Benchmarking Tabular Generative Models for Clinical Data Synthesis

Matteo Bonazzi*§, Maria Giulia Bacalini†‡, Enrico Giampieri*, Raffaele Lodi†‡, Gastone Castellani*, Claudia Sala*,
*Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, Italy
†Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna 40126, Italy
‡IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna 40139, Italy
§Corresponding author: matteo.bonazzi5@unibo.it

***Abstract*—Synthetic data generation (SDG) has emerged as a strategy to address data scarcity, class imbalance, and privacy constraints in medical machine learning. However, most tabular generative models are evaluated on large, balanced benchmarks, while their behavior on small, imbalanced, and privacy-sensitive clinical datasets has received limited attention.**

In this work, we conduct a comparative study of tabular generative models on a blood-chemistry cohort from the Alzheimer's Disease Neuroimaging Initiative (ADNI), characterized by limited sample size and uneven class distribution. We compare GAN-based (CTGAN, CTAB-GAN), VAE-based (TVAE), diffusion-based approaches (TabDDPM, STaSy), and Forest Diffusion, a regression tree-based framework leveraging conditional flow matching (CFM). Evaluation encompasses distributional fidelity, correlation and discrimination metrics, empirical privacy risks, and computational cost.

TVAE achieves the shortest training time (35.9 s), while Forest Diffusion attains the lowest Wasserstein distance (1.05) and shares the highest coverage (0.98) with TVAE. CTAB-GAN attains the highest distribution score (80), while TabDDPM achieves the highest correlation score (54). The highest discrimination performance (78) is observed for CTAB-GAN.

Privacy risks vary across methods: CTGAN exhibits the lowest risks, whereas TabDDPM shows consistently higher linkability and inference risks relative to the other approaches. To further assess class-specific fidelity, we compute the KL divergence between real and synthetic age distributions stratified by diagnosis, given the central role of age in Alzheimer's disease progression.

Overall, the results indicate that TVAE and Forest Diffusion provide competitive performance across the evaluated quality and privacy metrics, while CTGAN and STaSy consistently underperform, underscoring the importance of benchmarking generative models under realistic clinical conditions.

***Index Terms*—Synthetic Data Generation, Diffusion Models, Flow Matching, Generative Adversarial Networks**

I. Introduction

Synthetic data generation (SDG) is an emerging field in the application of artificial intelligence, especially in the medical field. By using deep learning (DL) methods to produce synthetic datasets that preserve key statistical properties of real data, SDG can alleviate data scarcity and class imbalance and enable safer data sharing [1].

These challenges are particularly pronounced in the medical domain, where data collection is costly, rare conditions create severe class imbalance, and access to patient records is restricted by privacy regulations. Synthetic data therefore offers three concrete benefits in clinical settings: reducing data scarcity by augmenting or replacing limited real samples, mitigating privacy risks by providing non-identifiable surrogate data while retaining utility, and addressing class imbalance by generating additional examples of rare conditions. Motivated by these reasons, we analyze the performance of different generation methods on a real-world dataset containing blood chemistry examination results of patients affected by Alzheimer's disease.

Several different studies have introduced tabular data generation methods based on different architectures: GANs [2] [3], VAE [3], diffusion models [4] [5] and flow-based models [6]. However, most of these methods have been evaluated on large, balanced, non-medical tabular benchmarks (thousands to tens of thousands of records). In contrast, clinical datasets are frequently small (fewer than 1000 entries) and highly imbalanced due to the presence of rare pathologies. The goal of this work is to explore the behavior and robustness of modern tabular generative models under such data-scarce and imbalanced conditions.

In this work, we systematically compare several representative generative approaches, including GAN-, VAE-, and diffusion-based models, within a small and imbalanced medical dataset setting, evaluating quality and privacy metrics. All models are trained and evaluated under identical experimental conditions, with consistent training splits, and evaluation protocols.

II. Generation methods

Multiple generation methods have been trained on the ADNI dataset, spanning across different architectures such as Generative Adversarial Networks (GAN), diffusion models and Variational Autoencoders (VAE).

A. CTGAN

CTGAN [3] is a GAN-based model for tabular data generation designed to handle multi-modal, skewed numerical distributions and rare categorical values. As in standard GANs, it consists of a generator G_θ and a discriminator D_ϕ, optimized adversarially so that the generator produces synthetic samples that the discriminator cannot reliably distinguish from real data.

CTGAN uses mode-specific normalization (MSN) to model each numerical feature with a Variational Gaussian Mixture (VGM). Each value x is assigned to the most probable component k^* and normalized as

$$\tilde{x} = \frac{x - \mu_{k^*}}{\sigma_{k^*}}, \tag{1}$$

while categorical features are one-hot encoded and concatenated with the continuous ones.

To better represent rare categories, CTGAN uses conditional generation: a conditional vector is provided as input to the generator, and an auxiliary cross-entropy term encourages consistency between the condition and the generated categorical output. Training follows the Wasserstein GAN with Gradient Penalty (WGAN-GP) objective [7], which improves stability and mitigates mode collapse.

B. CTAB-GAN

CTAB-GAN [2] extends CTGAN by improving both preprocessing and loss design. In addition to Mode-Specific Normalization (MSN), it introduces a General Transform (GT) that applies min–max scaling to map simple numerical features to the $[-1, 1]$ range, consistent with the *tanh* output activation. MSN is used for multi-modal variables, while GT is applied to features with simpler distributions.

CTAB-GAN also augments the WGAN-GP objective with information and downstream losses. The information loss encourages the generated data to match the first- and second-order statistics of the real data, while the downstream loss uses an auxiliary classifier to promote samples useful for the target task. As in CTGAN, the generator is additionally trained with a cross-entropy term on the conditional vector, whereas the discriminator objective remains unchanged.

C. TVAE

TVAE [3] is a variational autoencoder architecture adapted for tabular data. An encoder $q_\phi(z \mid x)$ maps each input x to a latent variable z, and a decoder $p_\theta(x \mid z)$ reconstructs x from z. The latent space follows a Gaussian prior, while categorical features are modeled with categorical distributions.
Model parameters are learned by maximizing the Evidence Lower Bound (ELBO) via gradient-based optimization using the reparameterization trick. The ELBO for a single example is

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}\big(q_\phi(z \mid x) \,\|\, p(z)\big). \tag{2}$$

Synthetic samples are generated by sampling $z \sim p(z)$ and decoding $x \sim p_\theta(\cdot \mid z)$.

D. TabDDPM

Tabular Denoising Diffusion Model [4] (TabDDPM) applies diffusion denoising probabilistic modeling to tabular data. A forward process progressively corrupts samples $x_0 \sim q(x_0)$ with noise according to a variance schedule $\{\beta_t\}_{t=1}^T$. The transition $q(x_t \mid x_{t-1})$ depends on the feature type: Gaussian noise for continuous variables,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I), \tag{3}$$

and multinomial noise for categorical variables,

$$q(x_t|x_{t-1}) = \mathrm{Cat}(x_t; (1-\beta_t)x_{t-1} + \beta_t/N_{\text{class}})\,. \tag{4}$$

A neural network is trained to approximate the reverse process and iteratively denoise a sample from x_T to obtain synthetic data. Training minimizes a timestep-dependent loss combining a mean squared error term for continuous variables and a Kullback–Leibler divergence term for categorical variables:

$$L_t = L_t^{\text{MSE}} + \frac{\sum_{i=1}^{N_{\text{class}}} L_t^{\text{KL}}}{N_{\text{class}}}. \tag{5}$$

E. STaSy

STaSy [5] is a score-based diffusion model that formulates data generation through stochastic differential equations (SDEs). The forward process perturbs data via

$$\mathrm{d}x = \mu_t(x)\,\mathrm{d}t + g(t)\,\mathrm{d}w, \tag{6}$$

while the reverse SDE,

$$\mathrm{d}x = \big(\mu_t(x) - g^2(t)\,\nabla_x \log p_t(x)\big)\,\mathrm{d}t + g(t)\,\mathrm{d}w, \tag{7}$$

depends on the score $\nabla_x \log p_t(x)$, which is approximated by a neural network $s_\theta(x, t)$. The network is trained via denoising score matching:

$$\arg\min_\theta \mathbb{E}\Big[\lambda(t)\,\|s_\theta(x(t), t) - \nabla_{x(t)} \log p(x(t) \mid x(0))\|_2^2\Big]. \tag{8}$$

STaSy further employs self-paced learning to improve training stability. Synthetic samples are generated by numerically solving the learned reverse SDE.

F. Forest Diffusion

Forest Diffusion generates synthetic tabular data using gradient boosted regression trees (XGBoost), leveraging their strong performance on tabular tasks [6]. The model adopts Independent Coupling Conditional Flow Matching (IC-CFM) [8], extending flow-matching to tree-based learners.
Given data $X \in \mathbb{R}^{n\times d}$ and Gaussian noise $\mathcal{Z}$, IC-CFM constructs intermediate states

$$X(t) = tX + (1-t)\mathcal{Z}, \quad t \in \left\{0, \frac{1}{n_t}, \ldots, 1\right\}, \tag{9}$$

defining discretized trajectories between noise and data. At each timestep, a boosted tree model is trained to approximate the conditional vector field by minimizing

$$\min_f \; \big\|f\big(X(t)\big) - \big(X - \mathcal{Z}\big)\big\|_2^2\,. \tag{10}$$

To account for the lack of mini-batch training in tree-based models, the dataset is paired with multiple noise realizations. Rather than encoding time explicitly, a separate model is

TABLE I
ADNI DATASET FEATURE OVERVIEW

Feature	Description	Type
Age	Patient age at examination (yr)	Num.
Albumin	Albumin blood protein (g/dL)	Num.
Creat	Creatinine (mg/dL)	Num.
Alp	Alkaline phosphatase (U/L)	Num.
Totchol	Total cholesterol (mg/dL)	Num.
Sbp	Systolic blood pressure (mmHg)	Num.
Bun	Blood urea nitrogen (mg/dL)	Num.
Uap	Uric acid plasma (mg/dL)	Num.
Lymph	Lymphocytes among white blood cells (%)	Num.
Mcv	Mean corpuscular volume (fL)	Num.
Wbc	White blood cell count ($\times 10^9$/L)	Num.
Glucose	Blood sugar level (mg/dL)	Num.
Dx.bl	Patient diagnosis (CN, LMCI, AD)	Cat.

trained for each noise level, yielding a discretized time-dependent vector field ($v_\theta(t, x)$). Synthetic samples are generated by solving the learned ordinary differential equation

$$\mathrm{d}x = v_\theta(t, x)\,\mathrm{d}t \tag{11}$$

using an explicit Euler solver.

III. Materials and Methods

A. Alzheimer's disease Neuroimaging Initiative Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

A subset of data containing results of hematological examinations has been extracted from ADNI: the dataset is composed of 561 patients divided into cognitive normal (CN, 58 entries), late mild cognitive impairment (LMCI, 392 entries) and Alzheimer's disease (AD, 111 entries).

Columns containing examination dates, patient identifiers, and acquisition sites were removed, as they are not relevant to the generative modeling objective. Additionally, the features *lncrp* and *lncreat* were discarded, since the former contained a large proportion of missing values (251 entries), while the latter was strongly correlated with *creat* (defined as $lncreat = \ln(creat{+}1)$), and therefore redundant. Rows containing missing values were removed rather than imputed. This resulted in the exclusion of 20 patients (approximately 3.6% of the dataset), yielding a final cohort of 541 individuals. Given the small proportion of removed samples and the potential bias introduced by imputation procedures in generative modeling settings, complete-case analysis was adopted for simplicity and consistency across models. A summary of features present in the dataset is reported in Table I.

The dataset contains 12 numerical features and one categorical feature (baseline diagnosis, Dx.bl). Categorical features were factorized, while numerical variables were retained in their original scale prior to model-specific preprocessing.

B. Evaluation metrics

Evaluation is organized into two categories: quality metrics and privacy metrics. Quality metrics quantify statistical similarity between real and synthetic data, while privacy metrics estimate the risk that synthetic data could reveal information about real subjects. We compute all quality metrics on an independent holdout set not used during model training, while privacy risks are computed on training data. For each model, we report the mean and standard deviation of the metrics computed over five-fold cross-validation, with five independent synthetic generations per fold.

1) Quality metrics: To the authors' best knowledge, there is no standard in the evaluation of synthetic tabular data. Three normalized (0-100) scores have been computed using the package Syndat, developed under the NFDI4Health project and extended in SYNTHIA, as an effort to introduce standardized and easily interpretable metrics. Furthermore, three metrics from [6] have also been computed, allowing for a better evaluation of data quality.

For scores annotated with an upward arrow (↑), higher values indicate better quality, while for metrics annotated with a downward arrow (↓) lower values indicate better quality:

- Distribution score (0-100, ↑): measure of marginal distribution similarity across features, computed as a normalized score inversely related to the mean Jensen–Shannon divergence over all features.
- Discrimination score (0-100, ↑): measures how difficult it is to distinguish synthetic from real samples, based on the ROC-AUC of a Random Forest classifier trained to discriminate real from synthetic data. The score is normalized so that higher values correspond to lower discriminability.
- Correlation score (0-100, ↑): normalized difference between correlation matrices for real and synthetic data.
- Wasserstein distance (↓): earth mover's distance between the marginal distributions of real and synthetic data, computed as the L_1 Wasserstein distance averaged over all features.
- Coverage (0-1, ↑): fraction of real samples that have at least one synthetic sample within a sphere of radius r, defined as the distance to the k-th nearest neighbor [6].
- Training time (↓): wall-clock training time to reach stopping criterion.

All the quality metrics have been computed on a separate set of data not used during training.

2) Privacy metrics: Privacy risks are assessed using three risk scores based on [9] and the Syndat framework:

- Singling out risk (0-100, ↓): risk that a record in the training set is uniquely identifiable by an attacker using a combination of attributes.
- Linkability risk (0-100, ↓): risk of linking together two or more data records belonging to the same patient or group.

TABLE II
QUALITY METRICS COMPUTED ON THE ADNI TEST SET. VALUES ARE REPORTED AS MEAN (STANDARD DEVIATION) OVER 5-FOLD CROSS-VALIDATION WITH 5 GENERATIONS PER FOLD. BEST IN BOLD.

Method	W Distance ↓	Coverage ↑	Time (s)	Distribution ↑	Correlation ↑	Discrimination ↑
TabDDPM	1.16 (0.04)	0.96 (0.02)	51 (2)	77 (1)	**54 (4)**	73 (10)
CTGAN	1.56 (0.08)	0.74 (0.08)	67.2 (0.3)	70 (2)	40 (5)	20 (6)
TVAE	1.08 (0.05)	**0.98 (0.02)**	**35.9 (0.1)**	78 (2)	53 (3)	69 (9)
CTAB-GAN	1.12 (0.04)	0.95 (0.03)	50 (1)	**80 (1)**	53 (4)	**78 (10)**
STaSy	1.32 (0.06)	0.90 (0.06)	323 (7)	76 (2)	52 (5)	56 (13)
Forest Diffusion	**1.05 (0.04)**	**0.98 (0.01)**	40 (3)	77 (2)	53 (4)	65 (11)

- Inference risk (0-100, ↓): risk of an attacker confidently guessing the value of an unknown attribute in the original data record.

IV. EXPERIMENTAL SETUP

The real dataset was evaluated using a five-fold cross-validation (CV) strategy, where splits were generated in a stratified manner based on the diagnosis variable to preserve class proportions across folds. For each fold, generative models were trained on the corresponding training partition, and evaluation was performed on the held-out test split. To account for stochasticity in the generation process, five independent synthetic datasets, each matching the size of the training set, were generated per fold.

TVAE and CTGAN were implemented using the Synthetic Data Vault (SDV) framework [10], while CTAB-GAN was implemented using the official repository provided in [2]. Preliminary experiments using the default number of training epochs resulted in low-quality synthetic samples. Increasing the number of epochs to 2000 led to stable convergence and substantially improved sample realism. Therefore, all GAN-based models and TVAE were trained for 2000 epochs, while keeping all other hyperparameters at their default values.
TabDDPM and STaSy were implemented following the official repositories from [4] and [5]. For diffusion-based models, we adopted the hyperparameter configuration proposed by [6], which has been shown to provide stable training on smaller tabular datasets. Forest Diffusion was implemented using the default hyperparameters from [6]. Across all models, hyperparameter adjustments were limited to training duration to ensure stable convergence on the small dataset. No architecture-specific tuning or data-driven optimization was performed.

All experiments were conducted locally in Python. Models supporting GPU acceleration were trained on an NVIDIA RTX 6000 Ada GPU, while CPU-based models were run on a 32-core processor. Training time was measured as wall-clock time under identical hardware conditions for all methods.

V. RESULTS

Results for quality metrics are reported in Table II and privacy metrics in Table III. All values are reported as mean (std) over five-fold cross-validation, with five independent synthetic generations per fold.

Among GAN-based approaches, CTGAN and CTAB-GAN show markedly different performance. CTGAN exhibits relatively poor distributional fidelity, with the highest Wasserstein distance (1.56 ± 0.08), low coverage (0.74 ± 0.08), and the lowest discrimination score (20 ± 6), indicating limited similarity between synthetic and real samples. CTGAN reports the lowest privacy risks across all metrics, although this may be partly explained by its low data fidelity.

TABLE III
PRIVACY METRICS ON THE ADNI TRAIN SET. VALUES ARE REPORTED AS MEAN (STANDARD DEVIATION) OVER 5-FOLD CROSS-VALIDATION WITH 5 GENERATIONS PER FOLD. BEST IN BOLD.

Method	Singling-Out ↓	Linkability ↓	Inference ↓
TabDDPM	13 (6)	30 (11)	49 (7)
CTGAN	**8 (3)**	**2 (1)**	**19 (2)**
TVAE	10 (3)	2 (1)	22 (2)
CTAB-GAN	9 (3)	2 (1)	20 (1)
STaSy	12 (5)	2 (1)	23 (1)
Forest Diffusion	11 (4)	6 (3)	31 (4)

CTAB-GAN achieves improved performance over CTGAN, with lower Wasserstein distance (1.12 ± 0.04), higher coverage (0.95 ± 0.03), and the best distribution score (80 ± 1). It also attains the highest discrimination score (78 ± 10), suggesting high utility. Privacy risks remain relatively low, with competitive linkability and inference risks, though not the absolute lowest.

TabDDPM reports competitive Wasserstein distance (1.16 ± 0.04) and high coverage (0.96 ± 0.02), achieving the highest correlation score (54 ± 4) and strong discrimination performance (73 ± 10). Its training time (51 ± 2 s) is moderate compared to other methods. However, TabDDPM exhibits the highest linkability (30 ± 11) and inference (49 ± 7) risks, along with elevated singling-out risk (13 ± 6), suggesting a tendency toward overfitting.

STaSy attains a Wasserstein distance of 1.32 ± 0.06 and coverage of 0.90 ± 0.06, with distribution and correlation scores comparable to other diffusion-based methods. However, it requires substantially longer training time (323 ± 7 s). Privacy risks are moderate, with low linkability and inference values comparable to most baselines, and a singling-out risk of 12 ± 5.

TVAE achieves low Wasserstein distance (1.08 ± 0.05) and the highest coverage (0.98 ± 0.02, tied), while also being the fastest method (35.9 ± 0.1 s). Its distribution (78 ± 2) and correlation (53 ± 3) scores are competitive with the strongest baselines. In privacy evaluation, TVAE shows low singling-out risk (10 ± 3), very low linkability (2 ± 1), and moderate inference risk (22 ± 2), providing a favorable trade-off.

Forest Diffusion reports the lowest Wasserstein distance

TABLE IV
KL DIVERGENCE BETWEEN REAL AND SYNTHETIC AGE DISTRIBUTIONS ACROSS DIAGNOSES. LOWER VALUES INDICATE BETTER SIMILARITY; BEST VALUES ARE IN BOLD.

Model	$\mathbf{KL_{AD}}$	$\mathbf{KL_{CN}}$	$\mathbf{KL_{LMCI}}$
TabDDPM	**0.015**	0.037	**0.010**
Forest Diffusion	0.017	**0.033**	0.012
TVAE	0.052	0.070	0.019

(1.05 ± 0.04) and the highest coverage (0.98 ± 0.01, tied), with distribution and correlation scores aligned with other top-performing methods. Its training time (40 ± 3 s) is comparable to TVAE and significantly lower than most diffusion-based approaches. In privacy metrics, Forest Diffusion shows moderate singling-out risk (11 ± 4), higher linkability (6 ± 3) compared to GANs and TVAE, and inference risk of 31 ± 4.

A. Age distribution analysis

To assess distributional fidelity, we compared age distributions (see Figure 1) for the top three models—TVAE, TabDDPM, and Forest Diffusion—selected based on their overall performance in terms of Wasserstein distance and coverage. Age was chosen due to its central role in Alzheimer's disease progression and clinical diagnosis.

KL divergence between real and synthetic age distributions (Table IV) provides a quantitative measure of similarity. Forest Diffusion achieves the lowest KL divergence for the CN group and competitive performance for AD, while TabDDPM attains the best results for AD and LMCI.

Visual inspection of the Gaussian kernel density estimates in Figure 1 supports these findings. Deviations from the real distribution are more pronounced in the less represented diagnostic groups (CN and AD), whereas all models closely match the LMCI distribution. This highlights the impact of class imbalance on generative performance, with better fidelity observed in the more represented class.

VI. DISCUSSION

Across all methods, no single model achieves the best performance across all quality and privacy metrics simultaneously. However, some models substantially underperform with respect to the best-performing baselines. CTGAN shows poor data fidelity, reflected by the highest Wasserstein distance, lowest coverage, and lowest discrimination score. Although measured privacy risks for CTGAN are the lowest among all models, these results should be interpreted as an effect of low data fidelity. Similarly, STaSy underperforms in Wasserstein distance, coverage and training time with respect to all the other models. Moreover, STaSy shows the highest variability in discrimination score, suggesting instability in the generation process.

CTAB-GAN improves on CTGAN performance across all metrics, achieving the highest distribution and discrimination scores, although it still shows slightly lower coverage than the best-performing baselines.

TabDDPM achieves the highest correlation score across all models, while maintaining competitive performance across the remaining quality metrics. Moreover, TabDDPM shows high privacy risk across all categories, suggesting possible overfitting to the training data.

TVAE and Forest Diffusion provide a balanced performance: both models show comparable performance across all quality metrics, with only a slight increase in inference and linkability risk for Forest Diffusion.

This study has some limitations that should be considered. The analysis is conducted on a single ADNI dataset; however, this is a deliberate choice to provide a controlled and clinically realistic case study rather than to derive universally generalizable conclusions. While no extensive hyperparameter tuning is performed, all models are trained under consistent and literature-supported settings to ensure a fair and reproducible comparison. Finally, the evaluation focuses on statistical fidelity and privacy, which are prerequisites for synthetic data use, while downstream utility is left for future work.

VII. CONCLUSIONS

In this work, we evaluated the performance of multiple tabular data generation methods in a clinically realistic setting characterized by limited sample size and class imbalance. While most generative models are commonly assessed on large, balanced benchmark datasets, real-world medical data often present substantially more challenging conditions. To address this gap, we conducted a comparative study in which all models were trained and evaluated under identical conditions on a blood chemistry dataset from ADNI.

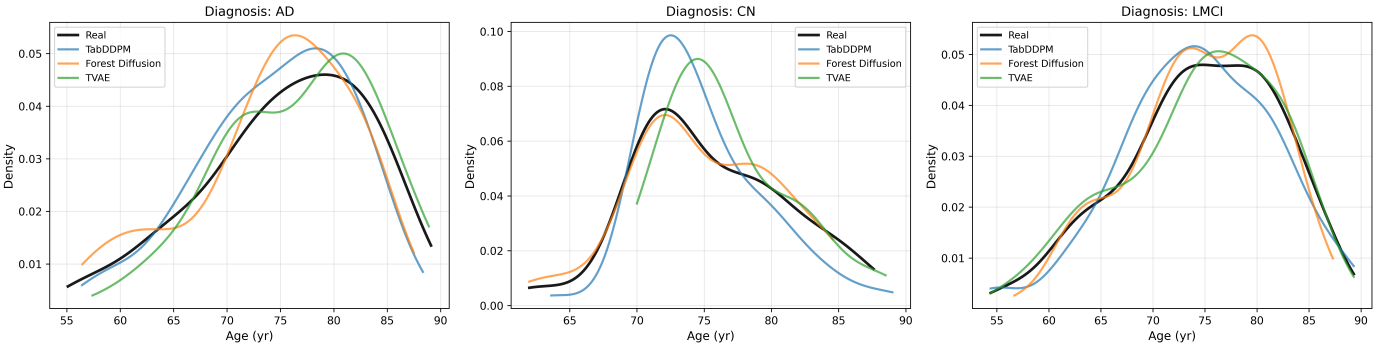
Our results indicate that no single model consistently outperforms the others across all metrics. TVAE achieves the shortest training time and shares the highest coverage with Forest Diffusion, while Forest Diffusion attains the lowest Wasserstein distance. CTAB-GAN obtains the highest distribution and discrimination scores, while TabDDPM achieves the highest correlation score. Privacy metrics show clearer differences across methods: CTGAN exhibits the lowest risks, whereas TabDDPM shows the highest singling-out, linkability, and inference risks, potentially indicating partial memorization of the training data.

The age distribution analysis indicates that class-specific variability persists even among the top-performing models. TabDDPM better reproduces age distributions in the AD and LMCI groups, while Forest Diffusion achieves the lowest divergence for the CN group.

Overall, our findings indicate that small and imbalanced medical datasets negatively affect the performance of several models: CTGAN and STaSy show the most consistent under-performance across quality metrics, CTAB-GAN exhibits reduced coverage and higher Wasserstein distance, and TabDDPM presents notable privacy risks. In contrast, our findings highlight TVAE and Forest Diffusion as more robust, maintaining a good balance between synthesis quality and privacy. These results highlight the importance of evaluating generative methods on clinically realistic datasets.

Future work will extend this analysis to additional medical

Fig. 1. Gaussian kernel density estimates of age distributions across diagnostic groups for real and synthetic samples generated by TVAE, TabDDPM, and Forest Diffusion. Visual alignment between real and synthetic curves indicates the ability of each model to preserve clinically relevant age patterns.



datasets, investigate subgroup-specific generation, and explore formal privacy-preserving mechanisms, as well as assessing synthetic data utility in downstream tasks.

Acknowledgments

This was conducted within the SYNTHIA project. SYNTHIA (Synthetic Data Generation framework for integrated validation of use cases and Al healthcare applications) is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them. The authors are grateful to the SYNTHIA consortium for its collaborative support.

The authors are grateful to AINDO for their collaborative support.

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; BristolMyers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

References

[1] F. Jacobs, S. D'Amico, C. Benvenuti, M. Gaudio, G. Saltalamacchia, C. Miggiano, R. De Sanctis, M. G. Della Porta, A. Santoro, and A. Zambelli, "Opportunities and challenges of synthetic data generation in oncology," *JCO Clinical Cancer Informatics*, no. 7, p. e2300045, 2023.

[2] 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. Volume 6 - 2023, 2024. [Online]. Available: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2023.1296508

[3] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

[4] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modeling tabular data with diffusion models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[5] J. Kim, C. Lee, and N. Park, "STasy: Score-based tabular data synthesis," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=1mNssCWt_v

[6] A. Jolicoeur-Martineau, K. Fatras, and T. Kachman, "Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees," in *International conference on artificial intelligence and statistics*. PMLR, 2024, pp. 1288–1296.

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio, "Improving and generalizing flow-based generative models with minibatch optimal transport," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=CD9Snc73AW

[9] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," *Proceedings on Privacy Enhancing Technologies (PoPETs)*, vol. 2023, no. 2, pp. 312–328, 2023.

[10] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410.