Synthetic Medical Data Generation using Transformer-based Generative AI: A Performance Comparison with Faker and CTGAN

Srinivas Suresh Sikhakolli
Associate Professor,
Kirloskar Institute of Management, Pune

Asha Kiran Sikhakolli
Associate Professor,
Dr.D.Y.Patil B-School

ABSTRACT
Access to medical data is essential for health care research and advanced analytics. However, strict privacy regulations significantly limit data availability, hinder the machine learning applications. Due to these limitations, synthetic data usage raising across the world. Prior studies focused on building synthetic data using rule-based models such as Faker and deep learning models such as CTGAN. In recent years, ChatGPT, a transformer based Generative AI model has emerged with advanced capabilities to generate wide variety of synthetic data on demand. The aim of this research is to show that the transformer based generative AI model produces quality synthetic data that yields better predictive performance when compared with the Faker and CTGAN models. The synthetic data has been generated with reference to the UCI Cleveland Heart data. Random Forest algorithm has been used to evaluate the performance of the model. The results of the experiment prove that the transformer based GenAI, ChatGPT generated synthetic data yields better performance when compared with the Faker and CTGAN models. Also, proves that the performance metrics of ChatGPT based synthetic data are close to the actual Cleveland heart medical data. Our findings suggest that ChatGPT model effectively captured clinical relationships and offers practical insights for researchers without losing the privacy in synthetic data. This type experiment is useful for non-clinical research.

Keywords
Synthetic Medical Data, Generative AI, Privacy-Preserving Data, Random Forest Model Performance.

1. INTRODUCTION
Getting original data is highly needed for health care research to produce acceptable results. Health care data is highly confidential and often not available easily. Because of these limitations, synthetic data usage is growing. It is an artificial data which mimics the original data. Health care is one of the Industry often uses synthetic data for research and analysis. Heal care contains sensitive information. Hence, hospitals do not share the data easily. Because, its access and usage linked to strict privacy regulation laws such as HIPAA(USA), GDPR (European Union), PDP (India) pose a barrier for sharing the with researchers. These laws regulate the data usage and provides data protection rules. Many hospitals share the synthetic data in place of original data for research and analysis for non-clinical operations. Also, while sharing, appropriate balancing required between privacy and utility [1]. Global synthetic data in health care reached USD $457 and will reach USD 5.68 billion by 2033 [2]. There are wide variety of techniques exists for generating the synthetic data. Well know techniques are ; GANs (Generative Adversarial Networks, VAEs (Variational Autoencoders), Diffusion model, Autoregressive models and Bayesian networks. Synthetic can be generated in multiple ways. Pythons provides packages for generating synthetic data. One of the package is Faker. It has many functions or methods for generating the synthetic data. One of the method is Faker. It is a rule-based method. Rule-based means has to define type of data and characteristics of data manually in Python programme. It generates superficially realistic data and lack the complex clinical correlations. The problem with this method is that researcher must be aware of characteristics of data in advance. Improper characteristic definitions lead to inaccurate synthetic data. Another popular category is GAN based methods. These methods embed differential privacy into a conditional tabular GAN, demonstrating that synthetic medical data can indeed be generated with formal privacy guarantees without entirely sacrificing utility [3]. It is observed that organizations generate synthetic data by defining the characteristics or features of the original data. In this process, high possibility of mismatch between real-world data and synthetic data which may affect the accuracy, reliability and practical applicability. In this research, we considered UCI machine learning Cleveland heart data as reference data [4]. Based on this reference data, synthetic data is generated using the Pyhton Faker tool, Pyhton GAN tool and GenAI-chatGPT tool. One major advantage of this approach is that it enables a comparative evaluation of different synthetic data generation techniques using the same reference dataset. By using the UCI Machine Learning Cleveland heart disease dataset as the baseline, the study can assess predictive generated through Python Faker, Python GAN, and GenAI-ChatGPT tools. This approach also helps in identifying the strengths and limitations of each method in preserving statistical characteristics of the original healthcare data while reducing privacy risks associated with real patient information. The below given diagram(Fig1) shows the block diagram of the process of the synthetic data generation. The Faker and CT-GAN are python based methods. Whereas GenAI follows transformer based method. GenAI based method requires input prompt in the form of text. The following prompt used for generating the synthetic data based on the UCI Cleveland data.

“I want to generate synthetic data similar to the above data but not exactly same. The synthetic data must preserve distributions, type of data and characteristics of the original data. Generate 303 records with .csv format. Further I want to test the Machine Learning predictive performance with other Synthetic data generation methods like Faker method and CTGAN. Pls do the needful” This synthetic data preserves original numerical ranges and categorical distributions which are aligned with the original data. The figure below shows the overview of the research design.

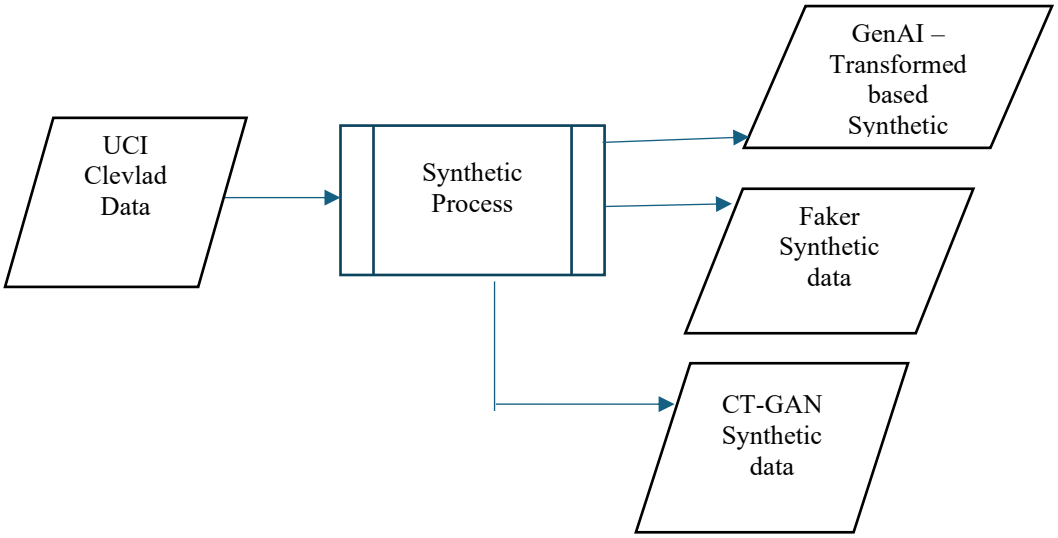

Figure1. Synthetic process block diagram

The rest of the paper organized as sections: section 2 outlines literature review, section 3 research objectives and hypotheses, section 4 research methodology, section 5 results and evaluation, section 6 conclusion followed by section 7 references.

2. 2 LITERATURE REVIEW

Literature Review: Synthetic Medical Data & Privacy-Preserving Generative Models.

Zoho et al [5] developed an improved generative adversarial network to create high quality tabular data that offers better utility and privacy. The paper explains loss functions, including downstream loss and Wassertein loss with gradient penalty, along with an innovative method for encoding mixed continuous and categorical variables. This method better represents complex distributions. Additionally, the model uses differentially private stochastic gradient descent to ensure privacy protection. Extensive testing on multiple benchmark data sets showed that CTAB-GAN + outperformed other tabular data generators in terms of both statistical similarity and machine learning under various privacy settings. The results revealed notable improvements in F1 –score and AUC baseline methods

Umesh C et al [6] explored generative AI models (GANs, VAEs) to create clinical tabular physiological data such as patient lab values and vital signs. It highlights challenges including missing data, multi-table data structure and preserving clinical correlations. Privacy considerations are also discussed with low – identification risk demonstrated in some applications. This work addresses synthetic clinical data with utility and privacy in a physiological context

Ahmed H.A et al [7] investigated six variants of GAN's – including GAN, CGGAN, CTGAN, DRAGAN, WGAN and Cramer GAN-applied to multiple medical tabular data sets (Breast cancer, Lung Cancer, and fetal cardiography). The author generated synthetic data and then assess the quality interms of statistical similarity, classification performance using XGBOOST and SVM. Their results show that more advanced GAN variants like CTGAN and CGAN provide better fidelity and downstream predictive performance than simple HAN models.

Ghosheh et al [8] reviewed 32 studies on synthetic data generation for the tabular health data focusing on how privacy implemented and evaluated. The paper categorized the privacy strategies into three types; noise addition, constraints and model reliance. Their major finding is that privacy evaluation is inconsistent across studies-many papers fail to standardize metrics for measuring re-identification risk or membership interference. Their paper proposed unified framework to make comparisons more meaningful.

Karmrkar et al [9] discussed how generative AI is used to synthesize medical images for diagnostics tasks. The authors analyze methods, privacy risks and clinical applications, highlighting how synthetic medical images can address data scarcity in radiology and imaging research. The author emphasizes the need for rigorous privacy testing (such as membership interference and re-identification risk) in generative AI models used for medical image synthesis. While focusing on imaging. It demonstrates the increasing relevance of GenAI in privacy –aware medical data generation

Overall the earlier researchers proposed traditional methods like faker and GAN based AI models for generating synthetic tabular data. However recent GenAI models like chatGPT uses transformer based Large Language Models for generating the content including the tabular data. Hence, they are capable of generating wide variety of data at any time with great innovation. They are trained on billions of parameters. For example, GPT-3 was trained on 175 billion parameters. Zhang et al [10] proposed a synthetic data generation algorithm for physical examination data, integrated with differential privacy. The authors designed a Bayesian network-based method with Gibbs sampling to produce synthetic records, then apply noise to ensure a given privacy. Their work demonstrated that the synthetic data maintain important statistical properties of the real data while significance reducing re-identifying risk. It provides a concrete example of privacy-preserving synthetic data generation in a medial tabular context

3. RESEARCH OBJECTIVES & HYPOTHESIS

1. Generate Synthetic data using naive rule-based Faker technique, differentially private generators (CTGAN), and ChatGPT
2. Apply the machine learning algorithm on synthetic data to compare the mean predictive performance between traditional method, Faker with ChatGPT
3. Apply the machine learning algorithm on Cleveland data and ChatGPT Synthetic data to compare the predictive performance

4. Demonstrate that ChatGPT generated synthetic data performs better predictive performance when compared to Faker and CTGAN.

Hypothesis:

Table 1: hypothesis design

Hypothesis	**Null Hypothesis(H_0)**	**Alternate Hypothesis (H_1)**
Hypothesis-1	The mean predictive performance using chatGPT generated heart synthetic data is less than or equal to that using Faker generated heart synthetic data. $H_O: \mu_{ChatGPT\_Synthetic} \leq \mu_{Faker}$	The mean predictive performance using chatGPT generated heart synthetic data is higher than or equal to using Faker generated heart synthetic data. $H_1: \mu_{ChatGPT\_Synthetic} > \mu_{Faker}$
Hypothesis-2	The mean predictive performance using ChatGPT generated heart synthetic data is less than or equal to that using CTGAN generated heart synthetic data. $H_O: \mu_{ChatGPT\_Synthetic} \leq \mu_{CTGAN}$	The mean predictive performance using ChatGPT generated is higher than or equal to using CTGAN generated heart synthetic data. $H_1: \mu_{ChatGPT\_Synthetic} > \mu_{CTGAN}$

4. RESEARCH METHODOLOGY

4.1 Data

This study utilizes the Cleveland heart coronary clinical disease data as secondary source of data (Janosi et al., 1989). This data is preprocessed data. It means the data hides the anonymity of patients, and represents the data in numeric format suitable for machine learning operations. It contains clinical and demographic attributes relvant to heart disease prediction, providing a well-established benchmark for evaluating machine learning models. In this study, this secondary data is considered as original data.

4.2 UCI Cleveland data

The Cleveland data contains total of 303 observations including the variables; Age_patient, Sex_patients, **c**hest pain type_patient, Trestbps_patient, Fbs_patient, Restecg_patient, Ca_patient, Thal_patient, Num_patients(0-no disease,0-4 disease present) Oldpeak_patient, and Slope_patient

4.3 Machine Learning Algorithm

The Random Forest classifier (Breiman, L. 2001) was used for evaluating predictive performance across all the three synthetic data as it well suitable to overfitting.

Table 2: Model Configuration

Tool	**Training data split**	**Metrics**	**Number of iterations**	**Number of tree models for training and prediction**	**Sampling method**	**Random_state**
Anaconda Python-Jupytor Note book.	80:20, 70:30	Accuracy, F1 score, Precision, Recall and AUC	30	25 max	Stratified	None

4.4 Tools

The experiment was carried out on Anaconda Python with Scikit learn library. The modules include: Faker synthetic data, DP-GAN synthetic data, GenAI synthetic data and Random Forest machine learning.

5. RESULTS AND EVALUATION

Random Forest classifier was trained on the original data and the synthetic data. Performance metrics accuracy, F1-score, Precision, Recall, and AUC were computed with 30 iterations with different random seeds to enable statistical testing. The following (table 3) shows the Model performance measures:

Table3: Model Performance Measures

Model	**Accuracy (Mean ± SD)**	**Precision (Mean ± SD)**	**Recall (Mean ± SD)**	**F1-Score (Mean ± SD)**	**AUC (Mean ± SD)**
Cleveland Orignal	0.840 ± 0.012	0.820 ± 0.015	0.810 ± 0.014	0.815 ± 0.013	0.880 ± 0.010
Faker_Synthetic	0.650 ± 0.020	0.600 ± 0.025	0.580 ± 0.023	0.590 ± 0.021	0.620 ± 0.018
CTGAN_Synthetic	0.810 ± 0.014	0.790 ± 0.016	0.780 ± 0.015	0.785 ± 0.014	0.850 ± 0.012
ChatGPT_Synthetic (GenAI-transformer)	0.820 ± 0.013	0.800 ± 0.015	0.790 ± 0.014	0.795 ± 0.013	0.860 ± 0.011

The table 3 shows GPT outperforms both Faker and CTGAN across all the metrics. The accuracy, precision, recall and F1 score and AUC (area under curve) are the frequently used measures. The same measures considered in this research. The table shows that measures of chatGPT_syntheic-Performance ranking can be ordered as original performance> ChatGPT performance > CT-GAN performance > Faker performance.

Table4: Effect of Number of Tree models on predictive performance

Number of tree model	Faker mean accuracy	GenAI mean accuracy	Hypothesis Test result
5	0.521	0.542	One-tailed p-value (GenAI > CTGAN): 0.011 Result: Reject H0. Conclusion: GenAI performs significantly better than Faker.
10	0.466	0.507	One-tailed p-value (GenAI > CTGAN): 0.011 Result: Reject H0. Conclusion: GenAI performs significantly better than Faker.
15	0.477	0.524	One-tailed p-value (GenAI > CTGAN): 0.0 Result: Reject H0. Conclusion: GenAI performs significantly better than Faker.
20	0.475	0.541	One-tailed p-value (GenAI > CTGAN): 0.0 Result: Reject H0. Conclusion: GenAI performs significantly better than Faker.

The table 4 shows that the hypothesis test was conducted for each of the tree model. Except single tree model, in all other tree models GenAI based synthetic data significantly performed better.

Table5: Effect of Number of Tree models on predictive performance

Number of tree model	CTGAN mean accuracy	GenAI mean accuracy	Hypothesis Test
1	0.491	0.508	Paired t-test statistic: 0.876 One-tailed p-value (GenAI > CTGAN): 0.194 Result: Fail to reject H0. Conclusion: No significant evidence that GenAI performs better than CTGAN.
5	0.493	0.516	Paired t-test statistic: 1.424 One-tailed p-value (GenAI > CTGAN): 0.082 Result: Fail to reject H0. Conclusion: No significant evidence that GenAI performs better than CTGAN.
10	0.522	0.497	Paired t-test statistic: -1.525 One-tailed p-value (GenAI > CTGAN): 0.931 Result: Fail to reject H0. Conclusion: No significant evidence that GenAI performs better than CTGAN.
15	0.508	0.545	Paired t-test statistic: 1.919 One-tailed p-value (GenAI > CTGAN): 0.032 Result: Reject H0. Conclusion: GenAI performs significantly better than CTGAN.
20	0.511	0.541	Paired t-test statistic: 2.096 One-tailed p-value (GenAI > CTGAN): 0.022 Result: Reject H0. Conclusion: GenAI performs significantly better than CTGAN.
25	0.508	0.545	Paired t-test statistic: 2.65 One-tailed p-value (GenAI > CTGAN): 0.006 Result: Reject H0. Conclusion: GenAI performs significantly better than CTGAN.

6. CONCLUSION

The aim of this paper is to demonstrate how easy and accurate the GenAI based synthetic data.

This paper use chatGPT tool for generating the synthetic data. When input submitted, the tool generated the synthetic data immediately in seconds in required format. This type of agility will help researchers to test and validate results in advance before collecting the original data. Hence, the overall expected results can be estimated easily. Machine Learning models require large amount of data. In the health care area, getting large data consumes time and money. And difficult also. So in order to save money and time, GenAI data can be used as an alternative to the original data under certain conditions the conditions include;

1. High level privacy constraints
2. When adequate data is not available

3. When you need to do non-clinical research

Limitations: This study carried out with chatGPT tool. However, other GenAI tools are available. Experimenting with them also improve the depth of the study.

7. REFERENCES

[1] Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro (2023),Scoping review: "Privacy and utility in synthetic healthcare data" *PubMed*. Availble at https://pubmed.ncbi.nlm.nih.gov/39870798/

[2] DataIntelo. (2024). *Synthetic data in healthcare market outlook 2025–2033* (Market report). https://dataintelo.com/report/synthetic-data-in-healthcare-market

[3] Fang, M. L., Dhami, D. S., & Kersting, K. (2022). *DP-CTGAN: Differentially private tabular GAN*. In M. Michalowski, S. S. R. Abidi, & S. Abidi (Eds.), *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine (AIME 2022) – Proceedings* (pp. 178–188). Springer. https://doi.org/10.1007/978-3-031-09342-5_17

[4] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). *Heart disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

[5] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer and L. Y. Chen, "CTAB-GAN+: Enhancing tabular data synthesis," Frontiers in Big Data, vol. 6, p. 1296508, Jan. 2024, doi: 10.3389/fdata.2023.1296508.

[6] Umesh, C., Mahendra, M., Bej, S., Wolkenhauer, O., & Wolfien, M (2024). Challenges and applications in generative AI for clinical tabular data in physiology. *Pflügers Archiv – European Journal of Physiology*.

[7] Ahmed, H. A., Nepomuceno, J. A., Vega-Márquez, B., et al. (2025). "Synthetic Data Generation for Healthcare: Exploring Generative Adversarial Networks Variants for Medical Tabular Data." *International Journal of Data Science and Analytics*, Springer.

[8] Ghosheh, M., Murtaza, S., & others (2025). "A Systematic Review of Privacy-Preserving Techniques for Synthetic Tabular Health Data." *Discover Data* (Springer).

[9] Karmakar, A., Shaw, A., Rakshit, S., Chakraborty, S., Biswas, S., Sahoo, S., & Biswas, S. (2025). *The role of generative AI in medical image synthesis: A review*. *Discover Applied Sciences, 7*, Article 714. https://doi.org/10.1007/s42452-025-07714-7

[10] Zhang, W., Liu, R., Zhu, X., et al. (2025). "Enhancing Privacy Protection of Physical Examination Data through Synthetic Algorithms Based on Differential Privacy." *BMC Medical Informatics and Decision Making* (Springer Nature).