Evaluating Synthetic Data Generation Techniques for Medical Dataset

Takayuki Miura[1][a], Eizen Kimura[2][b], Atsunori Ichikawa[1][c],
Masanobu Kii[1][d] and Juko Yamamoto[1]
[1] *NTT Social Informatics Laboratories, Tokyo, Japan*
[2] *Dept. Medical Informatics, Medical School of Ehime Univ., Ehime, Japan*

Keywords: Synthetic Data Generation, Differential Privacy, Real-World Data.

Abstract: Anticipation surrounds the use of real-world data for data analysis in medicine and healthcare, yet handling sensitive data demands ethical review and safety management, presenting bottlenecks in the swift progression of research. Consequently, numerous techniques have emerged for generating synthetic data, which preserves the features of the original data. Nonetheless, the quality of such synthetic data, particularly in the context of real-world data, has yet to be sufficiently examined. In this paper, we conduct experiments with a Diagonosis Procedure Combination (DPC) dataset to evaluate the quality of synthetic data generated by statistics-based, graphical model-based, and deep neural network-based methods. Further, we implement differential privacy for theoretical privacy protection and assess the resultant degradation of data quality. The findings indicate that a statistics-based method called Gaussian Copula and a graphical-model-based method called AIM yield high-quality synthetic data regarding statistical similarity and machine learning model performance. The paper also summarizes issues pertinent to the practical application of synthetic data derived from the experimental results.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Real-world data collected from healthcare settings has attracted attention for propelling new clinical research due to its non-invasive nature for patients and its potential to constitute big data, thereby reducing bias. Including personal information in the data necessitates a substantial investment of person-hours for ethical review procedures and data protection, thereby impeding the prompt progression of medical research. Anonymization techniques, which reduce the risk of identifying individuals, are crucial in providing data to third parties without patient consent and streamlining the research approval process. Unlike secure computation (Cramer et al., 2015; Shan et al., 2018), which facilitates data analysis in encrypted form, these techniques afford analysts the advantages of viewing anonymized data that possess similar properties to the original in a format equivalent to actual data and conducting analyses in an exploratory manner. However, conventional anonymization methods, such as k-anonymity (Sweeney, 2002), encounter an issue where the quality of the anonymized data significantly diminishes as the data becomes high-dimensional (Aggarwal, 2005).

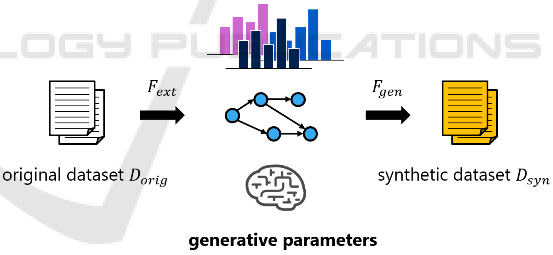

Figure 1: Overview of synthetic data generation.

The technology of synthetic data generation has been recognized for its ability to produce new data while preserving the original statistical properties of high-dimensional data (Hernandez et al., 2022; Tao et al., 2021; Sklar, 1959; Zhang et al., 2017; McKenna et al., 2022; McKenna et al., 2019; Xu et al., 2019). Specifically, this technology enables the expedited analysis of synthetic data in a relatively unrestricted environment, potentially abbreviating the approval process. Upon securing useful results, researchers can directly apply them to the original data, deriv-

[a] https://orcid.org/0000-0001-8694-312X
[b] https://orcid.org/0000-0002-0690-8568
[c] https://orcid.org/0000-0001-8013-7071
[d] https://orcid.org/0000-0003-1323-0983

Miura, T., Kimura, E., Ichikawa, A., Kii, M. and Yamamoto, J.
Evaluating Synthetic Data Generation Techniques for Medical Dataset.
DOI: 10.5220/0012314500003657
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024) - Volume 2*, pages 315-322
ISBN: 978-989-758-688-0; ISSN: 2184-4305
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

ing final results and potentially mitigating research costs (El Emam, 2020). Nevertheless, to the best of our knowledge, few studies have concurrently deployed various synthetic data generation techniques to authentic medical data (Barth-Jones, 2012; Culnane et al., 2017). Moreover, few studies have simultaneously applied various synthetic data generation techniques to real medical data, and insufficient knowledge has been accumulated on the differences among the techniques and the quality of the generated synthetic data.

In this paper, we generate synthetic data by using statistics-based, graphical-model-based, and deep-neural-network-based approaches and evaluate the quality of the resultant synthetic data. Utilizing the Diagnosis Procedure Combination (DPC) dataset from Ehime University Hospital as the original dataset, we evaluate generated synthetic data from three critical perspectives: distribution distances, machine learning model performances, and differences in correlation matrices. Furthermore, we incorporate differential privacy (DP) (Dwork, 2006) into each synthetic data generation method, serving as a theoretical privacy framework.

Consequent to the experimental results, we obtained the following conclusions:

- The incorporation of DP enhances privacy protection while concurrently diminishing the quality of synthetic data
- The magnitude of quality degradation is contingent upon the synthesis method employed. Gaussian Copula (Li et al., 2014) and AIM (McKenna et al., 2022) sustained comparatively superior quality even after applying DP.

2 RELATED WORK

2.1 Synthetic Data Generation

Numerous methods have been proposed for generating synthetic data, especially concerning tabular formatted data, while ensuring DP. Synthetic data generation approaches for tabular datasets can be categorized into three types. The first type is founded on basic statistics (Li et al., 2014; Asghar et al., 2020). The second type leverages graphical models (Zhang et al., 2017; Zhang et al., 2021; McKenna et al., 2022; McKenna et al., 2019). Tabular formatted data can be regarded as features extracted by humans. Since the graphical models learn relationships among attributes, they produce high-quality synthetic data (Tao et al., 2021). The third is the deep-neural-network-based method (Xu et al., 2019; Fang et al., 2022; Zhao et al., 2022; Chen et al., 2018; Lee et al., 2022; Kotelnikov et al., 2022; Liew et al., 2022). In this research, we evaluate one statistics-based method, three graphical-model-based methods, and one deep-neural-network-based method, utilizing a real medical dataset for the assessment.

2.2 Synthetic Data Generation for Medical Data

Researchers have directed substantial interest toward using synthetic data generation in the medical field, mainly focusing on image data (Guibas et al., 2017; Tajbakhsh et al., 2020). In these applications, practitioners employ synthetic data for data augmentation and privacy protection. However, the predominant methods, which are image-specific, present difficulties when applied to tabular data and do not account for DP. Although Hernandez et al. investigated a tabular healthcare dataset (Hernandez et al., 2022), their research concentrates exclusively on deep neural network-based synthetic data generation without considering DP. Our research evaluates several synthetic data generation techniques in conjunction with DP.

3 METHODOLOGY

Our experiment comprises three components: datasets, synthetic data generation algorithms, and evaluation methods. The experiment aims to evaluate the differences among synthesis algorithms and analyze DP's influence. An overview of the experiment is as follows:

- Apply a synthesis algorithm $F : \mathcal{D} \to \mathcal{D}$ to the original dataset D_{orig}. The generated synthetic dataset $F(D_{orig}) = D_{syn}$ is the same size as the original dataset D_{orig}.
- By using an evaluation method $E : \mathcal{D} \times \mathcal{D} \to R$, compare D_{syn} with D_{orig}.

3.1 Notations

In this paper, we focus on the tabular format datasets. A tabular dataset consists of several attributes $A_1, \ldots, A_d$. We can express a record as an element $x \in A := A_1 \times \cdots \times A_d$. If a dataset D contains N records, we can regard $D \in A^N$ and set a universe of datasets as $\mathcal{D} = A^N$. We set a probabilistic simplex $\Delta^d := \{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$.

Table 1: Names and types of attributes of DPC dataset. (n) means that the number of the attribute values is n.

	Name	Type
1	Gender	categorical (2)
2	Type of admission	categorical (7)
3	Emergency admission	categorical (2)
4	Length of Stay	numerical
5	Height	numerical
6	Weight	numerical
7	Smoking	categorical (2)
8	Pregnancy	categorical (2)
9	Independent eating	categorical (4)
10	Independence in Activities of Daily Living	categorical (4)
11	Independent Mobility	categorical (5)
12	Major diagnostic category	categorical (18)
13	Surgery	categorical (9)
14	Subclassification	categorical (10)
15	Secondary disease	categorical (3)

3.2 Dataset

This research uses a DPC dataset from Ehime University Hospital. This dataset has been extracted from the data warehouse, which encompasses DPC data from 2010 to 2013, to analyze the impact of 15 attributes on length of hospital stay: gender, type of admission, emergency admission, length of stay, height, weight, smoking, pregnancy, independent eating, independence in activities of daily living, independent mobility, major diagnostic category, surgery, subclassification, and secondary disease. Table 1 delineates the information for each category. All categorical data are encoded into one-hot vectors. Records containing missing values were excluded from the dataset, and the number of records became 9,666.

3.3 Synthesis Algorithm

In this research, we implement five synthesis algorithms, as listed in Table 2. Generally, a synthesis algorithm $F : \mathcal{D} \rightarrow \mathcal{D}$ is decomposed into two steps, as shown in Fig.1. The first step is to extract generative parameters $F_{ext} : \mathcal{D} \rightarrow \mathbb{R}^p$. Generative parameters are compressed information needed for the generation, such as basic statistics or trained machine learning model parameters. The second step is to generate synthetic data from the extracted generative parameters $F_{gen} : \mathbb{R}^p \rightarrow \mathcal{D}$.

Moreover, we use DP, which is known as the gold standard of the privacy protection framework (Dwork, 2006; Dwork et al., 2014). We add intentional noise to the generative parameter $\theta = F_{ext}(D)$ to satisfy DP. The formal definition is as follows.

Definition 3.1 (Differential privacy (Dwork, 2006; Dwork et al., 2014))**.** *A randomized function $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{Y}$ satisfies (ε,δ)-DP ((ε,δ)-DP) if for any neighboring $D, D' \in \mathcal{D}$ and $S \subset \mathcal{Y}$*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

In particular, $\mathcal{M}$ satisfies ε-DP if it satisfies $(\varepsilon, 0)$-DP.

If ε is smaller, it means that the output is more secure. We also interpret the case we do not add any intentional noise as $\varepsilon = \infty$. The stronger the protection, the worse the quality of outputs. δ can be regarded as a permissible error. This research investigates the case $\varepsilon = \infty, 8, 4, 2, 1$ and $\delta = 10^{-5}$.

3.3.1 Statistics-Based Methods

We evaluate the Gaussian Copula-based synthetic data generation as a statistics-based method (Sklar, 1959; Li et al., 2014). The Gaussian Copula's generative parameters are the original dataset's mean vector μ, the correlation matrix S, and the marginal distribution $H_1, \ldots, H_d$. For the DP version, we use the implementation by Li et al. (Li et al., 2014). We denote this method by `GCopula`.

3.3.2 Graphical-Model-Based Methods

We evaluate PrivBayes (Zhang et al., 2017), MWEM-PGM (McKenna et al., 2019), and AIM (McKenna et al., 2022) as graphical-model-based methods. PrivBayes trains important relations between attributes and expresses the relation as a directed acyclic graph. When generating data, attribute values are sampled in accordance with the graph. AIM and MWEM-PGM are similar methods that learn conditional probability tables to satisfy DP and sample data from them. These methods are denoted by `Bayes`, `MWEM`, and `AIM`.

3.3.3 Deep-Neural-Network-Based Methods

We evaluate Conditional Tabular Gan, CTGAN (Xu et al., 2019), as a deep-neural-network-based method. The differentially private version of CTGAN is implemented by smart-noise[1]. In this method, we train deep neural networks with DP-SGD (Abadi et al., 2016). This method is denoted by `CTGAN`.

3.4 Evaluation Methods (Quality of Synthetic Data)

In this research, we evaluate the quality of the synthetic dataset D_{syn}, which is the same size as the orig-

[1] https://docs.smartnoise.org/synth/index.html

Table 2: Synthesis algorithms in our experiment.

Synthesis algorithm	Description	Generative parameter
Gaussian Copula (Li et al., 2014)	`GCopula`	Statistics
PrivBayes (Zhang et al., 2017)	`Bayes`	Directed acyclic graph, conditional probability
MWEM-PGM (McKenna et al., 2019)	`MWEM`	Total joint distribution
AIM (McKenna et al., 2022)	`AIM`	Total joint distribution
CTGAN (Xu et al., 2019)	`CTGAN`	Model parameter of deep neural network

inal dataset D_{orig}, from three perspectives: distribution distances, machine learning model performances, and differences in correlations. Distribution distance is a broad measure, and machine learning model performance is a narrow measure (Drechsler and Reiter, 2009; Dankar et al., 2022). We also evaluate the absolute difference in correlations to compare relations explicitly. Let $E : \mathcal{D} \times \mathcal{D} \to R$ be an evaluation function.

3.4.1 Evaluation by Distribution Distances

The first evaluation is by statistical distribution distances $E_{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ between D_{orig} and D_{syn}. For each attribute, we evaluate the statistical distance of 1-way marginals. For the statistical distances, we use L1 distance, L2 distance, Hellinger distance, and Wasserstein distance. The definitions are as follows.

Definition 3.2 (L_p norm)**.** *For $x, y \in \Delta^d$, the L_p* **norm** *is defined as*

$$||x-y||_p := \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}.$$

We use the case when $p = 1$ or $p = 2$.

In a previous work, Hellinger distance was regarded as the best utility metric to rank synthetic data generation algorithms (El Emam et al., 2022).

Definition 3.3 (Hellinger distance)**.** *For $x, y \in \Delta^d$, the* **Hellinger distance** *is defined as*

$$\mathrm{Hel}(x,y) := \left(\sum_{i=1}^{d} \sqrt{x_i} - \sqrt{y_i}\right)^2.$$

Definition 3.4 (Wasserstein distance)**.** *For $x, y \in \Delta^d$, the* **Wasserstein distance** *or the* **Earth-Mover distance** *is defined as*

$$\mathrm{Was}(x,y) := \inf_{\gamma \sim \Gamma(x,y)} \mathbb{E}_{(a,b)\sim\gamma}[|a-b|],$$

where $\Gamma(x,y)$ is the set of all couplings of x and y. A coupling γ is a joint probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are x and y on the first and second factors, respectively.

3.4.2 Evaluation by the Difference of Machine Learning Model Performances

The second evaluation is the differences in machine learning model performances. Since DPC datasets are often used to predict the length of hospital stays, we train a regression model to predict length of stay (fourth attribute in Table 1) with LightGBM, which is a simple but high-performing machine learning model. We compare machine learning models trained by D_{syn} with D_{orig}.

The accuracy of models is evaluated by using the root-mean-square error (RMSE). For a trained model f, the error is defined by

$$RMSE(f,D) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2},$$

where $D = \{(x_i, y_i)\}_{i=1,\dots,n}$. We evaluate RMSE of a trained model with a synthetic dataset D_{syn}. Thus, the evaluation function $E_{ml} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is defined as $E_{ml}(D_{orig}, D_{syn}) = RMSE(f_{syn}, D_{orig})$, where f_{syn} is a trained model with D_{syn}.

3.4.3 Evaluation by the Difference of Correlation Matrices

The third evaluation is the difference in correlation matrices. The correlation matrix is defined as follows:

Definition 3.5 (Correlation matrix)**.** *For data samples $x^1, \dots, x^m \in \mathbb{R}^d$, set its mean vector as $\mu \in \mathbb{R}^d$. Then, a matrix $R \in \mathbb{R}^{d\times d}$ whose (i, j)-th component is*

$$R_{ij} = \frac{\sum_{k=1}^{d}(x_i^k - \mu_i)(x_j^k - \mu_j)}{\sqrt{\sum_{k=1}^{d}(x_i^k - \mu_i)^2}\sqrt{\sum_{k=1}^{d}(x_j^k - \mu_j)^2}}$$

is called the **correlation matrix***.*

We calculate the correlation matrices of D_{orig} and D_{syn}. We evaluate only numerical attributes and compute the absolute error of each component. Thus, the evaluation function $E_{cor} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n\times n}$ is defined as $(E_{cor}(D_{orig}, D_{syn}))_{i,j} = |R_{ij}^{orig} - R_{ij}^{syn}|$, where n is the number of numerical attributes.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

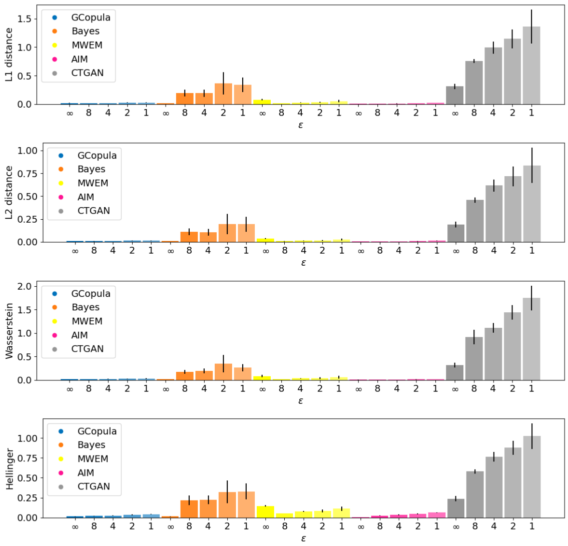

Figure 2: Result of categorical attributes distance. L1 distance, L2 distance, Hellinger distance, and Wasserstein distance from the top.

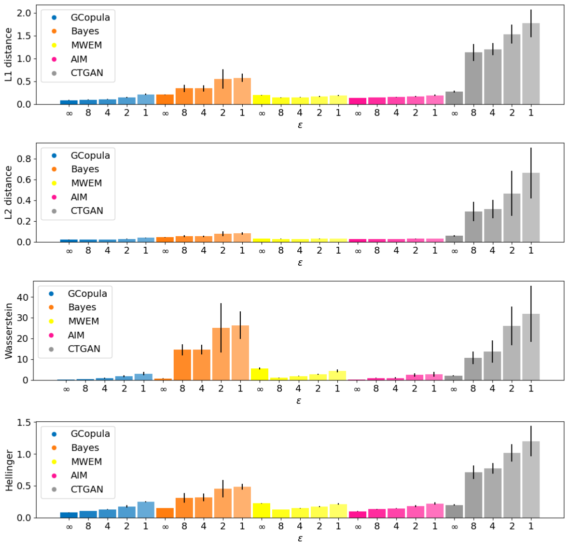

Figure 3: Result of numerical attributes distance. L1 distance, L2 distance, Hellinger distance, and Wasserstein distance from the top.

4 RESULTS

We generated synthetic data five times under the same conditions and calculated the average of the evaluation values. In this section, we report the results.

4.1 Distribution Distance Results

Figures 2 and 3 display the evaluation results by distribution distances, separating the graphs of categorical and numerical attributes due to differing scales. The results of all attributes are shown in Appendix. Values represent the means of all categorical or numerical attributes, respectively. Notably, the distance is regarded as a loss.

First, the losses for $\varepsilon = \infty$, representing a non-differentially private case, are small. Also, the losses significantly increase as the values of ε decrease, enhancing the robustness of the protection by DP.

`CTGAN` and differentially private `Bayes` exhibit more substantial losses when synthesizing algorithms are compared, while `GCopula`, `MWEM`, and `AIM` demonstrate lesser losses.

4.2 Machine Learning Model Performance Results

Fig. 4 illustrates the results of machine learning model performances, with the red line expressing RMSE for the original dataset. Non-differentially private results for each synthesis algorithm ($\varepsilon = \infty$) align closely with the original. The quality of the synthetic data discernibly declines as ε increases. Specifically, the results from differentially private `Bayes` and `CTGAN` are inferior, while those of `GCopula` and `AIM` remain proximate to the original results, even when differentially private.

4.3 Difference in Correlations Results

Fig. 5 presents the results in cases where $\varepsilon = \infty$, the absolute losses of `GCopula` and `Bayes` are small. Additionally, losses become more significant as ε increases, resulting in differentially private `CTGAN` being the worst.

5 DISCUSSION

5.1 Quality of Synthetic Data

The three evaluation methods reveal that the losses associated with non-differentially private synthesis remain sufficiently small, while DP diminishes the quality of synthetic data. In differentially private cases, the magnitude of the losses varies among synthesis methods. This indicates the potential for enhancing the quality of synthetic data by strategically devising DP. Notably, the recently proposed `AIM` achieves noteworthy experimental results consistently. `AIM` manifests negligible deterioration in the quality of the synthetic data when implementing DP.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

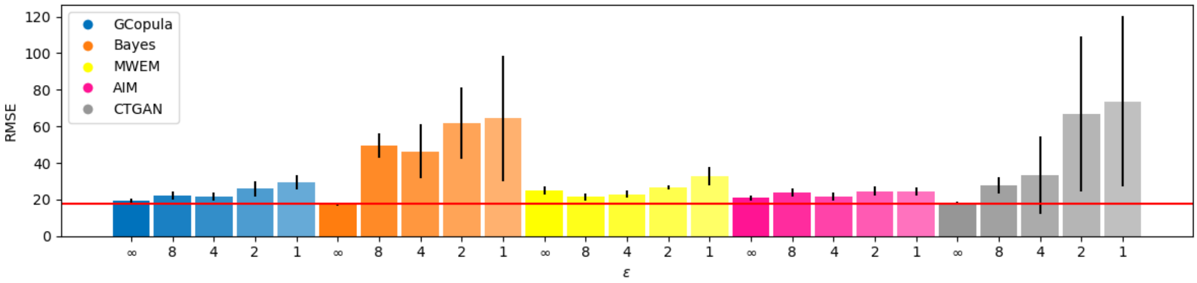

Figure 4: Results of machine learning model performances: RMSEs of a trained LightGBM regression model.

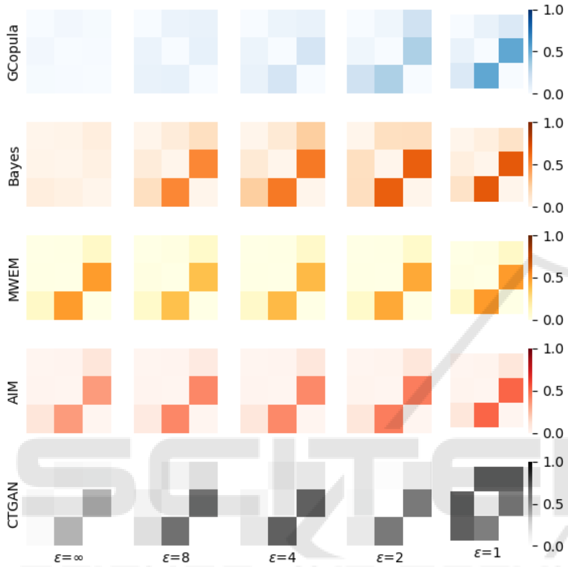

Figure 5: Results of differences in correlations

5.2 Evaluation Methods

This study employs L1 distance, L2 distance, Hellinger distance, and Wasserstein distance as evaluative metrics, which are widely utilized in studies measuring the quality of synthetic data and prove highly useful when assessing the ”relative” quality thereof. These metrics indicate that `AIM` exhibits notably superior results to other methods.

Conversely, to facilitate absolute evaluations with qualitative significance, it is necessary to assume realistic use cases for evaluations by machine learning performance and ascribe meaning to the magnitude of errors.

5.3 Towards Practical Use

Discussion has yet to emerge regarding whether using synthetic data for personal data is subject to the agenda of Ethics Review Committees. Conversely, Guo et al. have reported that they did not require an ethical review because the synthetic data contained no information that could lead to the identification of individual patients (Guo et al., 2020). It has been posited that, should synthetic data gain recognition as a viable option for privacy considerations, obtaining approval from ethics committees may become unnecessary (Azizi et al., 2021). In a case wherein an organization inadvertently disclosed the personal information of numerous individuals online while testing a cloud solution, the Norwegian Data Protection Authority (Datatilsynet) highlighted that testing could have been conducted by processing synthetic data or using less personal data [2]. This ruling also implies that synthetic data may be recognized as having the potential to exclude information that leads to personal identification.

Furthermore, DP can potentially enhance the security of such synthetic data. Therefore, DP is anticipated to minimize discussions concerning anonymous processing and expedite the progression of research. Nonetheless, studies have examined attacks that deduce the original data from synthetic data (Stadler et al., 2022), necessitating further research to ensure its security.

6 CONCLUSION

In this research, employing the a Diagnosis Procedure Combination (DPC) dataset, we experimentally evaluated synthetic data generation techniques’ effectiveness using statistic-based, machine-learning model-based, and deep neural network-based methods. The investigation clarified the differences in performance among the methods, attributing them to variations in the amount of source data and the degree of accuracy degradation when implementing differential privacy. Further, we discussed issues that must be addressed to apply synthetic data generation techniques more effectively.

[2] https://www.dataguidance.com/news/norway-datatilsynet-fines-nif-nok-12m-disclosing

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ETHICAL CONSIDERATIONS

The Ethics Review Committee of Ehime University Hospital approved this study (“Quality evaluation of synthetic data generation methods preserving statistical characteristics,” Permission number 2012001), and we conducted it in accordance with the committee’s guidelines.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *VLDB*, volume 5, pages 901–909.

Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S., and Kaafar, D. (2020). Differentially private release of datasets using gaussian copula. *Journal of Privacy and Confidentiality*, 10(2).

Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., and El Emam, K. (2021). Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497.

Barth-Jones, D. (2012). The’re-identification’of governor william weld’s medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *Then and Now (July 2012)*.

Chen, Q., Xiang, C., Xue, M., Li, B., Borisov, N., Kaarfar, D., and Zhu, H. (2018). Differentially private data generative models. *arXiv preprint arXiv:1812.02274*.

Cramer, R., Damgård, I. B., and Nielsen, J. B. (2015). *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press.

Culnane, C., Rubinstein, B. I., and Teague, V. (2017). Health data in an open world. *arXiv preprint arXiv:1712.05627*.

Dankar, F. K., Ibrahim, M. K., and Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158.

Drechsler, J. and Reiter, J. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of Official Statistics*, 25(4):589–603.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. *IEEE Security & Privacy*, 18(4):56–59.

El Emam, K., Mosquera, L., Fang, X., and El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*, 10(4):e35734.

Fang, M. L., Dhami, D. S., and Kersting, K. (2022). Dp-ctgan: Differentially private medical data generation using ctgans. In *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*, pages 178–188. Springer.

Guibas, J. T., Virdi, T. S., and Li, P. S. (2017). Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*.

Guo, A., Foraker, R. E., MacGregor, R. M., Masood, F. M., Cupps, B. P., and Pasque, M. K. (2020). The use of synthetic electronic health record data and deep learning to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. *Frontiers in digital health*, 2:576945.

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.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2022). Tabddpm: Modelling tabular data with diffusion models. *arXiv preprint arXiv:2209.15421*.

Lee, J., Kim, M., Jeong, Y., and Ro, Y. (2022). Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353.

Li, H., Xiong, L., Zhang, L., and Jiang, X. (2014). Dp-synthesizer: Differentially private data synthesizer for privacy preserving data sharing. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 7, page 1677. NIH Public Access.

Liew, S. P., Takahashi, T., and Ueno, M. (2022). PEARL: Data synthesis via private embeddings and adversarial reconstruction learning. In *International Conference on Learning Representations*.

McKenna, R., Mullins, B., Sheldon, D., and Miklau, G. (2022). Aim: An adaptive and iterative mechanism for differentially private synthetic data. *arXiv preprint arXiv:2201.12677*.

McKenna, R., Sheldon, D., and Miklau, G. (2019). Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444. PMLR.

Shan, Z., Ren, K., Blanton, M., and Wang, C. (2018). Practical secure computation outsourcing: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–40.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Stadler, T., Oprisanu, B., and Troncoso, C. (2022). Synthetic data–anonymisation groundhog day. In *31st*

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

USENIX Security Symposium (USENIX Security 22), pages 1451–1468.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., and Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693.

Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., and Miklau, G. (2021). Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4).

Zhang, Z., Wang, T., Li, N., Honorio, J., Backes, M., He, S., Chen, J., and Zhang, Y. (2021). {PrivSyn}: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. Y. (2022). Ctab-gan+: Enhancing tabular data synthesis. *arXiv preprint arXiv:2204.00401*.

APPENDIX

Results of All Attributes

The results of distribution distances for each attribute are shown in Fig. 6, 7, 8 and 9.

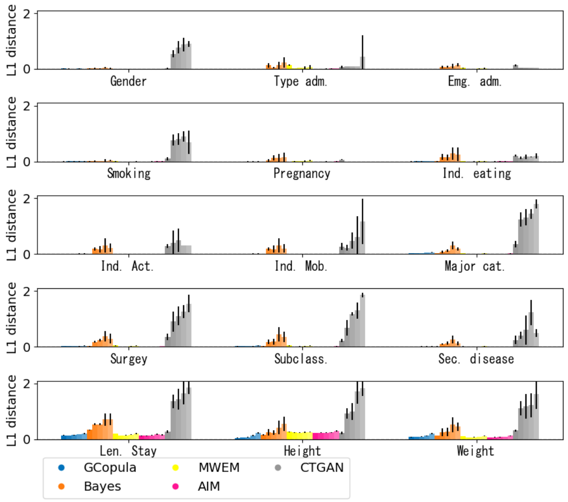

Figure 6: The values of L1 distance of each attribute.

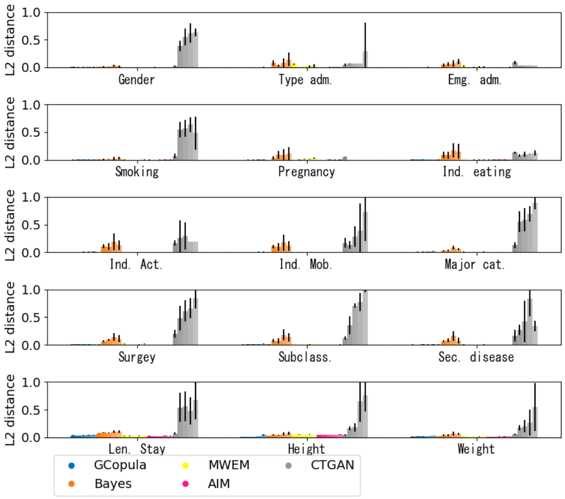

Figure 7: The values of L2 distance of each attribute.

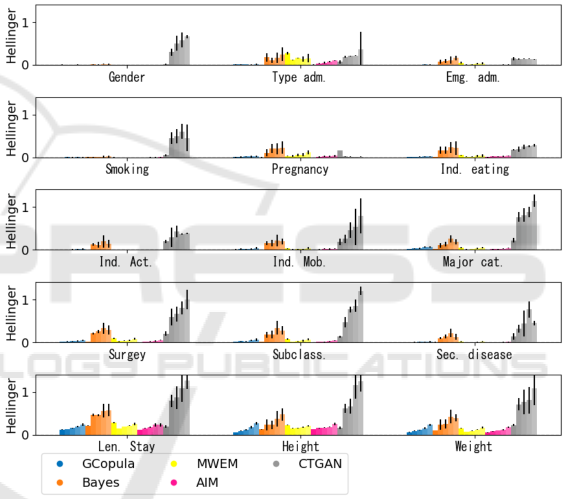

Figure 8: The values of Hel. distance of each attribute.

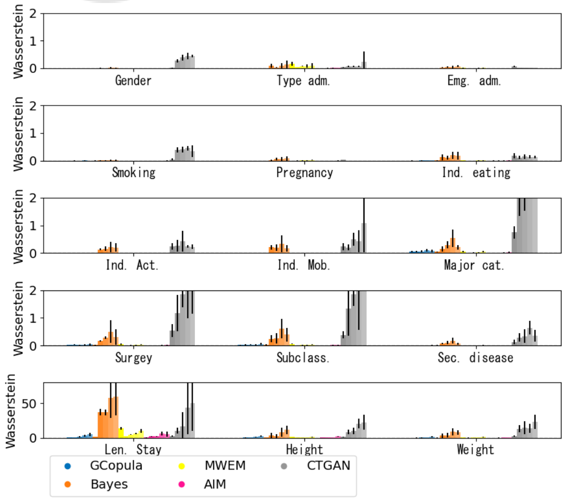

Figure 9: The values of Was. distance of each attribute.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS