Enabling Granular Subgroup Level Model Evaluations by Generating Synthetic Medical Time Series

Mahmoud Ibrahim[1,2,3](✉), Bart Elen[3], Chang Sun[1,2], Gökhan Ertaylan[3], and Michel Dumontier[1,2]

[1] Institute of Data Science, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands
[2] Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands
[3] VITO, Belgium
mahmoud.ibrahim@vito.be

Abstract. We present a novel framework for leveraging synthetic ICU time-series data not only to train but also to rigorously and trustworthily evaluate predictive models, both at the population level and within fine-grained demographic subgroups. Building on prior diffusion and VAE-based generators (TimeDiff, HealthGen, TimeAutoDiff), we introduce Enhanced TimeAutoDiff, which augments the latent diffusion objective with distribution-alignment penalties . We extensively benchmark all models on MIMIC-III and eICU, on 24-hour mortality and binary length-of-stay tasks. Our results show that Enhanced TimeAutoDiff reduces the gap between real-on-synthetic and real-on-real evaluation ("TRTS gap") by over 70 %, achieving $\Delta_{TRTS} \leq 0.014$ AUROC, while preserving training utility ($\Delta_{TSTR} \approx 0.01$). Crucially, for 32 intersectional subgroups, large synthetic cohorts cut subgroup-level AUROC estimation error by up to 50 % relative to small real test sets, and outperform them in 72–84 % of subgroups. This work provides a practical, privacy-preserving roadmap for trustworthy, granular model evaluation in critical care, enabling robust reliable performance analysis across diverse patient populations without exposing sensitive EHR data, contributing to the overall trustworthiness of Medical AI.

Keywords: Synthetic Data · Medical Data · Time-series

1 Introduction

Machine learning in critical care increasingly depends on large-scale ICU time-series data (continuous and irregular measurements of vitals, labs, and interventions) to develop predictive tools such as early-warning systems, mortality risk estimators, and length-of-stay predictors. Despite extensive repositories like eICU [2] and MIMIC [1], data sharing remains constrained by privacy regulations and limited access for underrepresented populations. Synthetic ICU time-series data offer a promising solution by providing realistic, privacy-preserving

alternatives to real patient records. Recent advances in generative models, from GANs [6] to diffusion frameworks [10], have shown strong fidelity in reproducing complex temporal patterns of physiological data.

Beyond enabling model training without privacy exposure, synthetic data can also support evaluation when access to real test data is restricted. For instance, a model trained on institutional EHRs may need to be evaluated externally by regulators or collaborators, without sharing the real data. In such cases, a synthetic hold-out set should faithfully approximate the real test distribution so that performance metrics (e.g., AUROC, RMSE) on synthetic samples reflect true performance on real patients. Most prior work, however, emphasizes training utility: the "Train on Synthetic, Test on Real" (TSTR) setup while neglecting evaluation utility, or "Train on Real, Test on Synthetic" (TRTS). For synthetic data to be a reliable test substitute, models evaluated on synthetic hold-outs must achieve performance comparable to those tested on genuine real data.

Beyond global model evaluation, a third, and often overlooked use case for synthetic ICU data is subgroup-level evaluation. ICU cohorts are inherently heterogeneous: patients differ in age, gender, ethnicity, comorbidities, socioeconomic status, and other factors. A high-impact clinical question is "How does my mortality prediction model perform on patients aged over 75, or on Black patients, or on females, or on black females aged over 75?" In this work, the focus is on proper model evaluation, specifically on addressing algorithmic bias through performance analysis on diverse population subgroups, which is recommended by [13]. Unfortunately, real EHR datasets frequently contain only a few hundred samples (or fewer) in any given fine-grained subgroup. Evaluating a downstream model on such a small test set yields wide confidence intervals and inaccurate evaluations that mask how a model truly performs on that subgroup. **Synthetic data, when conditioned on subgroup attributes, can overcome this limitation** by generating large, representative cohorts for fine-grained intersectional groups. This enables precise subgroup evaluation with narrow confidence intervals, improving transparency and fairness. Ultimately, such use of synthetic data contributes to more **trustworthy and equitable AI systems**, where performance metrics truly reflect clinical reality across all patient populations.

Main Contributions This work focuses on using synthetic ICU time-series data for robust and trustworthy model evaluation, both globally and across demographic subgroups. Our goal is to ensure that evaluation metrics computed on synthetic data faithfully reflect true performance on real ICU patients. To this end, we develop an enhanced generative framework and systematically compare it to existing models: TimeDiff, TimeAutoDiff, and HealthGen regarding their ability to produce synthetic data that serve as reliable test proxies.

1. **Enhanced TimeAutoDiff for Evaluation.** We introduce an improved version of the latent diffusion model TimeAutoDiff (optimized for model evaluation purposes), augmented with two distribution-alignment losses (MMD and consistency regularization). These additions reduce the TRTS (Train on

Real, Test on Synthetic) performance gap by over 70% while maintaining TSTR (Train on Synthetic, Test on Real) stability.
2. **Benchmarking Synthetic Hold-Outs.** We systematically compare Enhanced TimeAutoDiff, TimeDiff, TimeAutoDiff, and HealthGen to assess which generators produce synthetic hold-outs that most closely reproduce "real-on-real" (TRTR) evaluation metrics: which synthetic hold-outs can be most reliably trusted as proxies for real test data.
3. **Subgroup-Level Evaluation.** We evaluate whether synthetic data can support fine-grained, demographically stratified analyses by generating large, conditionally sampled subgroups, especially where real test data are insufficient for reliable statistics.
4. **Open-Source Release.** All code, model checkpoints, and subgroup-evaluation pipelines are publicly available at: `github.com/mahmoudibrahim98/icu-auto-diff`.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 summarizes the used architectures. Section 4 details the methodology and the experimental setup .Section 5 presents the experimental results. Section 6 discusses implications and limitations. Finally, Section 7 concludes.

2 Related Work

Very few studies explicitly use synthetic data for model evaluation or for assessing performance across intersectional subgroups. A structured regression framework was introduced in [14]. It pools information across demographic intersections to produce stable, confidence-aware subgroup accuracy estimates. SYNG4ME [15] similarly leverages synthetic cohorts to yield more reliable subgroup metrics than small held-out real subsets. However, this work does not address time-series data or medical datasets.

Synthetic EHR generation must account for the inherent heterogeneity of real-world ICU records, which combine static demographics, high-cardinality categorical codes (e.g., ICD), and multivariate vital-sign time series. Broadly, researchers distinguish four EHR formats: snapshot, aggregated, longitudinal, and time-dependent, each presenting unique modeling challenges [12]; our work focuses on the time-dependent setting. Early efforts like TimeGAN [5] adapted the GAN framework to capture temporal dynamics, while EHR-M-GAN [6] introduced separate encoders for each data modality to better synthesize mixed-type sequences. More recently, HALO [7] leveraged a hierarchical autoregressive language model to generate medical codes and visit sequences over long horizons, and HealthGen [8] adopted a VAE to produce continuous ICU trajectories. Diffusion-based approaches have also emerged: Diffusion-TS [9] uses a transformer-based latent diffusion to model complex time series, and TimeDiff [10] combines multinomial and Gaussian diffusion processes specifically for EHR data. Finally, FLEXGEN-EHR [11] and TimeAutoDiff [4] employs latent diffusion to synthesize heterogeneous longitudinal records end-to-end. It is noted that there is a noticeable lack of emphasis on conditional generation and incorporating patient and demographic context in the generation process.

3 Generative Models

3.1 Base Architecture

The base architecture used in this work is TimeAutoDiff [4], which consists of two main components:

Phase 1: β -VAE Training The autoencoder (consisting of GRU layers) maps mixed-type time-series features to a latent space z using the loss: $\mathcal{L}_{Auto} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KLD}$ where $\mathcal{L}_{recon}$ combines MSE loss for continuous features and cross-entropy for discrete features, and $\mathcal{L}_{KLD} = \mathrm{KL}(q(z|x)|p(z))$ enforces a Gaussian prior.

Phase 2: Conditional Diffusion Training A bidirectional RNN diffusion model operates on the latent space by corrupting and denoising z using the loss function $\mathcal{L}_{Diffusion} = \mathbb{E}_{t,\epsilon}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2]$, where c represents demographic conditioning information and classifier-free guidance balances condition adherence with diversity.

3.2 Enhanced TimeAutoDiff: Our Contributions

We introduce **Enhanced TimeAutoDiff**, an improved variant of the TimeAutoDiff generator that incorporates additional regularization to encourage a smoother and more structured latent space. Specifically, we augment each base loss with two auxiliary terms:

Maximum Mean Discrepancy (MMD) Loss:

$$\ell_{\mathrm{MMD}} = \frac{1}{B^2}\sum_{i,j} k(x_i, x_j) + \frac{1}{B^2}\sum_{i,j} k(y_i, y_j) - \frac{2}{B^2}\sum_{i,j} k(x_i, y_j),$$

where $k(\cdot,\cdot)$ is a radial basis function kernel. This term enforces distributional alignment between the latent representations of real and synthetic samples.

Consistency Regularization:

$$\ell_{\mathrm{consistency}} = \mathrm{MSE}\big(f(x), f(x+\delta)\big), \quad \delta \sim \mathcal{N}(0, \sigma^2 I),$$

This regularizer promotes local smoothness by constraining the model's output to remain consistent under small Gaussian perturbations of the input, thereby improving the robustness of generated samples.

The overall training objectives become:

$$\mathcal{L}^{Enhanced}_{\mathrm{Auto}} = \mathcal{L}_{\mathrm{Auto}} + \lambda^{MMD}_{Auto}\,\ell_{\mathrm{MMD}} + \lambda^{consistency}_{Auto}\,\ell_{\mathrm{consistency}},$$

$$\mathcal{L}^{Enhanced}_{\mathrm{Diffusion}} = \mathcal{L}_{\mathrm{Diffusion}} + \lambda^{MMD}_{diff}\,\ell_{\mathrm{MMD}} + \lambda^{consistency}_{diff}\,\ell_{\mathrm{consistency}}.$$

The model trained under these enhanced objectives constitutes our **Enhanced TimeAutoDiff** framework.

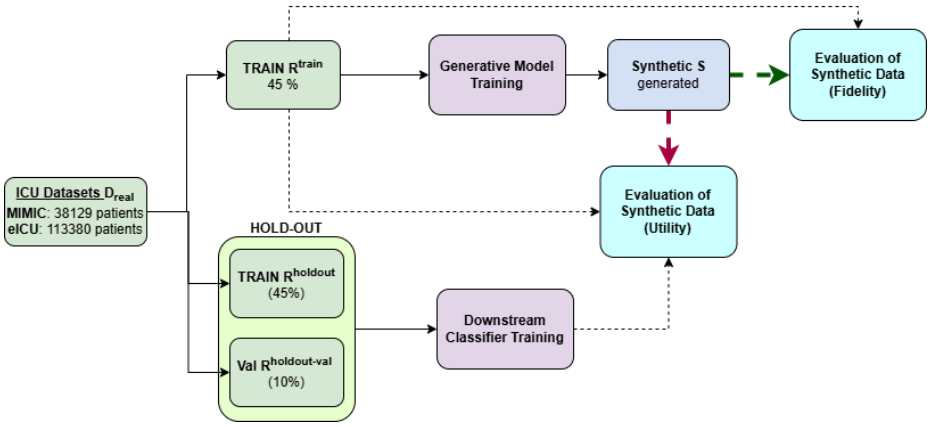

Fig. 1. Workflow for Generative Model Training and Evaluation of Synthetic ICU Data. The real ICU datasets (MIMIC and eICU) are split into training (R^{train}, 45%), holdout ($R^{holdout}$, 45%), and validation ($R^{holdout-val}$, 10%) subsets. A generative model is trained on R^{train} to produce synthetic data S. The synthetic data are assessed through two complementary evaluations: fidelity (green- intrinsic comparison with real data) and utility (purple- extrinsic evaluation via downstream classifier performance). Solid lines indicate training flows, while dashed lines denote evaluation processes.

3.3 Baseline Models for Comparison

In addition to the TimeAutoDiff baseline, we compare to two models: TimeDiff and HealthGen. **TimeDiff** [10] is a diffusion probabilistic model that also uses BiRNN architecture for realistic privacy-preserving EHR time series generation. To simultaneously generate both continuous and discrete-valued time series, TimeDiff introduces a mixed diffusion approach that combines multinomial and Gaussian diffusion for EHR time series generation. The **HealthGen** [8] model consists of a dynamical VAE-based architecture that allows for the generation of feature time series with informative missing values, conditioned on high level static variables and binary labels.

4 Methodology

4.1 Data

Let

$$D_{\text{Real}} = \{ (x_i, y_i, c_i) \}_{i=1}^{N}$$

be the real ICU dataset of size N, where x_i denotes the time-series data, y_i the outcome , and c_i the demographic attributes. As seen in fig. 1, we split D_{Real} into disjoint subsets: a training set R^{train} (45 % of $|D_{real}|$), a holdout set $R^{holdout}$ (45 %), and holdout validation set $R^{holdout-val}$ (10 %). The generative model G is trained on R^{train}, producing a synthetic dataset S of equal size N_{train} (possibly

conditioned on y and c). Since R^{train} serves both as the training source and reference for synthetic data evaluation, we omit the superscript hereafter and denote it simply as R.

The synthetic data S are evaluated through two complementary procedures: (i) *fidelity*, assessing how closely S matches the statistical properties of the real data R; and (ii) utility, evaluating how well S preserves downstream predictive performance when used to train or evaluate classifiers. The holdout subsets $R^{holdout}$ and $R^{holdout-val}$ are used respectively for training and validating the downstream classifiers. This splitting strategy ensures sufficient data for both training a stable generative model and downstream model while maintaining strict separation between training and testing data. All splits are stratified by outcome and demographic variables to preserve population balance.

We use the publicly available MIMIC-III [1] and eICU [2] ICU EHRs, which consist of deidentified ICU data linked by patient ID and timestamp. Following established cohort-selection criteria [3], we extract each patient's first 24 hours of hourly measurements, along with static demographics (age, sex, ethnicity). We frame two popular downstream tasks on this 24-hour window: ICU mortality prediction [3] and binary Length of Stays (LOS) ($>$ 3 days), and consider 8 hourly features for the prediction: four vital signs (heart rate, respiratory rate, oxygen saturation, and mean blood pressure), plus four binary indicators for whether each value is missing. Missingness itself is retained via a mask, and missing values are forward-filled. The missingness masks are used as input for the models. After preprocessing, eICU yields 113 380 patients (mortality = 5.51 %, LOS $>$ 3 d = 42.3 %), and MIMIC yields 38 129 patients (mortality = 8.13 %, LOS $>$ 3 d = 48.7 %).

4.2 Synthetic Data Generation

We adopt a standardized generation protocol across four architectures: Enhanced TimeAutoDiff, baseline TimeAutoDiff, HealthGen, and TimeDiff to ensure fair comparison. All models are trained on the same 45% training split (R_{train}) with identical demographic conditioning on age group, gender, ethnicity, and outcome. Each generator produces a synthetic dataset of equal size ($|S| = |R_{\text{train}}|$), preserving the conditioning distribution of the real data.

Since TimeDiff is not inherently conditional, we adapt it by including demographic and outcome variables as additional input features during training. Conditional samples are then obtained via rejection sampling, retaining only those that match the target demographic and outcome attributes. Although less efficient than native conditional generation, this ensures that TimeDiff yields synthetic data with population characteristics consistent with the conditional models.

4.3 Downstream Prediction Models

We use a one-layer Gated Recurrent Unit (GRU) network followed by a fully connected sigmoid output layer. Models are trained with binary cross-entropy

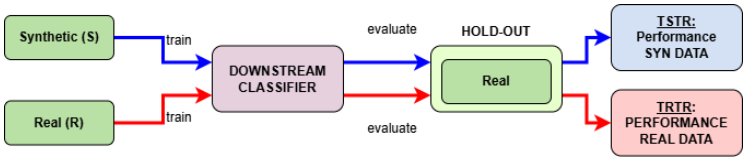

Fig. 2. Training Utility workflow.

loss using the Adam optimizer (learning rate $= 5 \times 10^{-4}$, batch size $= 64$) for up to 50 epochs, selecting the best checkpoint based on validation performance. Architectures and hyperparameters are fixed across all experiments for consistency and follow standard practices from the literature [8,3]. Ablation studies confirm that downstream model parameters have minimal influence on the relative performance of synthetic data generators, indicating that our findings reflect differences in synthetic data quality rather than downstream model optimization.

We implement a one-layer Gated Recurrent Unit (GRU) network followed by a fully connected layer with sigmoid output. Training is done using binary cross-entropy loss, an Adam optimizer (lr $= 5 \times 10^{-4}$), a batch size $= 64$. We fix these architectures across all experiments for consistency. In all evaluations, we train each downstream model for up to 50 epochs, selecting the best checkpoint via validation performance. We used standard hyperparameters and model architectures from the literature [8,3]. Our ablation studies demonstrate that downstream model parameters have minimal effect on the relative performance comparisons between synthetic data generators, ensuring that our conclusions are robust to these choices. The focus of our evaluation is on comparing synthetic data quality rather than optimizing downstream model performance.

4.4 Evaluation Metrics

Our goal is twofold: (1) to improve global evaluation utility (TRTS) so that real-trained models evaluated on synthetic data match performance on true held-out real data, and (2) to enable accurate subgroup-level evaluation by generating large synthetic cohorts conditioned on demographic labels.

Utility for Downstream Training Using identical downstream architectures and hyper-parameters, we train two downstream models, one on real data (red path) and another on synthetic data (blue path). Then, we evaluate both models on hold-out real data $R^{holdout}$ as seen in fig. 2. $\text{TRTR}_{\text{train}}$ and $\text{TSTR}_{\text{train}}$ refer to "Train on Real, Test on Real" and "Train on Synthetic, Test on Real," respectively, and are used to evaluate the training utility of synthetic data. This evaluation is repeated 5 times, each time on a differently trained downstream model.

Utility for Downstream Evaluation We train a downstream predictive model (mortality classifier or LOS classifier) on real holdout dataset $M_{\text{real}} = \text{train}\left(R^{holdout}\right)$.

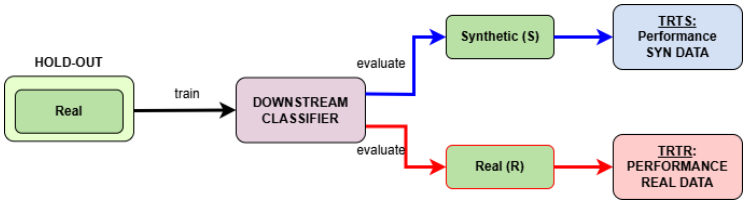

Fig. 3. Evaluation Utility workflow.

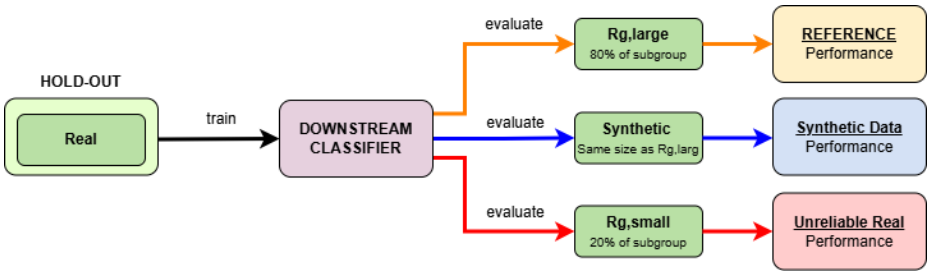

Fig. 4. Subgroup Utility workflow.

Then, we evaluate on both synthetic (blue) and real (red) datasets as seen in fig. 3. $\text{TRTR}_{\text{evaluate}}$ and $\text{TRTS}_{\text{evaluate}}$ refer to "Train on Real, Test on Real" and "Train on Real, Test on Synthetic," respectively, and are used to evaluate the evaluation utility of synthetic data. This evaluation is repeated 5 times, each time on a differently trained downstream model.

Utility for Subgroup Level Evaluation We define three categorical demographic dimensions: age bracket (< 30, 31-50, 51-70, > 70), sex (M and F), and ethnicity (White, Black, Asian, Other). This yields $4 \times 2 \times 4 = 32$ intersectional subgroups of age $\times$ sex $\times$ ethnicity.

For each demographic subgroup g, let $R_g \subset R_{\text{train}}$ be real samples with subgroup label g. To establish reliable reference evaluations, we allocate 80% of each demographic bucket as the "ground truth" set $R_{g,\text{large}}$. The downstream model is evaluated on this set to obtain stable performance estimates. The remaining 20% ($R_{g,\text{small}}$) simulates the small test samples typically available for minority subgroups in real datasets. The 80/20 split is repeated 5 times with different random seeds to ensure statistical robustness.

We generate synthetic cohort S_g with $|S_g| = |R_{g,\text{large}}|$ and compute:

$$\epsilon_g^{\text{naive}} = |\text{AUROC}(M_{\text{real}}, R_{g,\text{large}}) - \text{AUROC}(M_{\text{real}}, R_{g,\text{small}})| \quad (1)$$

$$\epsilon_g^{\text{synth}} = |\text{AUROC}(M_{\text{real}}, R_{g,\text{large}}) - \text{AUROC}(M_{\text{real}}, S_g)| \quad (2)$$

where $\epsilon_g^{\text{naive}}$ represents the error when using small real test sets, and $\epsilon_g^{\text{synth}}$ represents the error when using large synthetic cohorts, both compared to the ground truth performance on $R_{g,\text{large}}$.

Discriminative Fidelity Score We split the real and synthetic datasets into train and test subsets R_{train}, R_{test} and S_{train}, S_{test} respectively. We then label every example in R_{train} with 0 (real), and every example in S_{train} with 1 (synthetic). We train a discriminator on this balanced dataset containing R_{train} and S_{train}. Let the discriminator assign a score $p(\text{synthetic})$ to each sample in the combined dataset ($R_{\text{test}} \cup S_{\text{test}}$).

$$\text{DiscAUC}_S \;=\; \text{AUC}\big[y \in \{0,1\},\, p(\text{synthetic})\big].$$

4.5 Experimental Repetition

We generate 5 independent synthetic datasets from each generator and train 5 downstream models per dataset (real or synthetic), yielding 25 evaluation runs per experimental condition. Moreover, the process of splitting the subgroups into small and large groups is repeated 5 times. This accounts for variability in both synthetic data generation, downstream model training and evaluation, and data splits, enabling robust statistical analysis with reliable confidence intervals and significance testing, while balancing between statistical rigor and computational feasibility. All reported metrics include 95% CIs based on the independent runs.

4.6 Alignment weights search

We explored nine configurations that systematically vary the four alignment weights $\lambda_{AE\text{-}MMD}$,$\lambda_{AE\text{-}consist}$,$\lambda_{diff\text{-}MMD}$,$\lambda_{diff\text{-}consist}$ across three regimes (zero, light = 0.1, moderate = 0.5). In practice, each experiment requires full diffusion training (on eICU or MIMIC), synthetic sample generation, and downstream training/evaluation. Therefore, this choice of weights and number of configurations ensures sufficient coverage of "no regularization," "light touch," and "moderate emphasis" within a timely possible completion. Moreover, this choice helps isolate individual effects and probe pairwise combinations, avoiding a combinatorial explosion of configurations

5 Experimental Results

We present our results in three parts. First, we examine the discriminative fidelity score on eICU and MIMIC for both mortality and LOS tasks, then we present global evaluation utility (TSTR vs TRTS). Finally, we examine subgroup-level evaluation, comparing small real test sets to pooled real "ground truth" and to large synthetic cohorts conditioned on subgroup labels.

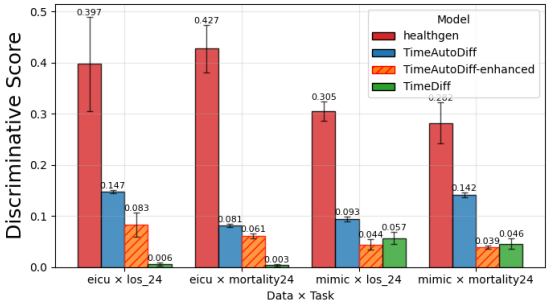

Fig. 5. Model Comparison - Discriminative Fidelity This figure displays bar charts comparing the Discriminative Score of HealthGen, TimeAutoDiff, Enhanced TimeAutoDiff and TimeDiff across different tasks and datasets. The lower the scores, the better.

5.1 Discriminative Fidelity Comparison

Fig. 5 presents discriminative AUCs for four synthetic data generators: HealthGen (conditional VAE), TimeAutoDiff (autoencoder + diffusion), Enhanced TimeAutoDiff (TimeAutoDiff with extra penalty), and TimeDiff (pure diffusion) on four Data × Task combinations (e.g. eICU LOS_{24}, MIMIC LOS_{24}, eICU $Mortality_{24}$, MIMIC $Mortality_{24}$).

HealthGen's discriminative scores range from 0.282 to 0.427, meaning a simple classifier can easily distinguish its outputs from real data. TimeAutoDiff improves on this with discriminative AUCs of 0.081–0.147, while TimeDiff performs best, with AUCs as low as 0.003–0.057. Across every setting, pure diffusion (TimeDiff) is hardest to distinguish from real, followed closely by the Enhanced TimeAutoDiff. The original TimeAutoDiff is an order of magnitude much better than HealthGen, but the added alignment losses yield a substantial drop (25-70%) in the discriminative score compared to the baseline.

5.2 Global Utility Comparison

We then evaluated the global downstream utility of the four synthetic data generators. As shown in the left-hand panel of fig. 6, HealthGen exhibits large Δ_{TSTR} gaps ($\approx$ 0.06-0.10), indicating that downstream models trained on HealthGen synthetic data perform substantially worse on real test data than fully real trained baselines. Both TimeAutoDiff and the enhanced variant exhibit similar performance and reduce Δ_{TSTR} to $\approx$ 0.006–0.011, and TimeDiff further drives it down to $\approx$ 0.003–0.009 across all tasks.

In the right-hand panel, HealthGen again performs worst with Δ_{TRTS} gaps ($\approx$ 0.13–0.19). TimeAutoDiff cuts those evaluation gaps by roughly 75 % (to $\approx$ 0.017–0.039), while TimeDiff drives it down to $\approx$ 0.003–0.023. The Enhanced TimeAutoDiff delivers the smallest evaluation gaps of all four: between 0.003 and 0.014 AUROC depending on the task. For instance, on eICU LOS_{24} it

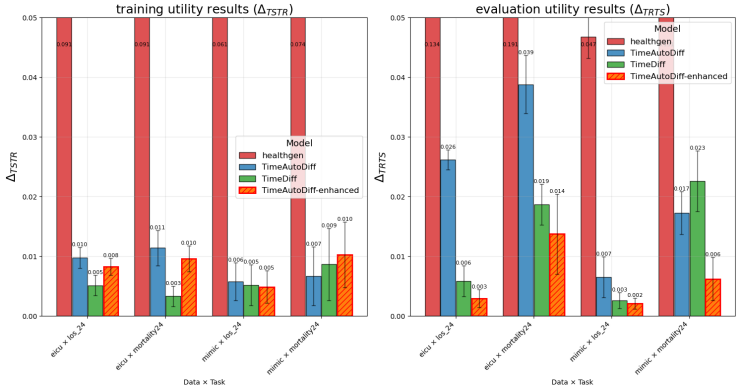

Fig. 6. Model Comparison- Global Downstream Utility This figure displays bar charts comparing the downstream utility for training (Δ_{TSTR}, left panel) and evaluating (Δ_{TRTS}, right panel) of HealthGen,TimeAutoDiff, Enhanced TimeAutoDiff, and TimeDiff across different tasks and datasets. The lower the scores, the better. For clarity, the y-axis is limited to 0.05, truncating larger values (e.g., HealthGen) to highlight differences among the other models.

achieves just 0.003, and on MIMIC $Mortality_{24}$ 0.006. This shows that our regularized variant of TimeAutoDiff yields the most faithful synthetic hold-outs for downstream evaluation, and when compared to the left-hand panel, this was done at a negligible cost to the generator's ability to produce useful training data.

Comparing the TimeAutoDiff and our enhanced variant of it, we see that the "train on real, test on synthetic" gap shrinks dramatically under the best-enhanced setting. For example, eICU LOS_{24} drops from a gap of 0.026 to 0.003, and MIMIC $Mortality_{24}$ falls from 0.017 to 0.006. In each case, the enhanced version (model trained with the optimized weights) reduces Δ_{TRTS} by roughly 70–90%. This means that the synthetic data can give a nearly identical performance as held-out real data for a trained model, all at almost zero cost to the generator's ability to produce synthetic data that trains a high-quality downstream model, as the Δ_{TSTR} remains very small (≈ 0.01) both before and after enhancement. Therefore, by adding MMD and consistency losses to TimeAutoDiff and by tuning their weights per Data x Task, we significantly reduce the evaluation gap Δ_{TRTS} without sacrificing Δ_{TSTR} or fidelity.

Table. 1 and fig. 6 together show that, even though a downstream model trained on synthetic data (TSTR) almost matches the performance of a model trained on real data, the reverse (training on real and testing on synthetic TRTS) shows a larger discrepancy. For example, looking at the data generated by TimeDiff, specifically for eICU $Mortality_{24}$, $\Delta_{TSTR} \approx 0.003$ while Δ_{TRTS} is $\approx$ 0.019, 6 times larger. This problem was substantially mitigated in the Enhanced TimeAutoDiff.

Table 1. Comparison of Downstream Evaluation and Downstream Training. This table illustrates some examples where the difference between training utility (Δ_{TSTR}) and evaluation utility (Δ_{TRTS}) across various tasks and datasets is significantly big.

Data	Task	Model	Δ_{TSTR}	Δ_{TRTS}	Relation
eICU	LOS_{24}	TimeAutoDiff	0.01 ± 0.002	0.026 ± 0.002	$\Delta_{TRTS} \approx 3 \times \Delta_{TSTR}$
eICU	$Mortality_{24}$	TimeAutoDiff	0.011 ± 0.003	0.039 ± 0.005	$\Delta_{TRTS} \approx 3.5 \times \Delta_{TSTR}$
eICU	$Mortality_{24}$	TimeDiff	0.003 ± 0.002	0.019 ± 0.003	$\Delta_{TRTS} \approx 6 \times \Delta_{TSTR}$

5.3 Subgroup-Level Evaluation

Table 2. Percentage of intersectional subgroups in which the synthetic evaluation error is strictly lower than the real-test error, for each generator:TimeAutoDiff (TA), TimeDiff (TD), and Enhanced TimeAutoDiff TA^+ across the different configurations.

Data	Task	% of Subgroups Where Synthetic < Test Error		
		TA	TD	TA^+
eICU	$Mortality_{24}$	48%	60%	76%
eICU	LOS_{24}	44%	52%	72%
mimic	$Mortality_{24}$	68%	68%	84%
mimic	LOS_{24}	72%	56%	76%

For each Data x Task, we compute the absolute AUROC distance between (a) a small real-test subgroup and the large "ground-truth" subgroup (red bars labeled "Test") versus (b) a synthetic, conditional-on-subgroup cohort and that same ground-truth subgroup (orange,green, and blue bars labeled "Synthetic"). Then, we average those errors across all subgroups to produce the red ("Test"), blue ("Synthetic TimeAutoDiff"), green ("Synthetic TimeDiff"), and dashed orange ("Enhanced TimeAutoDiff") bars in Fig. 7. In Table 2, we counted, across all subgroups, how often the Synthetic error was significantly smaller than the Test error.

Fig. 7 shows that, in every case, the synthetic cohorts cut the error roughly in half compared to the small real-test subsets (red bar), and our enhanced generator (Enhanced TimeAutoDiff) yields the lowest errors of all. For example, on eICU LOS_{24} the mean error drops from 0.044 (Test) to 0.039 (TimeAutoDiff), 0.033 (TimeDiff), and 0.028 (Enhanced TimeAutoDiff); on MIMIC $Mortality_{24}$ it falls from 0.142 to 0.081, 0.088, and 0.081 respectively; On eICU $Mortality_{24}$: $0.075 \rightarrow 0.061 \rightarrow 0.056 \rightarrow 0.050$, and on MIMIC LOS_{24}: $0.067 \rightarrow 0.043 \rightarrow 0.049 \rightarrow 0.042$.

Table 2 shows the fraction of subgroups in which each synthetic cohort outperforms the real-test subset (i.e. has a strictly smaller error). Although both TimeAutoDiff and TimeDiff reduce the mean Test error substantially as depicted

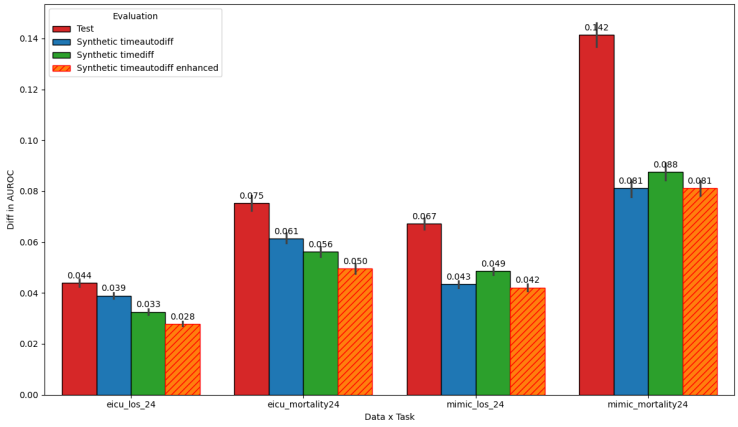

Fig. 7. Subgroup-Level Evaluation Mean absolute AUROC error of small-real vs. large-pool ("Test," red) compared to synthetic cohort vs. large-pool errors using TimeAutoDiff (blue) and TimeDiff(green), and Enhanced TimeAutoDiff (dashed orange) averaged over all subgroups for different data and task configurations. The lower the scores, the better.

in 7, almost halving it in most cases, their success rates remain under 50 % on the eICU tasks (TimeAutoDiff: 48 % for $Mortality_{24}$, 44 % for LOS_{24}; TimeDiff: 60 % for $Mortality_{24}$ and 52 % for LOS_{24}), meaning they only outperform the real-test evaluation in fewer than half of the subgroups. By contrast, our Enhanced TimeAutoDiff model outperforms Test in the majority of subgroups across every setting (72 %–84 % success). Moreover, the Enhanced TimeAutoDiff achieves the lowest average errors in every Data × Task.

Together, these results demonstrate that generating large conditional cohorts, even with different underlying generator architectures, yields substantially more accurate subgroup performance estimates than evaluating directly on small held-out subsets, and that our Enhanced TimeAutoDiff model turns synthetic cohorts into more reliable evaluators that deliver more accurate subgroup performance estimates in the vast majority of cases.

6 Discussion

This study shifts focus from synthetic data as a training resource to synthetic data as an evaluation tool, investigating how well synthetic ICU time-series can serve as reliable proxies for real data in model assessment. We systematically compared four generators across global and subgroup-level evaluation scenarios, with particular emphasis on the previously understudied evaluation utility (TRTS) gap.

6.1 Key Findings and Implications

Our experiments reveal three critical insights about synthetic medical data evaluation. First, **diffusion-based models substantially outperform VAE-based approaches** not only in terms of evaluation, but also in terms of training utility and discriminative fidelity. While HealthGen exhibits large evaluation gaps (Δ_{TRTS} = 0.13–0.19 AUROC), diffusion models reduce these gaps to 0.003–0.04 AUROC, with Enhanced TimeAutoDiff achieving the smallest gaps (0.003–0.014 AUROC). In terms of discriminative fidelity, pure diffusion (TimeDiff) achieved the lowest discriminative scores (AUC $\approx$ 0.003–0.057), with Enhanced TimeAutoDiff close behind (AUC $\approx$ 0.039–0.083), followed by TimeAutoDiff (0.081–0.147), while HealthGen remained easiest to distinguish from real data (AUC $\approx$ 0.28–0.43).

Second, **evaluation utility requires different optimization than training utility**. Although diffusion models achieve strong TSTR performance ($\Delta_{TSTR} \approx$ 0.01), they initially showed substantially larger TRTS gaps, revealing fundamental distributional mismatches that standard training objectives fail to address.

Third, **synthetic cohorts enable reliable subgroup evaluation where real data are insufficient**. Enhanced TimeAutoDiff outperforms small real test sets in 72–84% of demographic subgroups, reducing evaluation error by up to 50% compared to standard approaches that rely on limited real samples for minority populations.

6.2 Technical Innovation: Distribution Alignment

Although diffusion-based generators yield strong TSTR performance, they still exhibit substantially larger than TRTS gaps, revealing residual mismatches when evaluating real-trained models on synthetic hold-outs. This pronounced TSTR-TRTS gap reflects an inherent limitation of generative models: they tend to produce smoothed versions of training data, averaging out rare modes while preserving dominant patterns. During training (TSTR), downstream models can adapt to these smoothed distributions. However, during evaluation (TRTS), fixed models encounter mismatches between their learned decision boundaries and the synthetic distribution, leading to evaluation errors.

Enhanced TimeAutoDiff addresses this through explicit distribution alignment. We augment both autoencoder and diffusion losses with MMD penalties that enforce distributional matching and consistency losses that ensure smooth model behavior. This approach reduces Δ_{TRTS} by 70–90% (Across both datasets and both tasks, Δ_{TRTS} fell to 0.003–0.014 AUROC, outperforming all baselines) while preserving training utility ($\approx$ 0.01 AUROC), demonstrating that evaluation-specific optimization can substantially improve synthetic data reliability without sacrificing sample quality.

Moreover, the enhanced approach demonstrated improved subgroup evaluation. Our enhanced model outperforms the real test evaluation in 72–84 % of all subgroups, and yields the lowest mean errors in every Data$\times$Task. This confirms that generating large, conditionally matched synthetic cohorts using our

enhanced model is not only beneficial on average, but reliably so for intersectional groups where real data are scarce. In contrast, baseline models (TimeAutoDiff and TimeDiff) only succeeded in fewer than 50 % of eICU subgroups, meaning that in the majority of subgroups, real test subsets still outperform their synthetic counterparts.

6.3 Clinical Impact and Fairness Implications

Our subgroup evaluation results have direct implications for healthcare AI fairness. Traditional evaluation approaches often rely on small subgroup samples that yield unreliable performance estimates, masking true model behavior across demographic intersections. By generating large, balanced synthetic cohorts, we enable precise bias detection and fair performance assessment, facilitating more equitable AI systems while maintaining patient privacy.

6.4 Limitations and Future Directions

Several limitations warrant acknowledgment. Our hyperparameter optimization remains dataset-specific, requiring manual tuning for optimal performance. **Meta-learning approaches** could automate this process, improving robustness across different medical contexts.**Privacy guarantees** represent another critical gap. While MMD and consistency losses implicitly discourage memorization, formal differential privacy analysis is essential for deployment in sensitive healthcare settings. Incorporating DP-SGD training and conducting rigorous privacy audits should be immediate priorities. Finally, our evaluation focuses on two prediction tasks (mortality and LOS) within ICU settings. **Broader medical applications** including continuous outcomes, multimodal data, and other clinical domains require investigation to establish the generalizability of our approach.

7 Conclusion

We have introduced Enhanced TimeAutoDiff, a latent diffusion framework augmented with distribution-alignment losses, and demonstrated its superiority for both global and subgroup-level model evaluation on ICU time-series data. Our extensive benchmarks on MIMIC-III and eICU show that Enhanced TimeAutoDiff achieves the smallest evaluation gaps ($\Delta_{TRTS} \leq 0.014$ AUROC) and preserves training utility ($\Delta_{TSTR} \approx 0.01$). By generating large conditional cohorts, it also halved subgroup-level estimation errors and outperformed small real test sets in the majority of intersectional subgroups. These results establish a practical blueprint for using synthetic data to perform robust, privacy-preserving evaluations of predictive models across diverse patient populations. All code, checkpoints, and evaluation pipelines are publicly available to facilitate adoption and further research.

Acknowledgments. This work was funded by the European Union under the Horizon Europe grant 101095435. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A. (2021). Mark. R. MIMIC-IV (version 1.0). PhysioNet.
2. Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific data, 5(1), 1-13.
3. Van De Water, R., Schmidt, H., Elbers, P., Thoral, P., Arnrich, B., & Rockenschaub, P. (2023). Yet another ICU benchmark: A flexible multi-center framework for clinical ML. arXiv preprint arXiv:2306.05109.
4. Suh, N., Yang, Y., Hsieh, D.-Y., Luan, Q., Xu, S., Zhu, S., & Cheng, G. (2024). TimeAutoDiff: Combining autoencoder and diffusion model for time series tabular data synthesizing [Preprint]. arXiv. https://arxiv.org/abs/2406.16028
5. Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series Generative Adversarial Networks. In Advances in Neural Information Processing Systems, 32, 5508–5518.
6. Zhu, T., Li, J., Cairns, B. J., & Li, J. (2022). Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications. arXiv:2112.12047 [cs.LG].
7. Theodorou, B., Xiao, C., & Sun, J. (2023). Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. Nature Communications, 14, 5305. https://doi.org/10.1038/s41467-023-41093-0
8. Bing, S., Dittadi, A., Bauer, S., & Schwab, P. (2022). Conditional generation of medical time series for extrapolation to underrepresented populations. PLOS Digital Health, 1(7), e0000074. https://doi.org/10.1371/journal.pdig.0000074
9. Yuan, X., & Qiao, Y. (2024). Diffusion-TS: Interpretable Diffusion for General Time Series Generation. arXiv:2403.01742 [cs.LG].
10. Tian, M., Chen, B., Guo, A., Jiang, S., & Zhang, A. R. (2024). Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. Journal of the American Medical Informatics Association, ocae229. https://doi.org/10.1093/jamia/ocae229
11. He, H., Hao, W., Xi, Y., Chen, Y., Malin, B., & Ho, J. (2024). A Flexible Generative Model for Heterogeneous Tabular EHR with Missing Modality. ICLR 2024 poster.
12. Ibrahim, M., Al Khalil, Y., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., ... & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. Computers in biology and medicine, 189, 109834.
13. De Hond, A. A. H., Leeuwenberg, A. M., Hooft, L., Kant, I. M. J., Nijman, S. W. J., Van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., Van Smeden, M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., & Moons, K. G. M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. Npj Digital Medicine, 5(1), 2. https://doi.org/10.1038/s41746-021-00549-7

14. Herlihy, C., Truong, K., Chouldechova, A., & Dudík, M. (2024, June). A structured regression approach for evaluating model performance across intersectional subgroups. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 313-325).
15. van Breugel, B., Seedat, N., Imrie, F., & van der Schaar, M. SYNG4ME: Model Evaluation using Synthetic Test Data.