to open data for administrative data. Although progress has been made in establishing accepted evaluations for synthetic data models, missing are key holistic metrics for policymakers to aid their decision-making on open data initiatives. In this paper, we introduce and demonstrate a privacy risk with an identity disclosure risk (IDR) assessment, a quantitative measure of univariate distribution in Hellinger distance (HD), and a quantitative bivariate measure of differential pairwise correlation (DPC). By including our introduced privacy, univariate, and bivariate metrics in standard synthetic data evaluation, synthetic data models and methods can be better understood and utilized by policymakers in pursuit of open data.

1 Introduction

In many cases, the challenge for AI researchers and developers is not a shortage of data but instead getting access to data. Administrative data collected by public services such as health, education, or employment are increasingly being used by researchers, developers, and policymakers for the benefit of society [1]. As administrative data collected by public services often contain sensitive or personal data, lengthy approval processes, and restrictions are in place to access the de-identified data securely [1, 2]. As accessing administrative data for research has complex requirements and takes time, administrative data has become a critical component of open data initiatives [3–5].

Synthetic data is an approach to open data that can transform highly restrictive administrative data to instead be findable, accessible, interoperable, and reusable (FAIR standards) [6, 7]. Advancements in deep learning (DL) synthetic data generation such as CTGAN [8] and REalTabFormer [9] have enabled the pursuit of open data through the creation of synthetic administrative data. Although progress has been made in establishing accepted benchmarks and metrics for evaluating competing synthetic generation DL models such as machine learning efficiency (MLE), distribution plots, and discriminator measures [9–13], these accepted benchmarks and metrics are not designed to inform policymakers nor are motivated by the pursuit of open data initiatives [14]. Holistic measures on the utility, the suitability of the dataset for a task, quality, the integrity and completeness of the dataset, and the risks and recommendations are necessary metrics for policymakers in pursuit of open data initiatives [15]. Still missing from the accepted benchmark and evaluation of synthetic data methods in DL literature are measures of privacy risk, a quantitative measure of distribution, and a quantitative measure of correlations, that provide additional meaningful context to policymakers [14].

In this paper, we demonstrate additional measures to the evaluation of synthetic data DL methods in terms of privacy risk with an identity disclosure risk assessment (IDR) [16], a quantitative measure of distribution in Hellinger distance (HD) [17], and a quantitative bivariate measure of differential pairwise correlation (DPC) [18]. We demonstrate these additional metrics by training, generating, and evaluating synthetic data from a simulated real health dataset based on MIMIC-IV [7] using DL methods CTGAN [8] and REalTabFormer (RTF) [9]. We hope that with the measures introduced in this paper, synthetic data DL research can be better assessed and understood by policymakers in the pursuit of open data initiatives.

2 Methods

2.1 Simulated Real Dataset

A real health dataset is simulated from MIMIC-IV to be used to generate synthetic data. Fields of ethnicity, gender, death, religion, marital status, insurance, and age are sampled from MIMIC-IV to create a profile for each patient. Additional binary flags for select diagnoses of sepsis, birth, chest pain, hypertension, and overdose are recorded for each patient over all their admissions. The simulated real dataset contains 58,977 rows of patients.

2.2 Models

2.2.1 CTGAN

The base CTGAN model [8] is used with no modification. After training for 75 epochs, 58,977 synthetic data points are sampled from the model. Default learning rates of 2e-4 are used for both the generator and discriminator. Default decay rates of 1e-6 are used for both the generator and discriminator. A batch size of 500 is used during training.

2.2.2 REalTabFormer

The base REalTabFormer (RTF) model [9] is used with no modifications. After training for 100 epochs, 58,977 synthetic data points are sampled from the model. A batch size of 256 is used during training. Default hyperparameters other than epochs and batch size are used during training.

2.3 Additional Measures

2.3.1 Identity Disclosure Risk (IDR) Assessment

Identity disclosure risk (IDR) [16] is a framework to evaluate the privacy and re-identification risk of the generated synthetic data. Concerns about privacy are a key obstacle to the adoption of Open Data initiatives by organizations [19, 20]. Policymakers require measures of privacy risk such as IDR to assess privacy concerns of synthetically generated data for purposes of Open Data.

Identity disclosure risk takes into consideration privacy attacks defined as identity disclosure as well as attribution risk. Identity disclosure is the risk of correctly mapping a synthetic record to a real person and vice versa [16]. Attribution risk, conditional to identity disclosure, is defined as an adversary learning a certain characteristic about a real person [16]. The risk score can be simplified to two parts: Real-to-Synthetic Identification Risk, and Synthetic-to-Real identification Risk. The maximum of both of these risks is taken to be the overall risk of the synthetic dataset. Under the guidance of the European Medicines Agency (EMA) and Health Canada, an acceptable risk threshold of 0.09 is used [21]. The IDR risk is expressed in eq. 1 [16].

$$IDR = max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s} * I_s\right), \frac{1}{n}\sum_{s=1}^{n}\left(\frac{1}{F_s} * I_s\right)\right) \quad (1)$$

N, n is the number of records in the real dataset and synthetic datasets respectively, F_s, f_s is the size of the set of records with the same quasi-identifier values as record s in the real data and synthetic data respectively, and I_s is the binary indicator of whether a record s in the real data matches a record in the synthetic data. Quasi-identifiers are defined as a subset of variables that are known by an adversary

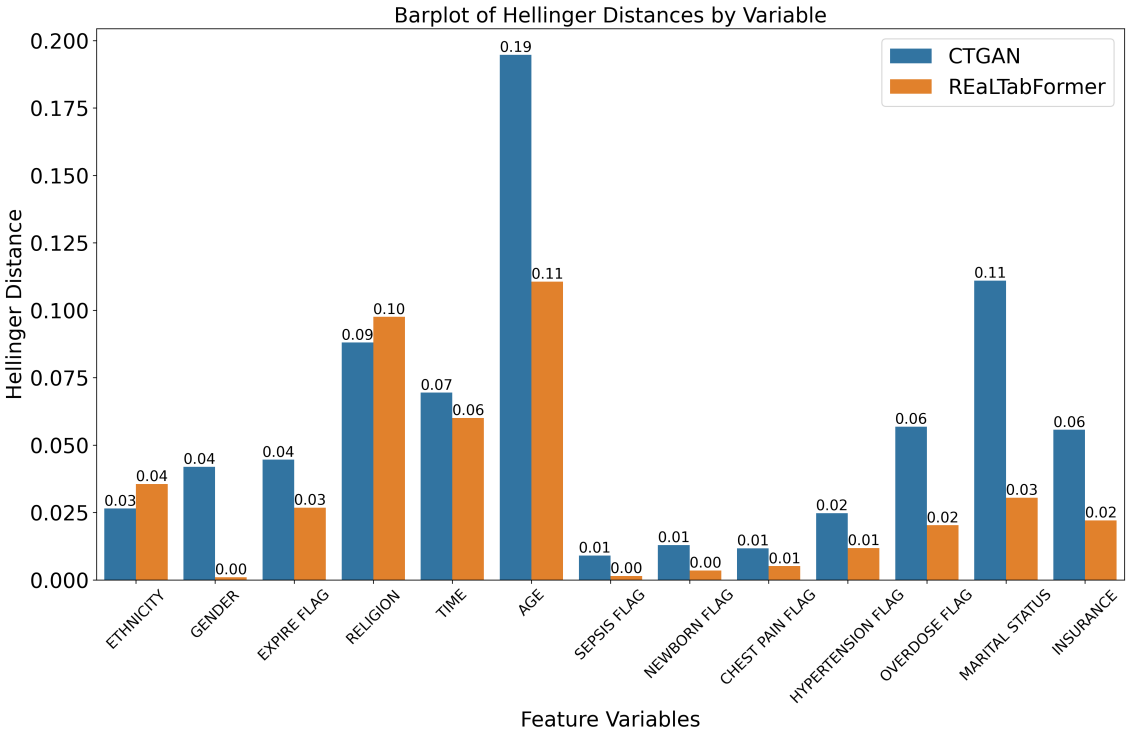

Fig. 1: Hellinger Distance for CTGAN and REalTabFormer (RTF) generated synthetic data over all data fields.

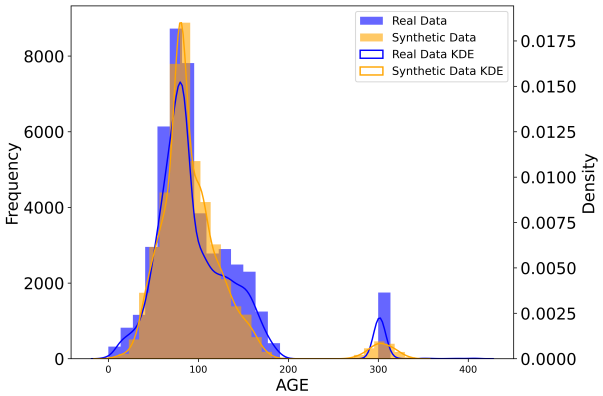

Fig. 2: Distributions of Age comparing synthetic data generated from CTGAN to the simulated real data.

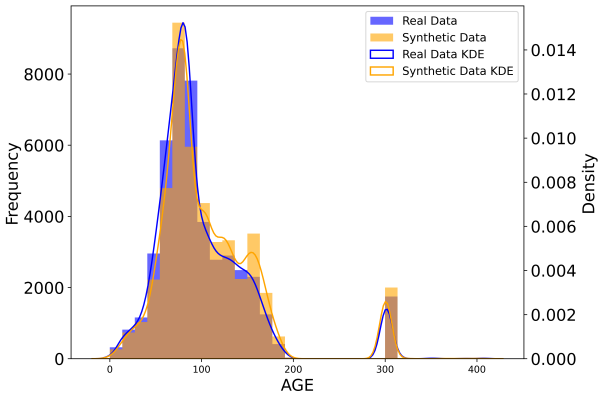

Fig. 3: Distributions of Age comparing synthetic data generated from RTF to the simulated real data.

[22]. The Synthetic-to-Real identification risk is the first sum in Eq. 1 while the Real-to-Synthetic risk is the latter. For the IDR computations completed in this paper, ethnicity, gender, death, religion, and marital status are considered quasi-identifiers that an adversary knows.

2.3.2 Hellinger Distance (HD)

Hellinger distance (HD) quantifies the similarity between two probability distributions [17]. Quantifiable and standard measures such as HD provide open data policymakers additional context alongside visual comparisons of real and synthetic data probability distributions. Given two discrete probability distributions $P = \{p_1, p_2, ..., p_n\}$ and $Q = \{q_1, q_2, ..., q_n\}$, the HD between P and Q is expressed in eq. 2.

$$HD^2(p,q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i}) \tag{2}$$

HD provides a summary statistic of differences between each variable in the real and synthetic datasets. HD scores range between 0 to 1, where values closer to 0 are desired as they indicate lower differences in the distribution between real and synthetic datasets [22].

After calculating the HD for each variable for the real and synthetic datasets, we carried out an overall assessment of the HD for all variables, the median and the interquartile range for the real and synthetic data were computed and assessed to check their proximity to 0. A high-utility dataset should have an overall average HD score closer to 0 [22].

2.3.3 Differential Pairwise Correlation

Synthetic data that closely resembles real data should have similar bivariate pairwise correlations. In combination with the univariate HD metric, DPC provides a bivariate metric for open data policymakers to utilize as a standard to better compare synthetic datasets. If the real and synthetic datasets had high fidelity (i.e., the synthetic dataset closely resembled the real dataset), then the absolute difference would be close to 0 or very small.

For any fields containing continuous variables, the differential pairwise correlations in the real and synthetic data were evaluated to obtain fidelity in terms of bivariate statistics as shown in eq. 3.

$$\Delta CV_{continuous_{XY}} = |\rho_{XY_{real}} - \rho_{XY_{synthetic}}| \tag{3}$$

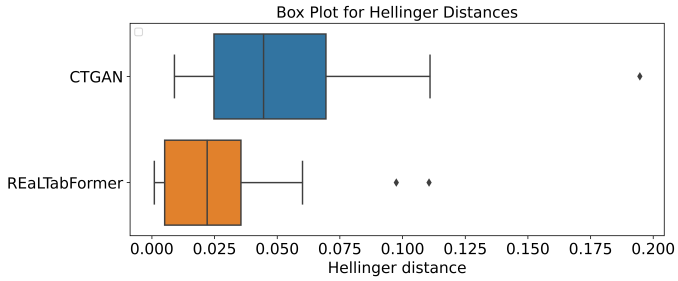

Fig. 4: Boxplot of Hellinger Distances for CTGAN and REalTabFormer generated synthetic data.

Table 1: Evaluations of Synthetic-to-Real (S-R) IDR, Real-to-Synthetic (R-S) IDR, average HD, and average DPC for synthetic data generated by CTGAN and REalTabFormer.

Model	S-R IDR	R-S IDR	Avg. HD	Avg. DPC
CTGAN	0.005	0.013	0.056 ± 0.048	0.029 ± 0.029
RTF	0.004	0.020	0.033 ± 0.034	0.014 ± 0.012

In eq. 3, X and Y denote the two continuous variables, whereas ρ_{XY} is the Pearson correlation coefficient for X and Y. The Pearson correlation coefficient is used as it can be well-defined over continuous variables in the real and synthetic data. In contrast, for categorical variables, the absolute differences for Chi-square statistics in the real and synthetic data are evaluated as shown in eq. 4

$$\Delta CV_{categorical_{XY}} = |\chi^2_{XY_{real}} - \chi^2_{XY_{synthetic}}| \tag{4}$$

In eq. 4, X and Y denote the two categorical variables, whereas χ^2_{XY} is the χ^2_{XY} statistic for X and Y. The χ^2 coefficient is used as it is well-defined over categorical variables in the real and synthetic data.

3 Results & Discussions

As discussed in the previous sections, the synthetic data generated using the two algorithms (CTGAN and RTF) are compared using evaluation metrics, which include Hellinger distance (HD), differential pairwise correlation heatmap, and Identity Disclosure Risk (IDR) Assessment scores. Summary statistics of HD, and DPC from Table 1 show that RTF is better able to create synthetic data that resembles the real data compared to CTGAN. The difference in average DPC between CTGAN and RTF is statistically significant with a p-value of $7.8E-4$. The difference in average HD between CTGAN and RTF was not found to be statistically significant. The improved performance of RTF over CTGAN could be due to mode collapse of the CTGAN model [8]. Comparing Figure 2 and Figure 3, we can see that RTF better models distributions of age given non-mode values compared to CTGAN.

Figure 4 shows that the average HD is much lower for RTF compared to CTGAN. Figure 1 shows that RTF outperforms CTGAN for HD in all fields except for ethnicity and religion. Comparing distributions in figure 2 and figure 3 for CTGAN and RTF, HD is able to quantify a 72% improvement in the synthetic data generated by RTF compared to CTGAN for age. In general, we can see that RTF is able to generate data that better univariately resembles the real data compared to CTGAN.

Although the overall IDR is higher for RTF compared to CTGAN (0.02 vs 0.013) (figure 1), given an acceptable privacy risk threshold of 0.09 as deemed by EMA and Health Canada, synthetic data produced by both CTGAN and RTF falls within this acceptable threshold. Both the synthetic data generated by RTF and CTGAN acceptably preserve the privacy of patients in the simulated data.

Comparing DPC for CTGAN and RTF (figure 5 and 6), we can see that RTF is better able to model correlations between fields compared to CTGAN. The improvement in average differential pairwise correlation between CTGAN and RTF is shown in table 1. In general, we can see that RTF is able to generate data that better bivariately resembles the real data compared to CTGAN.

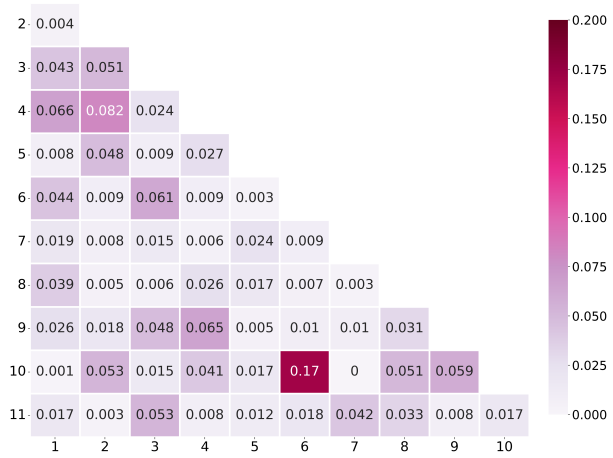

Fig. 5: Differential Pairwise Correlation Heatmap for feature correlation using CTGAN. The legend is as follows: (1) Ethnicity, (2) Gender, (3) Mortality, (4) Religion, (5) Sepsis, (6) Newborn, (7) Chest Pain, (8) Hypertension, (9) Overdose, (10) Marital Status, and (11) Insurance.

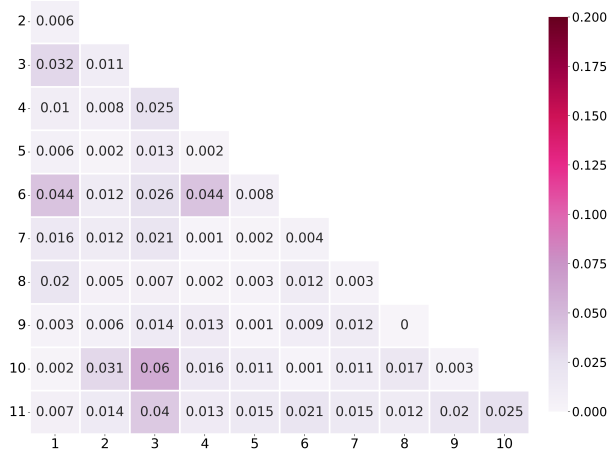

Fig. 6: Differential Pairwise Correlation Heatmap for feature correlation using RealTabformer. The legend is as follows: (1) Ethnicity, (2) Gender, (3) Mortality, (4) Religion, (5) Sepsis, (6) Newborn, (7) Chest Pain, (8) Hypertension, (9) Overdose, (10) Marital Status, and (11) Insurance.

From a policymaker's perspective, synthetic data from both RTF and CTGAN can be released given their acceptable IDR assessment below the 0.09 threshold deemed by Health Canada and EMA. Comparing HD and DPC scores between RTF and CTGAN, the synthetic data generated from RTF better models the real data both univariately and bivariately through pairwise correlations when compared to CTGAN. As an open data policy, based on our analysis, it is more favourable to release synthetic data generated from RTF as open data for the simulated real data compared to the CTGAN-generated synthetic data.

4 Conclusion

In this paper, we introduce and demonstrate additional privacy, univariate, and bivariate synthetic data evaluation metrics for the purpose of accelerating Open Data. Additional evaluations of a privacy metric of identity disclosure risk (IDR), a univariate measure of Hellinger distance (HD), and a bivariate measure of differential pairwise correlations (DPC) can aid policymakers in open data decision-making. By including our introduced privacy, univariate, and bivariate metrics in standard synthetic data evaluation, synthetic data models and methods can be better understood and utilized by policymakers in pursuit of open data.

References

[1] T. Kokosi, B. L. D. Stavola, R. Mitra, L. Frayling, A. R. Doherty, I. Dove, P. Sonnenberg, and K. L. Harron, "An overview of synthetic administrative data for research," *International Journal of Population Data Science*, vol. 7, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249036803

[2] N. Dattani, P. Hardelid, J. Davey, and R. Gilbert, "Accessing electronic administrative health data for research takes time," *Archives of Disease in Childhood*, vol. 98, pp. 391 – 392, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1803595

[3] J.-C. Burgelman, C. Pascu, K. Szkuta, R. Von Schomberg, A. Karalopoulos, K. Repanas, and M. Schouppe, "Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century," *Frontiers in big data*, vol. 2, p. 43, 2019.

[4] K. Armeni, L. Brinkman, R. Carlsson, A. Eerland, R. Fijten, R. Fondberg, V. E. Heininga, S. Heunis, W. Q. Koh, M. Masselink *et al.*, "Towards wide-scale adoption of open science practices: The role of open science communities," *Science and Public Policy*, vol. 48, no. 5, pp. 605–611, 2021.

[5] E. R. Chiware and L. Skelly, "Open science in africa: What policymakers should consider," *Frontiers in Research Metrics and Analytics*, vol. 7, p. 950139, 2022.

[6] S.-F. Tsao, K. Sharma, H. Noor, A. Forster, and H. H. Chen, "Health synthetic data to enable health learning system and innovation: A scoping review," *Studies in health technology and informatics*, vol. 302, pp. 53–57, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258785381

[7] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195767064

[9] A. Solatorio and O. Dupriez, "Realtabformer: Generating realistic relational and tabular data using transformers," *ArXiv*, vol. abs/2302.02041, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256615552

[10] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modelling tabular data with diffusion models," *ArXiv*, vol. abs/2209.15421, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252668788

[11] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE transactions on neural networks and learning systems*, vol. PP, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238353897

[12] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci, "Language models are realistic tabular data generators," *ArXiv*, vol. abs/2210.06280, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252846328

[13] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: a literature review," *Journal of Big Data*, vol. 10, no. 1, p. 115, 2023.

[14] D. G. Gomes, P. Pottier, R. Crystal-Ornelas, E. J. Hudgins, V. Foroughirad, L. L. Sánchez-Reyes, R. Turba, P. A. Martinez, D. Moreau, M. G. Bertram *et al.*, "Why don't we share data and code? perceived barriers and benefits to public archiving practices," *Proceedings of the Royal Society B*, vol. 289, no. 1987, p. 20221113, 2022.

[15] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible ai," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.

[16] K. El Emam, L. Mosquera, and J. Bass, "Evaluating identity disclosure risk in fully synthetic health data: Model development and validation," *J Med Internet Res*, vol. 22, no. 11, p. e23139, Nov 2020. [Online]. Available: http://www.jmir.org/2020/11/e23139/

[17] R. Beran, "Minimum hellinger distance estimates for parametri models," *The Annals of Statistics*, vol. 5, 05 1977.

[18] F. K. Dankar, M. K. Ibrahim, and L. Ismail, "A multi-dimensional evaluation of synthetic data generators," *IEEE Access*, vol. 10, pp. 11 147–11 158, 2022.

[19] B. Lorica and P. Nathan, *The state of machine learning adoption in the enterprise*. O'Reilly Media, 2018.

[20] G. A. Office and N. A. of Medicine, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development (Technology Assessment)*. U.S. GAO, 2019.

[21] J. Branson, N. Good, J.-W. Chen, W. Monge, C. Probst, and K. El Emam, "Evaluating the re-identification risk of a clinical study report anonymized under ema policy 0070 and health canada regulations," *Trials*, vol. 21, no. 1, pp. 1–9, 2020.

[22] K. El Emam, *Guide to the de-identification of personal health information*. CRC Press, 2013.