J. Mantas et al. (Eds.)

doi:10.3233/SHTI240511

Synthetic Generation of Patient Service Utilization Data: A Scalability Study

Joseph HOWIE [a,1], Sowmya BALASUBRAMANIAN [a], Jonas BAMBI [a],
Kenneth MOSELLE [a], Venkatesh SRINIVASAN [b] and Alex THOMO [a]
[a] *University of Victoria, BC, Canada*
[b] *Santa Clara University, CA, USA*
ORCiD ID: Joseph Howie https://orcid.org/0009-0009-5178-2499,
Sowmya Balasubramanian https://orcid.org/0009-0000-8959-0956,
Jonas Bambi https://orcid.org/0009-0006-4459-1158,
Venkatesh Srinivasan https://orcid.org/0000-0003-3819-3998,
Alex Thomo https://orcid.org/0000-0002-3020-2258

Abstract. To address privacy and ethical issues in using health data for machine learning, we evaluate the scalability of advanced synthetic data generation methods like GANs, VAEs, copulaGAN, and transformer models specifically for patient service utilization data. Our study examines five models on data from a Canadian health authority, focusing on training and generation efficiency, data resemblance, and practical utility. Our findings indicate that statistical models excel in efficiency, while most models produce synthetic data that closely mirrors real data, and is also useful for real-world applications.

Keywords. Health service data, synthetic data generation, Generative ML, Statistical models, Validation metrics

1. Introduction

In the data-driven landscape, the quest for high-quality data is paramount, with accuracy, timeliness, and relevance being key for insightful decision-making. Large datasets un- cover complex patterns, enhancing the robustness of machine learning models. However, challenges such as privacy concerns, high costs, and ethical issues impede the acquisition of vast, quality data [1]. Synthetic data generation (SDG) emerges as a solution, simulating real data's statistical properties without compromising sensitive information, addressing privacy concerns, and broadening data accessibility [2].

Early SDG methods such as duplicating existing data, sampling, and interpolation methods were simplistic, failing to capture real-world data complexities. In contrast, advanced approaches like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and transformer models have made significant strides [3, 4, 5, 6, 7]. These methods, leveraging sophisticated statistical methodologies and machine learning algorithms, produce highly realistic synthetic data while also addressing privacy

[1] Corresponding Author: Joseph Howie, Department of Computer Science, University of Victoria, Canada; E-mail: joehowie@uvic.ca.

concerns [8, 9, 10]. Yet, a systematic comparison of these methods' scalability, particularly for patient service utilization (PSU) data, remains unexplored.

This study fills this gap by evaluating five state-of-the-art SDG models on real patient data from a Canadian health service, focusing on four patient cohorts. We assess models based on training and generation times, resemblance to real data, and utility in practical scenarios. Our findings highlight the efficiency of statistical models such as FastML and Gaussian Copula, and the high resemblance and utility of synthetic data across most models, showcasing SDG's potential in addressing healthcare data acquisition challenges. The novel aspect of our study is its focus on the scalability of the SDG methods using real-world PSU data.

2. Methods

In our study, we delve into the intricacies of PSU data obtained from a comprehensive regional health service system in Canada. This data is unique as it encapsulates the sequential interactions of patients with various healthcare service classes, providing a granular view of healthcare usage patterns. We concentrated on four distinct patient cohorts, namely Schizophrenia Services (SS), Homeless–Ever (HE), Addictions Service – Post Withdrawal (AS), and Opioid Overdose (OO). Each patient's PSU is represented as a chronologically ordered series of service class IDs, reflecting their journey through the healthcare system. This sequential data is crucial for understanding patient pathways and service dependencies, making it a rich resource for healthcare analytics and planning.

Table 1. Cohorts

Cohort Name	Cohort Code	# Real Patients
Schizophrenia Services	SS	1829
Homeless – Ever	HE	2221
Addictions Service – Post Withdrawal	AS	2592
Opioid Overdose	OO	5381

To synthesize this complex and sequential PSU data, we evaluated a spectrum of five Synthetic Data Generation (SDG) models: Gaussian Copulas, Fast ML, CT- GAN, CopulaGAN, and a customized Generative Pre-trained Transformer (GPT) model (https://sdv.dev/). These models were selected due to their diverse methodological underpinnings, combining advanced statistical methodologies with cutting-edge machine learning algorithms. This blend allows for a comprehensive analysis of their efficacy in replicating the nuanced patterns found in PSU data.

The fidelity of the synthetic data to the original PSU sequences was evaluated using the robust validation metric called Jensen-Shannon Divergence (JSD) which is a symmetric version of the Kullback-Leibler Divergence (K-L Divergence):

$$D_{KL}(P||Q) = \sum_{x \in V} P(x) \log \frac{P(x)}{Q(x)}$$

$$JSD(P||Q) = \frac{1}{2}\left[D_{KL}(P||M) + D_{KL}(Q||M)\right]$$

where P and Q are probability distributions and M = (P + Q)/2. This metric is pivotal in quantifying the statistical resemblance between the synthetic and real datasets, ensuring that the synthetic data retains the statistical properties of the original data while preserving individual patient privacy. Furthermore, the utility of the synthetic data in real-world applications was assessed using a Binary Recurrent Neural Network (RNN) classifier. This classifier was tasked with distinguishing between cohorts based on their PSU sequences, trained on one dataset (real or synthetic) and tested on the other. This crossvalidation approach provides a comprehensive evaluation of the synthetic data's practical utility, and its potential to act as a surrogate for real data in healthcare applications.

3. Results and Discussion

We analyzed the scalability of five synthetic data generation methods on PSU data performing experiments on a secure server setup and describe our results below.

RQ1: How efficient are the models in training and data generation?

Training time analysis revealed significant variances among models (Table 2). The Gaussian Copula model was the fastest, whereas Fast ML, contrary to expectations, was slower. Both CTGAN and CopulaGAN models exhibited prolonged training times for larger cohorts, unlike the Transformer model, whose training time remained consistent across different cohort sizes.

Table 2. Time to fit each Cohort set to each model

Generative Method	SS	HE	AS	OO
Fast ML	1.75s	2.29s	2.91s	8.37s
Gaussian Copula	0.55s	0.53s	0.48s	1.20s
CTGAN	2973.24s	6748.23s	14176.77s	170854.28s
CopulaGAN	2973.01s	6740.78s	14039.20s	204707.17s
Transformer	469.23s	453.68s	454.34s	459.88s

In terms of data generation, Gaussian Copulas and Fast ML proved to be the most efficient, displaying super-linear time increases as dataset sizes grew. CTGAN and CopulaGAN also showed super-linear increases but were less efficient. The Transformer model lagged behind, with its data generation time increasing linearly with the size of the dataset. Figures 1, 2, and 3 visually illustrate the data generation times, underscoring the relative efficiency and scalability of the different models.

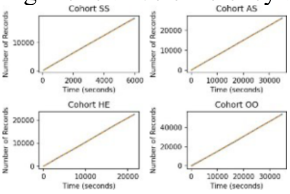

Figure 1. Time to Generate Data with the Transformer model.

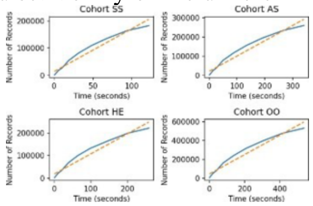

Figure 2. Time to Generate Data with the Gaussian Copulas model.

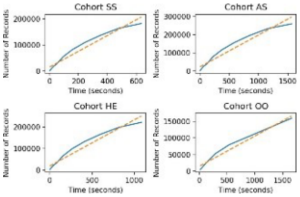

Figure 3. Time to Generate Datawith the CTGAN model.

RQ2: How do the generation models perform with respect to resemblance metrics?

Our evaluation of synthetic data's resemblance to real data revealed that marginal distributions across most models and features were closely aligned, with Jensen-Shannon Distances (JSD) mostly below 1%. Larger discrepancies were noted in joint distributions, particularly at smaller dataset sizes, but these distances diminished as dataset sizes increased, indicating improved resemblance. Detailed findings for cohort SS, as seen in Table 3, illustrate this trend, with age (Age), number of interactions (NI), and the ratio of usage to total interactions (U/T) showing notable convergence towards real data metrics at larger scales. Similar patterns were observed in other cohorts, underscoring the efficacy of synthetic data in mimicking real-world distributions (Refer to the full version [11] for comprehensive analysis across all cohorts and models).

Table 3. Transposed Consolidated JSD in percentages for Cohort SS across models as the dataset size scales

Scale	Gaussian Copulas JSD (%)			CTGAN Model JSD (%)			Transformer Model JSD (%)		
	Age	NI	U/T	Age	NI	U/T	Age	NI	U/T
1x	0.683	0.476	0.157	2.388	0.066	0.179	0.588	1.468	1.054
5x	0.513	0.002	0.04	1.323	0.003	0.043	0.601	0.441	0.771
10x	0.481	0.002	0.012	2.079	0.003	0.039	0.619	0.44	0.775

RQ3: How useful is the synthetic PSU data in practical scenarios?

We employed Binary Recurrent Neural Network (RNN) models to evaluate the practical utility of synthetic data. These models were tasked with differentiating between patients from the Schizophrenia Services (SS) and Addiction Services–Post Withdrawal (AS) cohorts based on their Patient Service Utilization (PSU) data. This approach allowed us to assess the synthetic data's effectiveness in scenarios that closely mimic real- world applications. The performance of these models was quantified using standard metrics: Accuracy, F-measure, Precision, and Recall.

Table 4. Performance metrics for SS vs. AS cohort classification using 10x datasets. The table presents resultsfrom three scenarios: Training (Training the RNN with real data), TSTR (Train on Synthetic, Test on Real), and TRTS (Train on Real, Test on Synthetic).

Scenario	Model	Accuracy	F-measure	Precision	Recall
Training	Real	0.959	0.958	0.951	0.966
	Transformer	0.961	0.961	0.971	0.951
	Gaussian Copulas	0.986	0.986	0.974	0.999
	CTGAN	0.988	0.988	0.978	1.0
TSTR	Transformer	0.956	0.956	0.95	0.962
	Gaussian Copulas	0.988	0.988	0.977	0.999
	CTGAN	0.988	0.988	0.978	0.999
TRTS	Transformer	0.938	0.937	0.96	0.915
	Gaussian Copulas	0.975	0.975	0.966	0.985
	CTGAN	0.978	0.978	0.97	0.986

The results shown in Table 4 showcase the synthetic data's practical utility, especially data generated by Gaussian Copulas and CTGAN models, which consistently matched real data performance in various scenarios. This highlights the potential of synthetic data to faithfully replicate the complex patterns of real healthcare datasets, making it a valuable asset for machine learning applications while addressing privacy concerns. Gaussian Copulas, in particular, produced synthetic data that proved highly effective in both the training and testing phases. This underscores the technique's capability to generate reliable synthetic data that can support a wide range of analytical tasks in healthcare, from predictive modeling to patient cohort analysis. The overall findings confirm synthetic data's promise as a practical tool in healthcare analytics, enabling researchers and clinicians to leverage data-driven insights while safeguarding patient privacy.

4. Conclusions

In our study, we assessed various synthetic data generation methods for patient service utilization data, focusing on statistical, machine learning-based, and hybrid approaches. Our findings reveal that Gaussian Copulas and Fast ML models excel in training and data generation speed. Additionally, through Jensen-Shannon Divergence, we confirmed that synthetic datasets closely resemble real data, with minor divergences. Despite larger discrepancies in joint distributions, Gaussian Copulas and Fast ML consistently outperformed others. Utilizing a Binary RNN for validation showed that synthetic data maintains high utility, with performance metrics surpassing 90% accuracy. Therefore, for generating PSU data, Gaussian Copulas and Fast ML models emerge as the top choices.

References

[1] Torfi A, Fox EA, and Reddy CK. Differentially private synthetic medical data generation using convolutional GANs. Information Sciences 2022; 586:485–500.

[2] El Emam K, Mosquera L, Fang X, and El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. JMIR medical informatics 2022; 10:e35734.

[3] Piacentino E, Guarner A, and Angulo C. Generating synthetic ECGs using GANs for anonymizing healthcare data. Electronics 2021; 10:389.

[4] Dahmen J and Cook D. SynSys: A synthetic data generation system for healthcare applications. Sensors 2019; 19:1181.

[5] Torfi A and Fox EA. CorGAN: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. arXiv preprint arXiv:2001.09346 2020.

[6] Goncalves A, Ray P, Soper B, Stevens J, Coyle L, and Sales AP. Generation and evaluation of synthetic patient data. BMC medical research methodology 2020; 20:1–40.

[7] Larrea X, Hernandez M, Epelde G, Beristain A, Molina C, Alberdi A, Rankin D, Bamidis P, and Konstantinidis E. Synthetic Subject Generation with Coupled Coherent Time Series Data. Engineering Proceedings 2022; 18:7.

[8] Wang Z, Myles P, and Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Computational Intelligence 2021; 37:819–51.

[9] McLachlan S, Dube K, Gallagher T, Simmonds JA, and Fenton N. Realistic synthetic data generation: The ATEN framework. Biomedical Engineering Systems and Technologies: 11th International Joint Conference, BIOSTEC 2018. 2019 :497– 523.

[10] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, and McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association 2018; 25:230–8.

[11] Howie J, Balasubramanian S, Bambi J, Moselle K, Srinivasan V, and Thomo A. Synthetic Generation of Patient Service Utilization Data: A Scalability Study. Technical Report, University of Victoria 2024.