Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241100

Use and Evaluation of GANs for Synthetic Data Generation in Pharmacogenetics

Dominic AESCHBACHER[a], Jessica MEISNER[a], Marko MILETIC[a] and Murat SARIYAR[a,1]
[a]*Bern University of Applied Sciences, Switzerland*
ORCiD ID: Murat Sariyar https://orcid.org/0000-0003-3432-2860

Abstract. Pharmacogenetics (PGx) explores the influence of genetic variability on drug efficacy and tolerability. Synthetic Data Generation (SDG) has emerged as a promising alternative to the labor-intensive process of collecting real-world PGx data, which is required for high-qualitative prediction models. This study investigates the performance of two Generative Adversarial Network (GAN) models, CTGAN and CTAB-GAN+, in generating synthetic PGx data. The benchmarking is based on utility metrics (Hellinger distance and Random Forest accuracy) and ε-identifiability. Results demonstrate that synthetic data generated by CTAB-GAN+ can surpass the original dataset in terms of utility. For instance, CTAB-GAN+ achieves higher Random Forest accuracy compared to the original data, indicating better predictive performance. These improvements suggest that synthetic data not only capture the essential patterns of the original data but also enhance model generalization and prediction capabilities, providing a more robust training ground for machine learning models. Consequently, SDG offers a promising solution to address data scarcity and imbalance in pharmacogenetic research.

Keywords. Pharmacogenetics (PGx), synthetic data generation, GAN, CTAB-GAN+, tabular data

1. Introduction

The efficacy and tolerance of medications vary among individuals. Pharmacogenetics (PGx) investigates the influence of genetic variations on drug effectiveness and adverse reactions. Variations in genes encoding proteins involved in pharmacologically relevant functions can alter drug metabolism, impacting their efficacy. Pharmacogenetic diagnostics aim to identify genetic predispositions to adverse reactions and therapeutic resistance [1]. A well-known example is the HIV medication Abacavir, which can cause severe reactions, such as fever, pain, or even respiratory distress, in individuals with the HLA-B*5701 allele. Such reactions can be fatal, so it is recommended that patients be tested for this gene before being prescribed Abacavir. If the test is positive, an alternative medication should be considered [2,3].

The Pharmaceutical Care Research Group at the University of Basel has initiated a pharmacogenetic observational study. The study collects data from patients experiencing adverse drug reactions or therapeutic failures to develop a reliable standard procedure for pharmacogenetic testing in hospital pharmacies. By analyzing pharmacogenetic

[1] Corresponding Author: Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

samples, the hereditary components contributing to patient susceptibility to treatment failure or adverse drug reactions are identified, followed by medication recommendations which may include dosage adjustments or alternative drugs. Subsequent consultations assess whether the recommendations were implemented [4].

Recruiting patients and collecting data is labor-intensive, and there is a significant need for pharmacogenetic data, especially to predict whether medication should be discontinued. Predictive models require sufficient data to make accurate predictions. Additionally, imbalanced class counts for the target variable are a common issue, and both data scarcity and imbalance should ideally be addressed. Synthetic Data Generation (SDG) methods offer a solution here. Various methods exist, with Generative Adversarial Networks (GANs) proving particularly effective in recent years. GANs can produce data with high statistical similarity to real data and are also used for anonymization, although their effectiveness in reducing disclosure risk is still debated [5,6]. Large Language Models, although a promising alternative in general, fall short in high dimensional settings due to the limitation of context length [7]. Common potentials and challenges of SDG methods are described in one of our recent papers [8].

This study investigates two GAN models for the PGx use case. First, the two approaches are introduced and benchmarked. Then, the results are presented and discussed in the context of their real-world potential and future steps. The primary research objective is to evaluate the extent to which synthetic data can replace or augment real data in a high-dimensional setting.

2. Methods

In the following sections we will describe the dataset, the SDG methods and the utility metrics as well as the risk measure used for evaluation.

The original dataset consists of 149 patients. It includes patient characteristics such as birthdate, gender, smoking status, medication-related data such as substance, dosage, frequency, and genetic data describing variants and corresponding phenotypes. Due to polypharmacy, patients may appear multiple times in the dataset, resulting in a total of 564 rows. This means that each row is treated as an independent entity, disabling follow up of patients. SDG methods are row-oriented and therefore cannot adequately take these aspects into account without an extension that would be too complex for this setting. For generating synthetic data, two cases are distinguished: (i) A high-dimensional dataset at the level of individual variants per gene (e.g., rs10509681 CYP2C8 A>G and rs11572080 CYP2C8 G>A) with 111 variables in total. This dataset is called “stratipharm” as the variants were determined with this provider. (ii) A dataset that only contains phenotypes, named “phenotype”, resulting in a total of 33 columns.

Two GANs specialized for tabular data, namely Conditional Tabular GAN (CTGAN) [9] and CTAB-GAN+ [10] are employed in the process of synthetic data generation. CTGAN is specifically designed to synthesize tabular data by employing a conditional GAN architecture. This approach addresses challenges inherent in mixed data types (continuous and categorical) and imbalanced class distributions. CTGAN utilizes mode-specific normalization and a conditional generator to produce synthetic data that closely approximates the distribution of real data. It promises to excel in scenarios characterized by highly imbalanced or complex feature sets [9]. CTAB-GAN+ represents an enhanced version of CTGAN tailored for generating synthetic tabular data. It builds upon the CTGAN framework by incorporating additional mechanisms to

improve data generation quality. These enhancements include: (1) integrating downstream losses into conditional GANs to optimize utility in both classification and regression tasks; (2) incorporating Wasserstein loss with gradient penalty to enhance training stability and convergence; (3) introducing novel encoders designed for mixed continuous-categorical variables and variables with skewed distributions; and (4) employing training with differentially private stochastic gradient descent to ensure rigorous privacy guarantees during model development [10].

The benchmarking of SDG methods is evaluated based on three criteria: specific utility, generic utility, and ε-identifiability. For specific utility, we employ random forests to predict medication changes (yes or no). Performance is assessed using Accuracy and AUC (Area under ROC curve). To obtain a comprehensive assessment across all variable types, we bin the numerical columns using quartile binning and subsequently use Hellinger distance to derive an overall metric for the generic utility. To approximate the re-identification risk inherent in synthetic data, we utilize ε-identifiability [11]. This metric measures the likelihood that the i-th record in the original dataset is closer to its nearest synthetic observation ($\hat{r}_i$) than to its nearest real observation (r_i), with ε representing the threshold of proximity. We conduct 10 sampling iterations for each combination of the SDG methods (CTGAN and CTAB-GAN+) with fixed training epochs of 400. The experiment was conducted on a NVIDIA RTX 4500 Ada Generation Graphics Card.

3. Results

Tables 1 and 2 present the general utility and ε-identifiability metrics on the phenotype and stratipharm datasets. Each table compares the two GAN methods with results on the original data, with variations in sample sizes (564 and 2000 samples). The results indicate notable improvements in utility metrics such as RF Acc and RF ROC for CTAB-GAN+ compared to CTGAN across both datasets and sample sizes, albeit with higher ε-identifiability values reflecting increased reidentification risk.

Increasing the sample size from 564 to 2000 in both the phenotype and stratipharm datasets demonstrates that synthetic data generated by CTAB-GAN+ can surpass the original dataset in terms of utility. This is evident from higher RF accuracy (RF Acc) and RF receiver operating characteristic (RF ROC) scores, which indicate better performance in predictive tasks. For example, in the phenotype dataset, CTAB-GAN+ achieves an RF Acc of .872±.021 and an RF ROC of 0.597±0.027 with 2000 samples, compared to the original dataset's RF Acc of .832±.000 and RF ROC of .448±.000. Similarly, for the stratipharm dataset, CTAB-GAN+ records an RF Acc of .902±.010 and an RF ROC of 0.517±0.023 with 2000 samples, outperforming the original dataset's RF Acc of 0.841±0.000 and RF ROC of .510±.000. These improvements suggest that the synthetic data not only captures the essential patterns and structures of the original data but also enhances the model's ability to generalize and predict. This could be due to the synthetic model's ability to generate more balanced and representative samples, reducing biases and overfitting, which are common in smaller, original datasets.

The variability observed in Tables 1 and 2 underscores the reliability and uniformity of SDG methods across diverse sample sizes and datasets. Specifically, the Hellinger distance exhibits consistent values across varying sample sizes and methods, indicating a stable statistical resemblance between synthetic and original data distributions. However, notable fluctuations are observed in RF Acc and RF ROC scores, with

improvements generally observed as sample sizes increase. In general, CTAB-GAN+ demonstrates higher variability compared to CTGAN, potentially linked to the complexity of the model, which tends to stabilize with larger training datasets. This complexity arises from CTAB-GAN+'s enhanced mechanisms aimed at optimizing data generation quality, including advanced loss functions and encoding techniques.

Table 1. Utility and ε-identifiability for the phenotype dataset variant.

method	samples	Hellinger distance	RF Acc	RF ROC	ε-identifiability
original	564	-	.832±.000	.448±.000	-
CTGAN	564	.138±.000	.803±.048	.517±.057	.003±.001
	2000	.130±.001	.782±.017	.490±.019	.004±.001
CTAB-GAN+	564	.120±.008	.864±.038	.579±.065	.008±.003
	2000	.109±.009	.872±.021	.597±.027	.013±.001

Table 2. Utility and ε-identifiability for the stratipharm dataset variant.

method	samples	Hellinger distance	RF Acc	RF ROC	ε-identifiability
original	564	-	.841±.000	.510±.000	-
CTGAN	564	.083±.001	.880±.030	.496±.004	.000±.001
	2000	.130±.001	.782±.017	.490±.019	.004±.001
CTAB-GAN+	564	.057±.004	.919±.035	.552±.070	.012±.002
	2000	.053±.003	.902±.010	.517±.023	.016±.001

4. Discussion and Conclusions

The results highlight that CTAB-GAN+ effectively captures the essential patterns and structures of the original data, enhancing model generalization and prediction capabilities. This superior performance is evident in the higher Random Forest accuracy and ROC scores achieved by CTAB-GAN+ across both the phenotype and stratipharm datasets. The ability of CTAB-GAN+ to generate synthetic data that closely resembles real data allows for improved training of machine learning models, leading to better predictive outcomes. This suggests that synthetic data can serve as a viable alternative to real data, particularly in scenarios where data collection is challenging or limited by privacy concerns. Despite the promising results, there are considerations regarding the reidentification risk associated with synthetic data. The higher ε-identifiability values for CTAB-GAN+ indicate increased reidentification risk compared to CTGAN. This reflects a trade-off between utility and privacy, where enhancements in data quality and utility come at the cost of increased potential for reidentifying individuals within the dataset. This trade-off is a critical factor to consider when deploying synthetic data in real-world applications, especially in sensitive fields like PGx where patient privacy is paramount.

Despite the promising advancements in SDG using CTAB-GAN+ for pharmacogenetic research, several challenges remain that warrant attention. Firstly, scalability and generalizability are crucial aspects that need refinement. Although CTAB-GAN+ shows promising results with increased sample sizes, ensuring scalability across diverse datasets and different pharmacogenetic applications remains a key hurdle. The robustness of SDG methods under varying data distributions and complexities needs to be rigorously evaluated. Secondly, the interpretability of synthetic data outputs is another area that requires attention. While CTAB-GAN+ excels in mimicking the statistical properties of real data, the interpretability of generated synthetic samples and their clinical relevance must be thoroughly validated. Establishing methods to validate

the biological plausibility of synthetic data outputs, especially in complex pharmacogenetic scenarios involving multiple genetic variants and clinical variables, remains an ongoing challenge. Lastly, the integration of synthetic data into real-world clinical decision-making processes poses implementation challenges. Despite its potential to mitigate data scarcity and improve predictive model training, the acceptance and adoption of synthetic data in clinical practice require robust validation and regulatory approval. Developing standardized guidelines and frameworks for the ethical use of synthetic data in pharmacogenetics is imperative. This includes establishing benchmarks for comparing synthetic and real data performance and ensuring transparency in the methods used for data generation and validation.

In conclusion, while CTAB-GAN+ shows promising advancements in SDG for pharmacogenetic research, addressing challenges related to privacy, scalability, interpretability, and clinical integration is crucial for realizing its full potential. Future research efforts should focus on refining GAN methodologies, enhancing privacy-preserving techniques, validating clinical relevance, and fostering regulatory acceptance to pave the way for safe and effective deployment of synthetic data in personalized medicine.

References

[1] Ensom MH, Chang TK, Patel P. Pharmacogenetics: the therapeutic drug monitoring of the future? Clin Pharmacokinet 2001;40:783–802. https://doi.org/10.2165/00003088-200140110-00001.
[2] Ma JD, Lee KC, Kuo GM. HLA-B*5701 testing to predict abacavir hypersensitivity. PLoS Curr 2010;2:RRN1203. https://doi.org/10.1371/currents.RRN1203.
[3] Mallal Simon, Phillips Elizabeth, Carosi Giampiero, Molina Jean-Michel, Workman Cassy, Tomažič Janez, et al. HLA-B*5701 Screening for Hypersensitivity to Abacavir. New England Journal of Medicine 2008;358:568–79. https://doi.org/10.1056/NEJMoa0706135.
[4] Study Details | Pharmacogenetic Testing of Patients With Unwanted Adverse Drug Reactions or Therapy Failure | ClinicalTrials.gov n.d. https://clinicaltrials.gov/study/NCT04154553 (accessed June 28, 2024).
[5] Emam KE, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O'Reilly Media, Incorporated; 2020.
[6] Stadler T, Oprisanu B, Troncoso C. Synthetic Data – Anonymisation Groundhog Day, 2022, p. 1451–68.
[7] Zhao Z, Birke R, Chen L. TabuLa: Harnessing Language Models for Tabular Data Synthesis 2023.
[8] Miletic M, Sariyar M. Challenges of Using Synthetic Data Generation Methods for Tabular Microdata. Applied Sciences 2024;14:5975. https://doi.org/10.3390/app14145975.
[9] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN 2019. https://doi.org/10.48550/arXiv.1907.00503.
[10] Zhao Z, Kunar A, Birke R, Chen LY. CTAB-GAN+: Enhancing Tabular Data Synthesis 2022. https://doi.org/10.48550/arXiv.2204.00401.
[11] Yoon J, Drumright LN, van der Schaar M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE J Biomed Health Inform 2020;24:2378–88. https://doi.org/10.1109/JBHI.2020.2980262.