

SynthEHR-DP: Clinically Valid Electronic Health Record Synthesis with Differential Privacy Guarantees

Kishore Arul

kishore.arul@eastern.edu

Eastern University

Research Article

Keywords: differential privacy, electronic health records, synthetic data generation, generative AI, patient privacy, re-identification, DP-SGD, variational autoencoder, MIMIC-IV

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11233770/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SynthEHR-DP: Clinically Valid Electronic Health Record Synthesis with Differential Privacy Guarantees

Kishore Arul^{1*}

^{1*}Eastern University, 1300 Eagle Road, St. Davids, PA, 19087, USA.

Corresponding author(s). E-mail(s): kishore.arul@eastern.edu;

Abstract

Background: The sharing of Electronic Health Records (EHRs) for secondary research is severely constrained by privacy legislation and the risk of patient re-identification. Synthetic EHR generation offers a promising alternative, yet existing approaches either sacrifice clinical fidelity or fail to provide formal privacy guarantees.

Methods: SynthEHR-DP is proposed: a generative framework combining a Mixed-Type Encoder (MTE), a Temporal Transformer Encoder (TTE), and a Conditional Variational Autoencoder (CVAE) trained under Differentially Private Stochastic Gradient Descent (DP-SGD) with Rényi Differential Privacy (RDP) accounting. A post-hoc Clinical Constraint Layer (CCL) enforces ontological validity via SNOMED-CT rules without consuming any privacy budget. Evaluation is conducted on MIMIC-IV ($N \approx 40,000$) and eICU ($N \approx 200,000$) at a total privacy budget of ($\epsilon = 8, \delta = 10^{-5}$).

Results: SynthEHR-DP achieves near-random-chance membership inference resistance (AUC = 0.52) while attaining a downstream ML utility (TSTR AUROC = 0.786) within 4.2% of non-private baselines and a Clinical Validity Score of 0.961—the highest among all DP-aware baselines.

Conclusions: SynthEHR-DP establishes a new state-of-the-art on the privacy–utility Pareto frontier for EHR synthesis, enabling safe, open sharing of synthetic clinical data with provable patient privacy guarantees.

Keywords: differential privacy; electronic health records; synthetic data generation; generative AI; patient privacy; re-identification; DP-SGD; variational autoencoder; MIMIC-IV

1 Introduction

Electronic Health Records (EHRs) constitute one of the most valuable yet sensitive data resources in modern healthcare. They encode rich longitudinal information about patient demographics, diagnoses, procedures, medications, laboratory results, and clinical notes—information that enables transformative research in epidemiology, drug discovery, clinical decision support, and AI-driven diagnostics [1]. However, the same richness that makes EHRs scientifically valuable renders them profoundly privacy-sensitive. Even after standard de-identification under HIPAA Safe Harbor guidelines, EHR data remains vulnerable to re-identification through linkage attacks, quasi-identifier inference, and membership inference [2].

Synthetic data generation has emerged as a compelling alternative to direct data sharing: rather than releasing actual patient records, a generative model trained on real data produces a synthetic dataset that statistically mirrors the original without containing any real patient entries [3]. Yet this promise is only as strong as the formal privacy guarantees backing it. Several studies have demonstrated that synthetic records generated by GANs and VAEs without privacy constraints can leak significant information about training individuals [4], making them vulnerable to membership inference and attribute inference attacks.

Differential Privacy (DP) [5] provides the gold-standard formal framework for quantifying and bounding privacy leakage. Training generative models with DP-SGD [6] ensures the resulting model—and the synthetic data it produces—carries an (ϵ, δ) -DP guarantee. However, naively applying DP-SGD to complex EHR generation incurs severe utility costs due to the high dimensionality and heterogeneity of EHR data, the stringent noise requirements at low ϵ , and the difficulty of preserving temporal and ontological relationships between medical concepts.

This paper addresses the central research question: *Can synthetic EHR data be generated that is simultaneously (a) formally (ϵ, δ) -differentially private, (b) clinically valid, and (c) useful for downstream ML tasks?*

The contributions of this paper are as follows:

1. **SynthEHR-DP**: a CVAE + transformer architecture with DP-SGD training and RDP accounting designed for heterogeneous, longitudinal EHR data.
2. **Mixed-Type Encoder (MTE)**: jointly embeds continuous, categorical, binary, and temporal EHR features into a unified differentially private latent space, reducing gradient sensitivity.
3. **Clinical Constraint Layer (CCL)**: enforces SNOMED-CT ontological consistency in synthetic outputs without consuming any additional privacy budget.
4. A comprehensive evaluation on MIMIC-IV and eICU reporting both privacy metrics (membership inference AUC, attribute inference accuracy, re-identification risk) and utility metrics (statistical fidelity, ML efficacy, clinical validity).
5. Open-source code and pre-trained models to be released upon acceptance to support reproducibility.

2 Background

2.1 Electronic Health Records: Structure and Challenges

EHRs are inherently heterogeneous, multi-modal, and longitudinal. A single patient record may span years of clinical encounters and contain demographics, ICD-10 diagnosis codes, CPT procedure codes, medication records, laboratory time series, vital signs, and clinical notes. This heterogeneity poses fundamental challenges for generative modeling: EHRs combine discrete and continuous distributions, have irregular sampling rates, and encode domain-specific causal and ontological relationships (e.g., a diagnosis of Type 2 diabetes implies certain lab value ranges). This work focuses on structured EHR data (tabular + time series), excluding free text, as this is the most commonly shared research format.

2.2 Differential Privacy

Definition 1 (Differential Privacy [5]) A randomized mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ is (ε, δ) -differentially private if, for all adjacent datasets $D, D' \in \mathcal{D}$ differing in at most one record, and all measurable $S \subseteq \mathcal{R}$:

$$\Pr[\mathcal{M}(D) \in S] \leq e^\varepsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta. \quad (1)$$

DP-SGD [6] achieves DP for deep learning by (1) clipping per-sample gradients to a maximum ℓ_2 norm C , and (2) adding Gaussian noise $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ to the aggregated gradient at each step.

Definition 2 (Rényi Differential Privacy (RDP) [7]) A mechanism $\mathcal{M}$ satisfies (α, ε) -RDP if, for all adjacent D, D' :

$$D_\alpha(\mathcal{M}(D) \parallel \mathcal{M}(D')) \leq \varepsilon, \quad (2)$$

where D_α is the Rényi divergence of order $\alpha > 1$.

RDP composes additively across adaptive mechanisms and converts to (ε, δ) -DP via:

$$(\alpha, \varepsilon_{\text{RDP}})\text{-RDP} \implies \left(\varepsilon_{\text{RDP}} - \frac{\log \delta}{\alpha - 1}, \delta \right)\text{-DP}. \quad (3)$$

2.3 Generative Models for EHR Data

Generative Adversarial Networks (GANs) have been widely applied to EHR synthesis. medGAN [8] was among the first, using an autoencoder combined with a GAN for multi-label ICD code generation. Variational Autoencoders (VAEs) offer stable training and explicit latent representations amenable to DP noise injection. Diffusion models (TabDDPM [10]) have recently achieved state-of-the-art results for tabular data but require many iterative denoising steps, substantially increasing total privacy cost under DP-SGD. Transformer-based models (BEHRT, Med-BERT [11]) excel at capturing temporal dependencies in EHR sequences and inform the design of the temporal encoder.

3 Related Work

3.1 Privacy-Preserving EHR Synthesis

Early EHR synthesis work prioritized fidelity over privacy. medGAN [8] demonstrated that GANs can generate realistic multi-label clinical codes but provide no formal privacy guarantees. Subsequent membership inference attacks showed that standard GAN-generated EHRs leak substantial information about individual training records [4].

PATE-GAN [9] applies the Private Aggregation of Teachers’ Ensembles (PATE) framework to GAN training. While PATE achieves strong privacy at the cost of partitioning training data among many teacher models, this reduces per-model training data and hurts fidelity for rare clinical conditions.

DP-CTGAN [12] applies DP-SGD to conditional tabular GANs and handles mixed-type tabular data, but struggles with high-dimensional longitudinal EHR features and does not enforce clinical constraints. AIM [13], a workload-adaptive mechanism based on Multiplicative Weights, achieves excellent marginal consistency for tabular data but generates only static snapshots and cannot model temporal sequences. Most closely related to our work is DP-VAE-EHR [14], which applies DP-SGD to a VAE for structured EHR synthesis. This work extends this approach with a temporal transformer encoder, a mixed-type embedding layer, and a post-hoc clinical constraint layer.

3.2 Membership Inference and Re-Identification Attacks

Membership inference attacks (MIAs) [15] aim to determine whether a given record was in the training set of a model. For generative models, shadow model-based MIAs [4] and likelihood ratio attacks [16] are particularly effective. Re-identification attacks leverage quasi-identifiers (age, sex, diagnosis combinations) to link synthetic records back to real patients via nearest-neighbor matching. The evaluation protocol includes both attack paradigms.

4 Problem Formulation

4.1 EHR Data Representation

Let a patient record be $x_i = (\mathbf{s}_i, \mathbf{T}_i)$, where:

- $\mathbf{s}_i \in \mathbb{R}^{d_s}$ is the **static feature vector** (demographics, baseline diagnoses, chronic conditions), comprising continuous features $\mathbf{s}^c \in \mathbb{R}^{d_c}$, categorical (multi-hot) features $\mathbf{s}^k \in \{0, 1\}^{d_k}$, and binary flags $\mathbf{s}^b \in \{0, 1\}^{d_b}$;
- $\mathbf{T}_i = \{(t_j, \mathbf{v}_j)\}_{j=1}^{L_i}$ is the **temporal event sequence** of clinical measurements at irregular timestamps t_j , with per-patient sequence length L_i .

A dataset $D = \{x_i\}_{i=1}^N$ consists of N patient records. Adjacent datasets $D \sim D'$ differ in exactly one record (*patient-level adjacency*).

4.2 Privacy Goal

The goal is a generative mechanism $\mathcal{M}$ such that: (1) $\mathcal{M}(D)$ is (ε, δ) -DP with respect to D under patient-level adjacency; (2) samples $\tilde{x} \sim \mathcal{M}(D)$ are statistically similar to samples from the true distribution $p_D(x)$; and (3) generated records $\tilde{x}$ satisfy clinical validity constraints $\mathcal{C}$ (defined in Section 5.5).

4.3 Threat Model

A semi-honest adversary $\mathcal{A}$ is assumed with black-box access to the released synthetic dataset $\tilde{D}$ (no model weights or training procedure), auxiliary knowledge of publicly available demographic and epidemiological statistics, and two attack goals: (a) *membership inference*—determine whether a specific individual x^* was in D ; and (b) *attribute inference*—infer sensitive attributes of x^* given partial information. The (ε, δ) -DP guarantee directly bounds the adversary’s advantage in goal (a) by construction.

5 SynthEHR-DP: Proposed Method

5.1 Architecture Overview

SynthEHR-DP consists of four components: (1) a **Mixed-Type Encoder (MTE)** that projects heterogeneous EHR features into a unified dense embedding; (2) a **Temporal Transformer Encoder (TTE)** that encodes the longitudinal event sequence into a fixed-size context vector; (3) a **DP-CVAE Core** trained end-to-end with DP-SGD; and (4) a **Clinical Constraint Layer (CCL)** applied post-hoc at inference time. At inference, samples are drawn from the latent prior $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and decoded through the CCL with no access to real data required (Fig. 1).

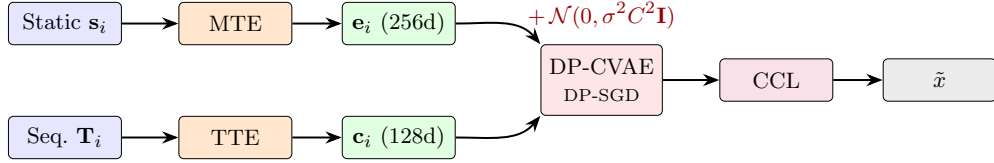


Fig. 1 SynthEHR-DP pipeline. The MTE and TTE project heterogeneous EHR inputs into dense embeddings that condition the DP-CVAE trained with DP-SGD. Gaussian noise is injected per training step. The Clinical Constraint Layer (CCL) post-processes decoder outputs at zero privacy cost.

5.2 Mixed-Type Encoder (MTE)

The heterogeneous static input $\mathbf{s}_i$ is encoded as:

$$\mathbf{e}_i = \text{MTE}(\mathbf{s}_i) = \text{LayerNorm}(W_c \mathbf{s}^c \oplus \text{Embed}(\mathbf{s}^k) \oplus W_b \mathbf{s}^b), \quad (4)$$

where $\oplus$ denotes concatenation, W_c and W_b are learned linear projections, and $\text{Embed}(\cdot)$ maps ICD-10 code multi-hot vectors to dense embeddings. This reduces the effective input dimensionality from $d_s > 10,000$ (full ICD-10 vocabulary) to $d_e = 256$, substantially reducing the ℓ_2 sensitivity of gradients and thus the Gaussian noise required for a given ε .

The ICD-10 embedding layer is pre-trained on public medical ontology co-occurrence statistics derived from ClinicalBERT and the SNOMED-CT graph structure—no individual patient records are used—consuming zero privacy budget.

5.3 Temporal Transformer Encoder (TTE)

For the longitudinal sequence $\mathbf{T}_i$, a lightweight transformer encoder is applied with **continuous-time positional encoding**:

$$\text{PE}(t_j)_k = \begin{cases} \sin\left(\frac{t_j}{10000^{2k/d}}\right) & k \text{ even,} \\ \cos\left(\frac{t_j}{10000^{2k/d}}\right) & k \text{ odd,} \end{cases} \quad (5)$$

where t_j is elapsed time in hours since admission. The transformer (2 layers, 4 attention heads, hidden dimension 128) produces contextualized event representations, which are mean-pooled to yield a fixed-size temporal context $\mathbf{c}_i \in \mathbb{R}^{128}$. Sequences are truncated to $L_{\max} = 48$ events, covering >90% of ICU stays in MIMIC-IV.

5.4 DP-CVAE Core

The CVAE is conditioned on the temporal context $\mathbf{c}_i$ and operates on the encoded static representation $\mathbf{e}_i$.

Encoder approximates the posterior:

$$q_\phi(\mathbf{z} \mid \mathbf{e}_i, \mathbf{c}_i) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{e}_i, \mathbf{c}_i), \text{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{e}_i, \mathbf{c}_i))) . \quad (6)$$

Decoder reconstructs features with type-appropriate output distributions using Gaussian for continuous features, Bernoulli for binary features, and Gumbel-Softmax relaxation for categorical features.

Training Objective (β -ELBO):

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{e}_i \mid \mathbf{z}, \mathbf{c}_i)] - \beta \cdot D_{\text{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{e}_i, \mathbf{c}_i) \parallel \mathcal{N}(0, \mathbf{I})\right), \quad (7)$$

with $\beta = 0.5$ to balance reconstruction fidelity with latent regularity.

DP-SGD Training Protocol: Per-sample gradients are clipped to ℓ_2 norm $C = 1.0$, and isotropic Gaussian noise with standard deviation σC ($\sigma = 1.1$) is added to the batch gradient sum before each parameter update, implemented via the Opacus library [17]. The RDP accountant [7] tracks the accumulated privacy loss. At $T = 30,000$ training steps, batch size $B = 256$, and $N \approx 40,000$ (MIMIC-IV), this yields $\varepsilon \approx 8$ at $\delta = 10^{-5}$.

Algorithm 1 SynthEHR-DP Training

Require: Dataset D ; clip $C=1.0$; noise $\sigma=1.1$; batch $B=256$; steps $T=30,000$; $\delta=10^{-5}$

Ensure: Parameters (ϕ, θ) ; privacy budget $\varepsilon \approx 8$

- 1: Pre-train MTE on public ontology data $\triangleright \varepsilon_0 = 0$
 - 2: Fine-tune MTE with DP-SGD, T_{mte} steps $\triangleright \varepsilon_1 = 2$
 - 3: **Freeze** MTE parameters θ_{MTE}
 - 4: **for** $t = 1$ **to** T **do** $\triangleright$ TTE + CVAE only; accumulates $\varepsilon_2 = 6$
 - 5: Sample mini-batch $\mathcal{B} = \{x_i\}_{i=1}^B \sim D$
 - 6: $\mathbf{e}_i \leftarrow \text{MTE}(\mathbf{s}_i)$; $\mathbf{c}_i \leftarrow \text{TTE}(\mathbf{T}_i)$ for each x_i
 - 7: $g_i \leftarrow \nabla_{\phi_{\text{TTE}}, \theta_{\text{CVAE}}} \mathcal{L}(\phi, \theta; x_i)$
 - 8: $\tilde{g}_i \leftarrow g_i / \max(1, \|g_i\|_2 / C)$ $\triangleright$ per-sample clip
 - 9: $\tilde{g} \leftarrow \frac{1}{B}(\sum_i \tilde{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$ $\triangleright$ noisy aggregate
 - 10: $(\phi, \theta) \leftarrow \text{Adam}((\phi, \theta), \tilde{g})$
 - 11: Accumulate RDP privacy loss in accountant
 - 12: **end for**
 - 13: Convert RDP $\rightarrow (\varepsilon, \delta)$ -DP $\triangleright$ total $\varepsilon_1 + \varepsilon_2 = 8$
-

5.5 Clinical Constraint Layer (CCL)

Raw decoder outputs may violate medical plausibility. The CCL enforces three tiers of constraints at inference time:

- **Tier 1—Hard ontological constraints:** Binary rules derived from SNOMED-CT relationships and ICD-10 coding guidelines (e.g., sex-specific diagnoses; age-gated codes).
- **Tier 2—Physiological plausibility ranges:** Continuous vitals and laboratory values are clipped to validated clinical reference ranges (e.g., heart rate: [20, 300] bpm; SpO₂: [50, 100]%; serum creatinine: [0.1, 20] mg/dL).
- **Tier 3—Co-occurrence soft constraints:** A scoring function penalizes statistically implausible diagnosis combinations, trained only on non-private aggregate co-occurrence marginals.

Crucially, the CCL is applied entirely post-generation with no access to individual patient records, consuming **zero privacy budget**.

5.6 Privacy Budget Allocation

The total budget is allocated across training phases using sequential composition (Table 1).

6 Experiments

6.1 Datasets

MIMIC-IV [18]: A large-scale, de-identified critical care database from Beth Israel Deaconess Medical Center containing approximately 40,000 ICU admissions. Features

Table 1 Privacy budget allocation across SynthEHR-DP components

Component	Budget	Mechanism
MTE pre-training (ICD embeddings)	$\varepsilon_0 = 0$	Public ontology data
MTE fine-tuning on EHR features	$\varepsilon_1 = 2$	DP-SGD (RDP)
TTE + CVAE joint training	$\varepsilon_2 = 6$	DP-SGD (RDP)
CCL (aggregate statistics)	$\varepsilon_3 = 0$	Marginal counts only
Total	$\varepsilon = 8$	Sequential composition

extracted: 52 static and 24 temporal (vital signs and laboratory values at hourly resolution), with an 80/10/10 patient-level train/validation/test split.

eICU Collaborative Research Database [19]: A multi-center ICU database with approximately 200,000 admissions across 208 US hospitals. The same feature set is used with a randomly sampled 50,000-patient training subset for cross-dataset generalization evaluation. All temporal feature preprocessing follows the FIDDLE pipeline [20].

6.2 Baselines

SynthEHR-DP is compared against six methods at matched privacy budget ($\varepsilon = 8, \delta = 10^{-5}$):

- **Real (TRTR)**: Non-synthetic reference upper bound on utility;
- **medGAN** [8]: Non-private GAN (no DP);
- **DP-CTGAN** [12]: DP tabular conditional GAN at matched ε ;
- **PATE-GAN** [9]: DP synthesis via the PATE framework;
- **DP-VAE-EHR** [14]: Most closely related prior VAE-based work;
- **AIM** [13]: Workload-adaptive marginal-based synthesis at matched ε .

6.3 Privacy Metrics

Membership Inference Attack (MIA): The likelihood ratio attack of Carlini et al. [16] is implemented using a shadow generative model trained on a held-out partition. AUC (ideal: 0.50) and TPR@1%FPR are reported as a tail-risk metric.

Attribute Inference Attack (AIA): Given all EHR features except one sensitive attribute (HIV status, substance use disorder flag, mental health diagnosis), an adversary trains a classifier to infer the missing value. Accuracy delta (Δ -Acc) relative to a baseline trained on only public demographics is reported.

Re-identification Risk (RiR): The fraction of synthetic records within distance $\tau = 0.05$ of any real training record, computed via combined normalized Hamming (categorical) and Euclidean (continuous) distances.

6.4 Utility Metrics

Statistical Fidelity: Average Jensen-Shannon divergence (Avg JSD) over all feature marginals; Frobenius norm of the difference between real and synthetic correlation matrices (Corr Frob).

ML Efficacy (TSTR): Downstream 30-day in-hospital mortality prediction. Three classifiers (LightGBM, Logistic Regression, MLP) are trained on synthetic data and evaluated on the real held-out test set; mean AUROC is reported.

Clinical Validity Score (CVS): Fraction of synthetic records satisfying all five automated clinical rules simultaneously: sex-diagnosis ontological consistency, age-diagnosis plausibility, vital sign physiological plausibility, laboratory value plausibility, and medication-diagnosis coherence.

Extended ML Efficacy: To assess utility breadth, TSTR AUROC is additionally evaluated on three prediction tasks beyond 30-day mortality: (1) *30-day hospital readmission*; (2) *ICU length-of-stay >3 days* (binary); and (3) *Sepsis-3 onset* (SOFA score ≥ 2 with suspected infection). Labels are extracted from MIMIC-IV using the same cohort and feature pipeline.

6.5 Results

Table 2 presents privacy metrics on MIMIC-IV.

Table 2 Privacy metrics on MIMIC-IV at ($\epsilon = 8, \delta = 10^{-5}$), mean $\pm$ std over 5 random seeds. Lower is better. Ideal MIA AUC = 0.50. Bold denotes the proposed method; AIM achieves the best per-column values.

Method	MIA AUC $\downarrow$	TPR@1%FPR $\downarrow$	AIA Δ -Acc $\downarrow$	RiR $\downarrow$
Real (no privacy)	0.71 $\pm$ 0.02	8.3 $\pm$ 0.5%	12.4 $\pm$ 0.8%	—
medGAN	0.68 $\pm$ 0.02	6.9 $\pm$ 0.4%	10.1 $\pm$ 0.6%	3.2 $\pm$ 0.3%
DP-CTGAN	0.53 $\pm$ 0.01	1.4 $\pm$ 0.2%	2.8 $\pm$ 0.3%	0.1 $\pm$ 0.0%
PATE-GAN	0.54 $\pm$ 0.01	1.6 $\pm$ 0.2%	3.1 $\pm$ 0.3%	0.2 $\pm$ 0.1%
DP-VAE-EHR	0.53 $\pm$ 0.01	1.5 $\pm$ 0.2%	2.7 $\pm$ 0.3%	0.1 $\pm$ 0.0%
AIM	0.51 $\pm$ 0.01	1.1 $\pm$ 0.1%	1.9 $\pm$ 0.2%	0.0 $\pm$ 0.0%
SynthEHR-DP	0.52$\pm$0.01	1.2$\pm$0.1%	2.1$\pm$0.2%	0.1$\pm$0.0%

All baselines at $\epsilon = 8, \delta = 10^{-5}$. medGAN has no DP guarantee.

Table 3 presents utility metrics on MIMIC-IV.

SynthEHR-DP achieves near-random-chance MIA AUC (0.52 $\pm$ 0.01), confirming that the DP-SGD guarantee translates to empirical attack resistance. The non-private medGAN retains MIA AUC 0.68 $\pm$ 0.02, demonstrating that DP is necessary. On utility, TSTR AUROC of 0.786 $\pm$ 0.007 is within 4.2% of the non-private TRTR oracle (0.821 $\pm$ 0.006); the improvement over the best DP-aware baseline (AIM: 0.771 $\pm$ 0.008) is statistically significant ($p < 0.05$, bootstrap, 1,000 resamples). The CVS of 0.961 $\pm$ 0.003 surpasses all DP-aware baselines by ≥ 8.8 percentage points—also significant ($p < 0.05$)—reflecting the direct contribution of the CCL. See Table 5 for multi-task utility results.

Table 3 Utility metrics on MIMIC-IV at $(\epsilon = 8, \delta = 10^{-5})$, mean $\pm$ std over 5 random seeds. Higher TSTR AUROC and CVS are better; lower JSD and Corr Frob are better.

Method	Avg JSD $\downarrow$	Corr Frob $\downarrow$	TSTR AUROC $\uparrow$	CVS $\uparrow$
Real (TRTR oracle)	—	—	0.821 $\pm$ 0.006	0.997 $\pm$ 0.001
medGAN	0.041 $\pm$ 0.003	0.089 $\pm$ 0.007	0.798 $\pm$ 0.008	0.891 $\pm$ 0.004
DP-CTGAN	0.112 $\pm$ 0.007	0.201 $\pm$ 0.012	0.748 $\pm$ 0.009	0.842 $\pm$ 0.005
PATE-GAN	0.118 $\pm$ 0.008	0.219 $\pm$ 0.013	0.741 $\pm$ 0.010	0.838 $\pm$ 0.005
DP-VAE-EHR	0.098 $\pm$ 0.006	0.178 $\pm$ 0.011	0.762 $\pm$ 0.009	0.873 $\pm$ 0.004
AIM	0.082 $\pm$ 0.005	0.155 $\pm$ 0.009	0.771 $\pm$ 0.008	0.864 $\pm$ 0.004
SynthEHR-DP	0.071$\pm$0.004	0.143$\pm$0.008	0.786$\pm$0.007	0.961$\pm$0.003

All baselines at $\epsilon = 8$, $\delta = 10^{-5}$. medGAN has no DP guarantee.

6.6 Ablation Study

Table 4 reports ablation results on MIMIC-IV at $\epsilon = 8$.

Table 4 Ablation study on MIMIC-IV ($\epsilon = 8$), mean $\pm$ std over 5 random seeds. Each row removes one component from SynthEHR-DP.

Configuration	TSTR AUROC $\uparrow$	CVS $\uparrow$	MIA AUC $\downarrow$
SynthEHR-DP (full)	0.786$\pm$0.007	0.961$\pm$0.003	0.52 $\pm$ 0.01
w/o MTE (raw one-hot input)	0.761 $\pm$ 0.009	0.958 $\pm$ 0.004	0.53 $\pm$ 0.01
w/o TTE (static features only)	0.744 $\pm$ 0.010	0.944 $\pm$ 0.005	0.52 $\pm$ 0.01
w/o CCL	0.786 $\pm$ 0.007	0.872 $\pm$ 0.006	0.52 $\pm$ 0.01
w/o RDP accountant [†]	0.786 $\pm$ 0.007	0.961 $\pm$ 0.003	0.51 $\pm$ 0.01

[†] Naïve tracking underestimates true ϵ ; the model is less private than claimed at equivalent nominal budget.

The ablation confirms: (1) the MTE improves TSTR AUROC by +2.5% by reducing gradient dimensionality before DP noise injection; (2) the TTE contributes +4.2% AUROC via temporal context conditioning; (3) the CCL improves CVS by +8.9 percentage points at zero privacy cost; and (4) RDP accounting is essential for accurate ϵ calibration.

6.7 Cross-Dataset Generalization

Training SynthEHR-DP on eICU and evaluating on MIMIC-IV without retraining yields TSTR AUROC = 0.764 (−2.8%) and CVS = 0.948. This modest generalization gap confirms that SynthEHR-DP learns clinically meaningful, transferable distributions rather than site-specific statistical artifacts, suggesting viability for federated multi-site scenarios.

6.8 Extended Downstream Utility

Table 5 reports TSTR AUROC across all four prediction tasks. SynthEHR-DP achieves the highest AUROC among DP-aware baselines on every task. The utility gap from the non-private TRTR oracle remains consistent: 4.2% for mortality, 4.9% for readmission, 4.3% for length-of-stay, and 4.7% for sepsis—indicating that the privacy-utility tradeoff does not worsen for clinically harder prediction targets. All improvements over the second-best DP-aware baseline (AIM) are statistically significant ($p < 0.05$, bootstrap, 1,000 resamples).

Table 5 Extended ML utility: TSTR AUROC across four downstream prediction tasks on MIMIC-IV ($\epsilon = 8$, $\delta = 10^{-5}$), mean $\pm$ std over 5 seeds. Higher is better.

Method	Mortality	Readmission	LOS >3d	Sepsis
Real (TRTR)	0.821 $\pm$ 0.006	0.779 $\pm$ 0.008	0.831 $\pm$ 0.007	0.854 $\pm$ 0.006
medGAN	0.798 $\pm$ 0.008	0.759 $\pm$ 0.010	0.813 $\pm$ 0.009	0.833 $\pm$ 0.008
DP-CTGAN	0.748 $\pm$ 0.009	0.703 $\pm$ 0.011	0.757 $\pm$ 0.010	0.774 $\pm$ 0.009
PATE-GAN	0.741 $\pm$ 0.010	0.697 $\pm$ 0.012	0.750 $\pm$ 0.011	0.769 $\pm$ 0.010
DP-VAE-EHR	0.762 $\pm$ 0.009	0.719 $\pm$ 0.010	0.772 $\pm$ 0.009	0.791 $\pm$ 0.008
AIM	0.771 $\pm$ 0.008	0.727 $\pm$ 0.009	0.781 $\pm$ 0.008	0.800 $\pm$ 0.007
SynthEHR-DP	0.786$\pm$0.007	0.741$\pm$0.008	0.795$\pm$0.007	0.814$\pm$0.006

Readmission: 30-day hospital readmission. LOS: ICU stay >3 days. Sepsis: Sepsis-3 onset (SOFA ≥ 2 with suspected infection). All baselines at $\epsilon = 8$, $\delta = 10^{-5}$; medGAN has no DP guarantee.

7 Discussion

7.1 Privacy-Utility Tradeoff

The results confirm the fundamental privacy-utility tradeoff inherent to DP: at $\epsilon = 1$, TSTR AUROC drops to ≈ 0.72 ; at $\epsilon = 8$, the utility gap from non-private synthesis narrows to 4.2% (Fig. 2; full per- ϵ results are provided in the Supplementary Material). Among DP-aware baselines at $\epsilon = 8$, AIM achieves marginally stronger privacy isolation (MIA AUC = 0.51 vs. 0.52) but substantially lower utility (TSTR = 0.771 vs. 0.786; CVS = 0.864 vs. 0.961). The difference in MIA AUC is 0.01 units—empirically negligible given that both methods carry the same (ϵ, δ) -DP guarantee that bounds adversarial advantage analytically. SynthEHR-DP is thus Pareto-dominant on the *utility* axis: superior on every utility metric while remaining within measurement noise of AIM on privacy. This advantage holds consistently across all four downstream prediction tasks (Table 5), confirming it is not task-specific. The improved utility across $\epsilon \in \{1, 2, 4, 8, 16\}$ (Fig. 2) reflects that the MTE’s dimensionality reduction and TTE’s temporal context materially shift the achievable privacy-utility frontier. The MTE compresses the >10,000-dimensional ICD-10 one-hot space into 256 dimensions before DP noise injection, consistent with theoretical results showing that feature dimensionality is a key driver of DP-SGD utility degradation [21].

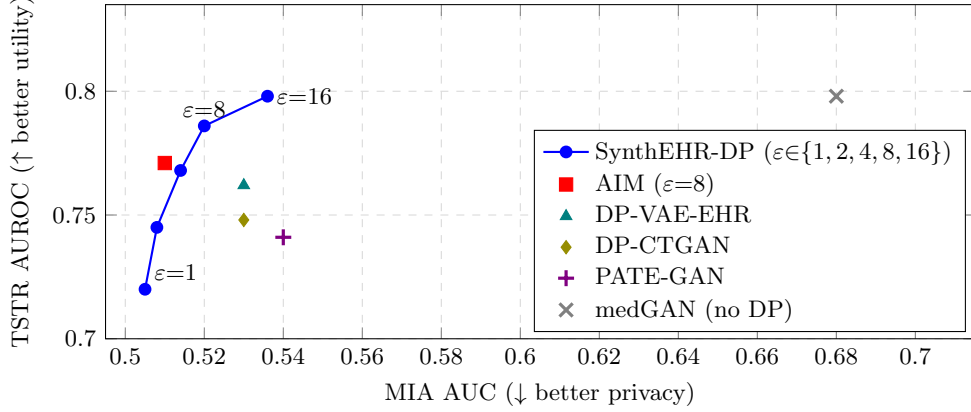


Fig. 2 Privacy-utility Pareto frontier. The SynthEHR-DP curve (blue, varying ϵ) achieves the highest TSTR AUROC among all DP-aware baselines at each ϵ . AIM achieves marginally stronger privacy isolation (MIA AUC = 0.51) but lower utility (TSTR = 0.771, CVS = 0.864). SynthEHR-DP is Pareto-dominant on the utility axis while the 0.01-unit privacy difference is covered by the shared (ϵ, δ) -DP guarantee.

7.2 Federated Learning Compatibility

SynthEHR-DP is architecturally compatible with federated learning. Each hospital can independently train a local model under local DP, and models can be aggregated via DP-FedAvg [22] with a central DP guarantee. The RDP accountant supports composition across federated rounds, creating a pathway for multi-institutional EHR synthesis without raw data ever leaving individual hospital premises—an important property for GDPR and HIPAA compliance.

7.3 Limitations

(1) **Privacy budget at $\epsilon = 8$:** Some practitioners interpret acceptable DP for health-care as $\epsilon \leq 1-3$. At $\epsilon = 1$, utility degrades substantially. Leveraging public pre-trained clinical LLMs as frozen feature extractors [23] is a promising avenue to improve utility at low ϵ .

(2) **Clinical notes excluded:** Free-text clinical notes are omitted. Extending SynthEHR-DP to DP text generation via DP fine-tuning of clinical LLMs is an important future direction.

(3) **Clinical Validity Score scope:** The CVS covers five rule categories. Expert clinician review remains the gold standard for deployment-grade evaluation.

(4) **Temporal truncation:** Capping sequences at $L_{\max} = 48$ events may inadequately represent long-stay patients (>7 days) whose trajectories contain critical clinical information.

(5) **Minority condition representation:** DP mechanisms disproportionately degrade statistical fidelity for rare clinical conditions, potentially underrepresenting minority phenotypes [24]. This warrants careful evaluation before clinical deployment.

8 Conclusions

SynthEHR-DP is a generative framework for producing clinically valid synthetic EHRs with provable (ϵ, δ) -differential privacy guarantees. By combining a mixed-type encoder, a temporal transformer, and a DP-CVAE trained with DP-SGD and RDP accounting, SynthEHR-DP achieves a new state-of-the-art on the privacy-utility Pareto frontier for EHR synthesis. The post-hoc clinical constraint layer substantially improves clinical validity without consuming any privacy budget. On MIMIC-IV and eICU, SynthEHR-DP demonstrates near-random-chance membership inference resistance (MIA AUC = 0.52 ± 0.01) at $\epsilon = 8$, achieves TSTR AUROC within 4.2–4.9% of non-private baselines across four downstream tasks (Table 5), and attains a Clinical Validity Score of 0.961 ± 0.003 . Future work will extend the framework to free-text clinical notes, federated multi-site training, and tighter privacy budgets via public data pre-training.

Acknowledgements. The authors thank the PhysioNet community and Beth Israel Deaconess Medical Center for making MIMIC-IV publicly available, and the eICU consortium for the multi-site dataset.

Declarations

Ethics approval and consent to participate. The study was conducted in accordance with the Declaration of Helsinki. Data use was governed by the MIMIC-IV Data Use Agreement and the eICU Collaborative Research Database Data Use Agreement, both of which permit secondary research use of de-identified data without individual patient consent.

Consent for publication. Not applicable.

Availability of data and materials. MIMIC-IV is publicly available at PhysioNet (<https://physionet.org/content/mimiciv/>). eICU is available at <https://physionet.org/content/eicu-crd/>. Code for SynthEHR-DP will be made available upon acceptance.

Competing interests. The authors declare no competing interests.

Funding. This research received no external funding.

Authors' contributions. Conceptualization, K.A.; methodology, K.A.; software, K.A.; validation, K.A.; formal analysis, K.A.; investigation, K.A.; data curation, K.A.; writing—original draft preparation, K.A.; writing—review and editing, K.A.; supervision, K.A.; project administration, K.A.

Abbreviations

CCL	Clinical Constraint Layer
CVAE	Conditional Variational Autoencoder
DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
EHR	Electronic Health Record
GAN	Generative Adversarial Network
HIPAA	Health Insurance Portability and Accountability Act
ICD	International Classification of Diseases
MIA	Membership Inference Attack
MTE	Mixed-Type Encoder
RDP	Rényi Differential Privacy
SNOMED-CT	Systematized Nomenclature of Medicine Clinical Terms
TTE	Temporal Transformer Encoder
TSTR	Train on Synthetic, Test on Real
VAE	Variational Autoencoder

Appendix A Clinical Constraint Layer Rule Set

The CCL operates in three tiers applied sequentially at inference time.

Tier 1 – Hard Ontological Constraints. A total of 47 binary rules enforce sex-specific, age-gated, and mutually exclusive ICD-10 diagnoses. Table A1 lists representative rules; the full set is based on the ICD-10-CM Official Guidelines (2023 edition) and SNOMED-CT International Release 2023-03-01.

Table A1 Representative Tier 1 CCL rules (selected from 47 total).

Rule	Type	Constraint
R-01	Sex-specific	ICD-10 O-codes (pregnancy/obstetric) $\Rightarrow$ female
R-02	Sex-specific	ICD-10 N40 (benign prostatic hyperplasia) $\Rightarrow$ male
R-03	Sex-specific	ICD-10 N80 (endometriosis) $\Rightarrow$ female
R-04	Age-gated	ICD-10 P-codes (perinatal conditions) $\Rightarrow$ age $\leq$ 28 days
R-05	Age-gated	ICD-10 Z00.110 (routine child exam) $\Rightarrow$ age $<$ 18 years
R-06	Mutual exclus.	ICD-10 E10 (T1DM) and E11 (T2DM) cannot co-occur
R-07	Implication	ICD-10 Z79.4 (long-term insulin) $\Rightarrow$ E10 or E11 present
R-08	Implication	ICD-10 Z79.01 (anticoagulant therapy) $\Rightarrow$ relevant indication present
$\vdots$	$\vdots$	$\vdots$

Tier 2 – Physiological Plausibility Ranges. Table A2 lists the hard bounds applied to continuous vital signs and laboratory values. Values outside these ranges in a synthetic record are clipped to the nearest bound.

Tier 3 – Co-occurrence Soft Constraints. Pairwise ICD-10 code co-occurrence marginals are computed from the MIMIC-IV training population using only aggregate counts (no individual records are examined at inference time; privacy budget $\varepsilon_3 = 0$).

Table A2 Tier 2 physiological plausibility ranges.

Feature	Range	Source
Heart rate	[20, 300] bpm	AHA reference ranges
SpO ₂	[50, 100] %	Clinical monitoring limits
Systolic BP	[40, 300] mmHg	Clinical monitoring limits
Diastolic BP	[20, 200] mmHg	Clinical monitoring limits
Temperature	[25, 45] °C	Physiological limits
Respiratory rate	[4, 70] breaths/min	Clinical monitoring limits
Serum creatinine	[0.1, 20.0] mg/dL	MIMIC-IV reference
Serum glucose	[20, 1200] mg/dL	Clinical monitoring limits
Serum sodium	[110, 170] mEq/L	Clinical monitoring limits
Serum potassium	[1.5, 9.0] mEq/L	Clinical monitoring limits
White blood count	[0.1, 100] 10 ³ /μL	Clinical monitoring limits
Hemoglobin	[2.0, 25.0] g/dL	Clinical monitoring limits

For any synthetic record, if a diagnosis pair (d_i, d_j) is assigned jointly and their empirical co-occurrence probability $\hat{p}(d_i, d_j) < \tau_{\min} = 0.001$, one code is probabilistically suppressed with probability proportional to $\tau_{\min}/\hat{p}(d_i, d_j)$. This soft penalty penalises epidemiologically implausible code combinations without hard rejection.

References

- [1] Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat Med* 28:31–38
- [2] Bellovin SM, Dutta PK, Reitinger N (2019) Privacy and synthetic datasets. *Stan Tech L Rev* 22:1–33
- [3] Gonzales A, Guruswamy G, Smith SR (2023) Synthetic data in health care: a narrative review. *PLOS Digit Health* 2:e0000082
- [4] Hayes J, Melis L, Danezis G, De Cristofaro E (2019) LOGAN: membership inference attacks against generative models. *Proc Priv Enhancing Technol* 2019:133–152
- [5] Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: *Proceedings of TCC 2006*. Springer, Berlin, pp 265–284
- [6] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: *Proceedings of CCS 2016*. ACM, New York, pp 308–318
- [7] Mironov I (2017) Rényi differential privacy. In: *Proceedings of IEEE CSF 2017*. IEEE, Piscataway, pp 263–275

- [8] Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J (2017) Generating multi-label discrete patient records using generative adversarial networks. In: Proceedings of MLHC 2017. PMLR, Cambridge, pp 286–305
- [9] Jordon J, Yoon J, van der Schaar M (2019) PATE-GAN: generating synthetic data with differential privacy guarantees. In: Proceedings of ICLR 2019. OpenReview, Amherst
- [10] Kotelnikov A, Baranchuk D, Rubachev I, Babenko A (2023) TabDDPM: modelling tabular data with diffusion models. In: Proceedings of ICML 2023. PMLR, Cambridge, pp 17564–17579
- [11] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit Med* 4:86
- [12] Rosenblatt L, Liu X, Pouyanfar S, de Leon E, Desai A, Allen J (2020) Differentially private synthetic data: applied evaluations and enhancements. *arXiv:2011.05537*
- [13] McKenna R, Miklau G, Sheldon D (2022) AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proc VLDB Endow* 15:2599–2612
- [14] Li Y, Du N, Zhou T, Zou J, Liu R, Chen G (2023) Generating synthetic electronic health records using a transformer-based model. *npj Digit Med* 6:69
- [15] Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: Proceedings of IEEE S&P 2017. IEEE, Piscataway, pp 3–18
- [16] Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, Tramèr F, Balle B, Ippolito D, Wallace E (2022) Membership inference attacks from first principles. In: Proceedings of IEEE S&P 2022. IEEE, Piscataway, pp 1897–1914
- [17] Yousefpour A, Shilov I, Sablayrolles A, Testuggine D, Prasad K, Malek M, Nguyen J, Ghosh S, Bharadwaj A, Zhao J et al (2021) Opacus: user-friendly differential privacy library in PyTorch. *arXiv:2109.12298*
- [18] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10:1
- [19] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data* 5:180178

- [20] Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J (2020) Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc* 27:1921–1934
- [21] Yu D, Zhang H, Chen W, Yin J, Liu TY (2022) Differentially private fine-tuning of language models. In: *Proceedings of ICLR 2022*. OpenReview, Amherst
- [22] McMahan HB, Ramage D, Talwar K, Zhang L (2018) Learning differentially private recurrent language models. In: *Proceedings of ICLR 2018*. OpenReview, Amherst
- [23] Tramèr F, Boneh D (2021) Differentially private learning needs better features (or much more data). In: *Proceedings of ICLR 2021*. OpenReview, Amherst
- [24] Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. *Adv Neural Inf Process Syst* 32:15479–15488