TempEHR: A Temporal Dependency-Based Approach for Synthesizing Electronic Health Records

Emmanuella Budu[1]✉, Amira Soliman[1], Farzaneh Etminani[1], and Thorsteinn Rögnvaldsson[1]

Halmstad University, Halmstad 301 18, Sweden
{emmanuella.budu,amira.soliman,farzaneh.etminani,thorsteinn.rognvaldsson}@hh.se

Abstract. Synthetic Electronic Health Records (EHRs) provide a viable means of accessing EHR data while addressing the privacy concerns related to the use of EHRs. A key characteristic of EHRs is the irregular timing of clinical events, admissions, and associated temporal trends. Many existing models for generating synthetic EHRs overlook these temporal irregularities, often assuming uniform intervals between clinical events for each patient and neglecting the time component, which hinders the representation of true temporal dynamics. To address these limitations, we propose TempEHR, a framework designed to synthesise EHRs, emphasising temporal awareness. We employ a time-aware Variational Autoencoder (VAE), specifically a Maximum Mean Discrepancy VAE (MMD-VAE), leveraging Time-aware Long Short-Term Memory (T-LSTM) layers to generate temporal synthetic EHRs along with time information. Simultaneously, we enhance the temporal awareness of our proposed model with a novel network we refer to as a TrendFinder. TrendFinder leverages a moving average to extract the temporal patterns inherent in irregular longitudinal EHR data. This approach seeks to enhance the fidelity and usefulness of synthetic EHRs for research and clinical applications. We assess the effectiveness of TempEHR using EHRs from the Medical Information Mart for Intensive Care (MIMIC-IV) repository. Our results demonstrate the potential of the proposed method in capturing the temporal patterns present in EHRs in utility, fidelity and privacy evaluations.

Keywords: synthetic data · Electronic Health Records (EHRs) · temporal data · time-series analysis.

1 Introduction

Recent advancements in artificial intelligence (AI) have accelerated research in studies on data-driven medicine, where electronic health records (EHRs) serve as the primary data source for these studies. EHRs are a valuable source of information to facilitate such research studies and enhance patient care outcomes [1] as they encompass time-sequenced records of various clinical events and interactions between patients and healthcare providers over time [2].

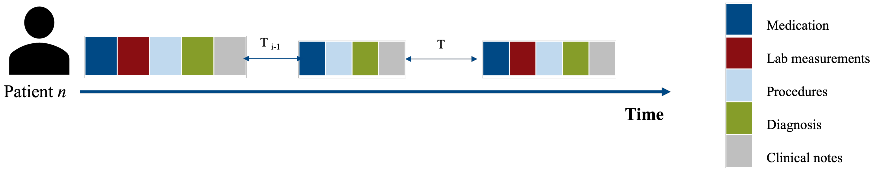


Fig. 1: An illustration of the irregularities in patient records. Time intervals are not regular, i.e., $T_{i-1} \neq T$, and different actions and measurements can be done at each visit.

One promising area in the work on EHRs is the generation of synthetic EHRs using deep generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and most recently, Large Language Models. This is being investigated as a viable means of obtaining EHRs that do not compromise the privacy regulations surrounding real EHRs [1]. Significant efforts have been made in recent years in several studies [3–8] to generate diverse and temporal synthetic EHRs. Although viable, these existing approaches ignore one fundamental characteristic of EHRs and assume regularity between clinical events in patient records, which is an unrealistic perception.

EHRs are inherently irregular due to varied patient visits, admissions and interactions, as illustrated in Figure 1. In healthcare, we encounter multivariate EHRs with irregularities attributed to missing recordings, irregular patient visits, and varying lab measurements. This irregularity can also be attributed to economic and social factors [5], such as accessibility to healthcare facilities in underprivileged societies.

Existing EHR generation methods proposed in several studies [1, 4, 6] employ recurrence methods such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) [9, 10] that assume regular intervals between successive elements. This is a significant limitation because it compromises the validity of the generated EHRs, as the temporal dependency is inaccurately modelled. Several studies [9–11] have shown that using vanilla recurrence models such as LSTMs to model irregular data leads to subpar and inconsistent performance in modelling progression patterns in EHRs. In addition, some studies [12] have demonstrated that traditional recurrence architectures also struggle with representing or learning the trends present in temporal data.

More recent studies in synthetic EHR generation have begun addressing the challenges posed by the inherent temporal irregularities in EHRs. For example, Yoon et al. [1] explored incorporating irregular time information as an additional feature in observed patient data before synthesising the data. While this method attempts to model temporal irregularity, it treats the irregularities associated with clinical events as just another feature in the observed EHR data. As a result, the approach limits the generative model's ability to fully capture and leverage the temporal dependencies between events, as the model does not fully account for the time irregularities in the data.

Other research efforts [13, 14] have sought to address the challenge of modelling temporal irregularity in synthetic EHRs by transforming time information into embedding vectors that are concatenated with the latent space representation during training. However, modelling the complexities of temporal irregularity in synthetic EHR generation remains an underexplored area. Moreover, several studies have emphasised the critical role of modelling irregular time intervals when synthesising EHRs [5,6,14], highlighting the importance of advancing research in this domain.

Building on these efforts, we present TempEHR, which explores a time-aware approach that leverages time information to enhance the generative model's ability to capture temporal dependencies within EHRs. Our approach employs an information-maximising variational autoencoder (MMD-VAE) [15] with time-aware long short-term memory networks (T-LSTMs) [9], transforming time intervals between clinical events into weights. These weights are used to adjust the memory in the recurrence layers, emphasising close events. Inspired by the work of Lin et al. [16], we supplement this with a novel module, which we call TrendFinder, which learns the underlying temporal trends in the real data to enhance the quality of the synthetic EHRs.

To the best of our knowledge, this is the first work to combine time-aware models, such as T-LSTMs, with trend-extracting methods within a generative framework for synthetic EHRs. To this end, the contributions of this work are as follows:

- We propose TempEHR, a generative framework for synthesising temporal EHRs that accounts for the irregular patterns and associated temporal trends inherent in EHRs.
- We leverage the time intervals between clinical events as weights to prioritise recent data and capture the temporal patterns in irregular EHRs.
- We introduce a novel TrendFinder module that leverages a time exponential moving average to learn the underlying trends in irregularly timed EHRs.
- We evaluate the effectiveness of TempEHR on a real-world publicly available EHR dataset on fidelity, utility and privacy measures.

2 Related Work

2.1 Modelling Irregular Data in Deep Learning

Irregularity in sequenced or temporal data refers to temporal data characterised by non-uniform intervals between successive time points. This may arise from irregular sampling, variable observations, and misaligned time points [10]. Irregular data is particularly common in healthcare, where data is often recorded at inconsistent intervals due to varying processes, patient behaviours, and conditions. This paper specifically addresses the challenge of modelling irregularity related to varying intervals between successive time points in EHRs.

In deep learning for temporal data, recurrent models such as RNNs and LSTMs are among the most widely used architectures. Despite their success across various tasks, these models face challenges when dealing with irregular data as they typically assume fixed or relatively small intervals between successive data points [9, 10]. As a result, they are often inefficient in handling data with irregular intervals, struggling to capture the evolving patterns in temporal data [11]. Common approaches to handling this include incorporating additional information about irregularities, such as indicators or time intervals, before using a recurrence model [17]. In contrast, some studies have examined methods that modify the underlying LSTM model by adding time-gating mechanisms [10] or transforming the time intervals between consecutive observations into weights, which adjust the contents of the hidden state accordingly [9]. Moreover, attention models have been employed to help the model focus on recent parts of the input by using positional encodings [10].

2.2 Synthesising Temporal EHRs

Existing methods for generating temporal EHRs often utilise recurrent architectures within deep generative models, such as GANs and VAEs. Recurrence-based GAN models, including TimeGAN [3], TAP-GAN [4], DAAE [7], and EHR-Safe [1], initially employ an autoencoder to learn the underlying representations of real EHRs that are used in combination with a GAN to generate synthetic EHRs.

Similarly, Biswal and Ghosh [5] utilise a VAE with convolutions, while Nikolentzos et al. [8] introduced an approach that employs a Variational Graph Autoencoder (VGA) to model patient trajectories. Other research efforts, such as EHR-M-GAN [6], combine a VAE and a GAN with Coupled Recurrent Networks (CRN) to generate mixed-type longitudinal EHRs.

Additionally, some studies have explored the use of diffusion models [14] to generate synthetic temporal EHRs. However, most assume a fixed interval for clinical events for the EHRs.

An emerging focus in recent studies is the generation of synthetic EHRs that handle the time irregularity inherent in EHRs. For instance, EHR-Safe [1] employs an autoencoder-based framework that models the irregular timing of events as a feature. However, it does not use these time intervals to guide the overall generation process.

IGAMT [13] and EHRPD [14] also address irregularities by learning the relationships between features across different time steps. IGAMT utilises a transformer-based encoder-decoder framework, while EHRPD employs a diffusion-based autoencoder structure. Both models calculate time intervals between events, converting these increments into embedding vectors concatenated with the latent space representation of the EHR data.

EHR-Safe [1], IGAMT [13], and EHRPD [14] all explore different approaches for addressing temporal irregularity in EHR generation. However, the field is still under-explored and presents opportunities for further research.

2.3 Moving Averages

Moving averages (MA) are key concepts in temporal data analysis, defined as a time series generated by averaging consecutive values from the original data [18]. For a given time series $x_i, x_{i+1}, \ldots, x_k$, the moving average over k time steps can be expressed as:

$$\mathrm{MA}_t = \frac{1}{k} \sum_{i=t-k+1}^{t} x_i \tag{1}$$

In time series analysis, moving averages serve two main purposes: they help to track the behaviour of a time series by smoothing out fluctuations and revealing underlying trends, and they are also used to forecast future values [18,19].

Different variations of moving average methods exist, with the most common being the simple moving average, which computes the arithmetic mean of a set of observations. The cumulative moving average calculates an average by including all previous data points, while exponential moving averages apply a smoothing factor to prioritise recent observations.

Moving averages can also be adapted for irregular temporal data. In such cases, the moving averages are adjusted to account for the time intervals between observations. Menth and Hauser [19] proposed various techniques for incorporating these time intervals into the moving average calculations for irregular temporal data.

3 TempEHR

3.1 Problem Definition

Let $D = \{(x_i, a_i)\}_{i=1}^{N}$ represent the multivariate EHR dataset, where N is the total number of patient records. Each record x_i is a sequence of events for different variables, and a_i is the corresponding sequence of inter-admission intervals, representing the time between consecutive events. Specifically, $x_i = \{x_i^1, x_i^2, \ldots, x_i^T\}$, where x_i^t denotes the event at time step t for the i-th patient, and $a_i = \{a_i^1, a_i^2, \ldots, a_i^{T-1}\}$, where a_i^t denotes the associated time interval at time step t for the i-th patient. The goal is to learn an approximate distribution $\hat{p}(D)$, from which we can sample data points to create a synthetic dataset, $\hat{D}$.

3.2 TempEHR

TempEHR [1] comprises a time-aware VAE combined with an additional moving average-based neural network, which we call the TrendFinder, to handle the irregularity inherent in EHRs as illustrated in Figure 2. TempEHR simultaneously learns a latent representation of the irregular EHRs and the underlying trend across time. This yields an enriched latent space that can be used to synthesise temporal EHRs and the associated time intervals.

[1] Code repository for TempEHR: https://github.com/EmmanuellaBudu/TempEHR

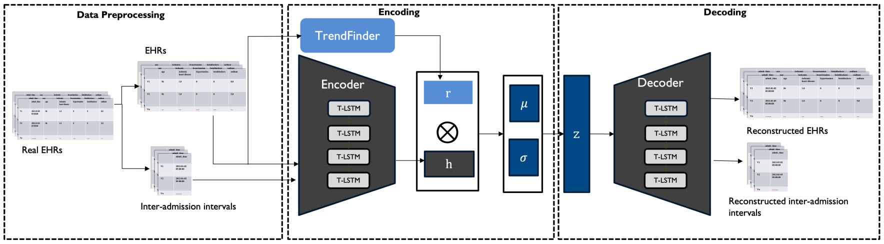


Fig. 2: Overview of TempEHR. TempEHR leverages a T-LSTM-based VAE to generate synthetic EHRs and associated time intervals.

3.3 Learning temporal dependencies

TempEHR utilises T-LSTM layers [9] within the encoder and decoder modules of a VAE framework. The Encoder takes as input the multivariate EHRs containing clinical events along with the associated time intervals, Δt, between visits/admissions. The time intervals Δt are transformed into weights using a time decay function in the Encoder. These weights are then used to adjust the cell state contents, subsequently influencing the learned hidden representation h, from the T-LSTM-based Encoder.

Additionally, we incorporate a TrendFinder network inspired by the work of Lin et al. [16] that employs a time exponential moving average (TEMA)to identify trends, r. The motivation for using TrendFinder is to learn the temporal dependencies in irregularly-timed data better. By enhancing the latent representation, we aim to improve overall model performance in generating high-fidelity synthetic EHRs. TEMA is described in Equation 2:

$$\mathrm{TEMA}_t = e^{-\Delta t/\tau} \cdot x_t + \left(1 - e^{-\Delta t/\tau}\right) \cdot \mathrm{TEMA}_{(t-1)} \tag{2}$$

Where: x_t is the value at time t, τ is the time constant, Δt is the time difference between t and $t-1$. We employ TEMA after evaluating various moving average methods to identify the most effective one. Additionally, it aligns with the inherent mechanisms of T-LSTMs, emphasising recent values. This approach enables us to capture the underlying trend despite the irregularities effectively.

Next, the representations r and h are concatenated and passed through linear layers to generate μ and σ vectors, which parameterise the latent space z. The latent space z is regularised with Maximum Mean Discrepancy (MMD) to obtain a more informative prior [15] as compared to the traditional Kullback-Leibler (KL) divergence. Next, the Decoder uses samples from z to reconstruct the input EHRs and the associated time intervals.

Table 1: Description of datasets.

	MIMIC-CHF	MIMIC
Total Admissions	8835	21380
Total Patients	1767	2138
Total Features	24	23
Max Seq. Length	5	10

3.4 TempEHR Loss

TempEHR employs three loss functions during training to generate synthetic EHRs. First, we utilise a VAE loss, which is composed of a reconstruction loss and an MMD loss on the latent space, where $p(z)$ is the prior distribution and $q(z)$ is the posterior distribution over the EHRs:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{Reconstruction}} + \lambda_1 \mathcal{MMD}\left[p(z), q(z)\right] \tag{3}$$

Second, we incorporate a mean squared error (MSE) loss to model the time intervals explicitly. This loss penalises the difference between the real-time and generated time intervals. This explicit focus on time intervals provides finer control over time, which is critical in this context:

$$\mathcal{L}_{\text{Time}} = \lambda_2 \text{MSE}(a_{\text{real}}, a_{\text{syn}}) \tag{4}$$

The total loss is a weighted sum of the reconstruction loss, MMD loss, and time loss, where the weights, λ_1 and λ_2 are empirically determined to balance the contributions of each term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Reconstruction}} + \lambda_1 \mathcal{MMD}\left[p(z), q(z)\right] + \lambda_2 \mathcal{L}_{\text{Time}} \tag{5}$$

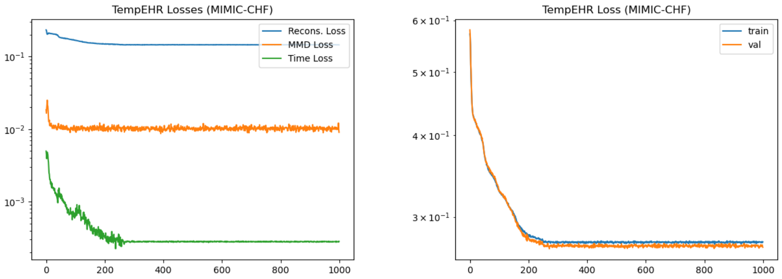
4 Experiments

4.1 Datasets

We utilise real-world EHRs from the Medical Information Mart for Intensive Care (MIMIC-IV) data repository. This publicly accessible database contains de-identified patient records from the Beth Israel Deaconess Medical Centre, covering the years from 2001 to 2012 [20].

For this study, we extracted two separate sets of EHRs from the hospital module, which includes records from general hospital stays. We target a cohort of heart failure patients (MIMIC-CHF) alongside a general group of patients without specific disease diagnoses (MIMIC) as done in previous studies [1]. We extract demographics, vital signs, lab measurements, and co-morbidity flags. The characteristics of the datasets are described in Table 1.

The dataset is organised so that each patient is represented by a time-ordered sequence of admissions containing data recorded during that admission, often



(a) (b)

Fig. 3: TempEHR training and validation losses on MIMIC-CHF dataset.

known as a patient trajectory. The dataset includes both discrete and continuous values. During preprocessing, discrete values are transformed into continuous values by mapping them to the range [0,1], following the approach described in [21]. The interval [0,1] is divided into sections based on the cumulative probability of each unique discrete value. Each discrete value is then mapped to a point within its corresponding section. Additionally, all continuous values are normalised to the range [0,1].

4.2 Baselines

We consider two baseline temporal generative models: (i) EHR-M-GAN [6], a dual-VAE model with a Coupled Recurrent Network (CRN) generator for generating heterogeneous temporal EHRs, and (ii) TimeGAN [3], a state-of-the-art GAN-based time-series data generator that employs LSTMs or Gated Recurrent Units (GRUs). For these baselines, we model the irregular timing of events as a feature [1]. Due to some code unavailability, technical challenges, and time constraints with the implementations of some other related works [1,13,14], we were unable to directly include the models as baselines. Despite these limitations, the selected baselines are useful benchmarks for evaluating our proposed method.

4.3 Training Details

TempEHR is implemented in PyTorch and trained with the Adam optimiser and a learning rate of 0.001. We employ a reduce-on-plateau learning rate scheduler to adapt the learning rate during training. Furthermore, we partition the datasets using a 70:30 split; the loss from training on the MIMIC-CHF is illustrated in Figure 3, showing the combined losses and individual losses. The rest of the model parameters are as follows: hidden size:128, latent size:64, encoder and decoder layers:3, λ_1: 4.97,λ_2:0.39.

4.4 Evaluation

This study employs a comprehensive evaluation framework to assess the synthetic EHRs generated by TempEHR and the baseline models, focusing on *fidelity*, *utility* and *privacy* [22] evaluations. Fidelity refers to the faithfulness of the synthetic data to the real data. We employ fidelity measures to assess whether the structural and temporal relationships in the real EHRs have been replicated in synthetic EHRs. Specifically, we assess structural similarity in low dimensions using Uniform Manifold Approximation and Projection (UMAP), discriminative accuracy [3, 6], and temporal dynamics [23] using Dynamic Time Warping (DTW) and trend similarity. In contrast, utility focuses on assessing the usefulness of synthetic data compared to real data. Specifically, we assess the performance of a predictive model trained on the synthetic EHRs and tested on the real EHRs. Lastly, privacy measures evaluate whether real patient information has been leaked into the synthetic data. We employ the membership inference attack to determine whether real patient records used to train the generative model can be inferred based on the synthetic data. We report results averaged over three independently generated synthetic datasets.

5 Results and Discussions

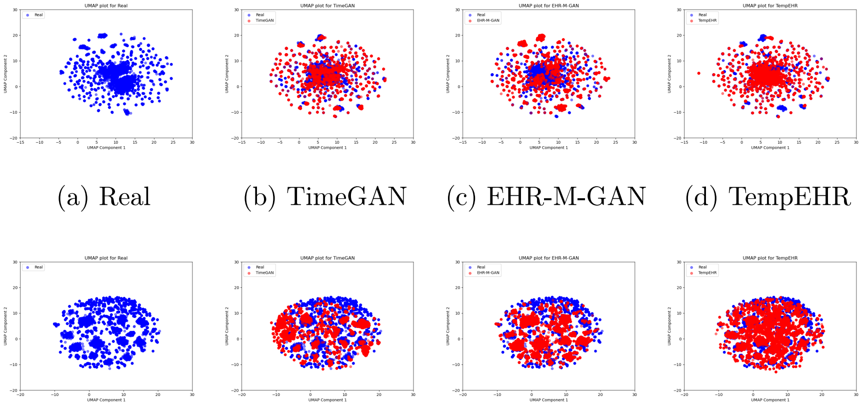
5.1 Structural Similarity with Dimensionality Reduction

We first assess the structural similarity between real and synthetic EHRs by visualising a low-dimensional representation of the data with UMAP, presented in Figure 4. The blue points indicate real patient admissions, while the red points represent synthetic patient admissions. To ensure comparable visualisations, we combine the real and synthetic EHR datasets, add an indicator variable, and transform them into a lower-dimensional space.

From the visualisation, we observe that the coverage of the synthetic EHRs generated by TempEHR overlaps the real EHRs better than the synthetic EHRs generated by TimeGAN and EHR-M-GAN for the MIMIC-CHF and MIMIC datasets. TempEHR captures most of the significant and minor clusters present in the real data. This demonstrates that the TempEHR is better at replicating the underlying distribution of the real EHR data, which is important given the complexity of healthcare datasets such as EHRs. A synthetic EHR dataset should effectively replicate the global structural properties of real EHRs.

5.2 Discriminative Accuracy

To assess how distinguishable the synthetic patient trajectories are from the real patient trajectories, we compute the discriminative accuracy [6] of a T-LSTM-based classifier evaluated on data generated by TimeGAN, EHR-M-GAN, and TempEHR. A classifier that cannot distinguish between real and synthetic EHRs would achieve an accuracy of 0.5.



(a) Real (b) TimeGAN (c) EHR-M-GAN (d) TempEHR

(e) Real (f) TimeGAN (g) EHR-M-GAN (h) TempEHR

Fig. 4: UMAP visualisation for real and synthetic admission data from MIMIC-CHF (top row) and MIMIC (bottom row) patients.

As shown in Table 2, TempEHR achieves a lower discriminative accuracy of 0.751 (±0.042) in the MIMIC-CHF dataset compared to TimeGAN (0.778 ±0.011) and EHR-M-GAN (0.757 ±0.028). The classifier struggles more to distinguish TempEHR-generated trajectories from real EHRs, indicating that TempEHR better preserves the sequential structure of real EHRs, making them appear more realistic.

For the MIMIC dataset, the discriminative accuracies for all models are close to 0.5, with TempEHR achieving an accuracy of 0.534 (±0.032). This indicates that the classifier has difficulty differentiating between real and synthetic trajectories across the models. All models generate synthetic EHRs that are nearly indistinguishable from the real EHRs.

Table 2: Discriminative scores and predictive errors. In both cases, lower values indicate better performance.

Metric	**Model**	**MIMIC-CHF**	**MIMIC**
Discriminative Accuracy	TimeGAN	0.778 ± 0.011	$\mathbf{0.525 \pm 0.030}$
	EHR-M-GAN	0.757 ± 0.028	0.540 ± 0.011
	TempEHR	$\mathbf{0.751 \pm 0.042}$	0.534 ± 0.032
Predictive Error (MAE)	Real	$\mathbf{0.243 \pm 0.000}$	$\mathbf{0.230 \pm 0.000}$
	TimeGAN	0.268 ± 0.014	0.243 ± 0.010
	EHR-M-GAN	0.266 ± 0.007	$\mathbf{0.238 \pm 0.004}$
	TempEHR	$\mathbf{0.262 \pm 0.002}$	0.242 ± 0.009

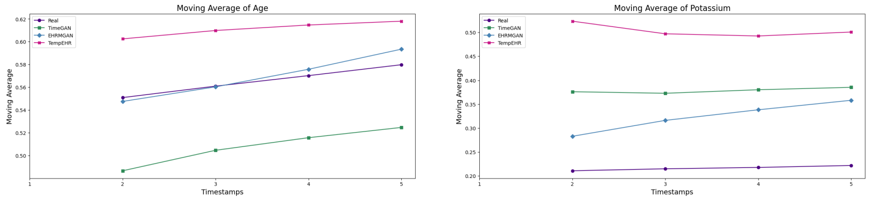


Fig. 5: Moving average across time for the features, Age and Potassium in real and synthetic data. Moving average starts at timestep 2.

5.3 Trend Similarity

Trends describe the overall change in data across time. We compute feature-level trend similarity scores, focusing on continuous variables such as patient age, lab values and vital signs.

First, we classify trends by assessing the probability of variables increasing, decreasing, fluctuating or remaining constant over time within a predefined normal range according to the MIMIC repository [20] (except for age). We then compare these distributions between the real and synthetic EHRs. As an example, we illustrate the moving average over time along with the probability distribution of possible trends for Age and Potassium from the MIMIC-CHF dataset, as shown in Figures 5 and 6.

We employ Jensen-Shannon Divergence (JSD) between the distributions, shown in Table 3, to quantify the similarity. Scores range from 0 (identical distributions) to 1(dissimilar distributions).

As shown in Table 3, TempEHR shows more consistent trend similarity scores compared to EHR-M-GAN and TimeGAN, indicating better trend replication. For patients' ages with clear trends over time, TempEHR consistently outperforms both models due to its TrendFinder Network, which models the temporal dependency.

However, some inconsistencies are observed in Figure 6, such as deviations in age trends among a small subset of the synthetic EHRs. This originates from the limitations of modelling temporal relationships in the generative process [24]. While these can be addressed through further processing of the generated EHRs, our analysis focuses on the raw generated EHRs to evaluate the trends in the data.

In reality, EHRs often contain complex, nonlinear temporal patterns that generative models struggle to fully capture [24]. TimeGAN and EHR-M-GAN exhibit variabilities across the different variables, while TempEHR demonstrates a more consistent performance. This reflects TempEHR's potential to model temporal trends and maintain the global structure of the real EHRs.

5.4 Temporal Dynamics

We employ a multivariate DTW-based assessment [23] to evaluate the similarity in overall temporal dynamics between patient trajectories in the real and syn-

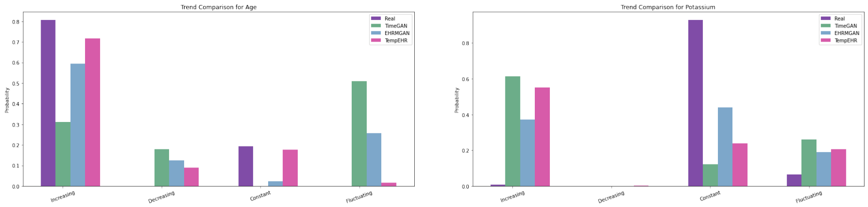


Fig. 6: Histogram showing the probability of belonging to different trends over time for the features Age and Potassium.

thetic EHRs. DTW is a time-series similarity measure that quantifies the similarity between time series that may have different local shifts and speeds [25]. We first compute pairwise DTW similarity scores [26] in a real-vs-synthetic setting to assess the similarity, comparing real patient trajectories with synthetic trajectories. We then report the average DTW similarity score of the most similar synthetic patient trajectories in Table 4. This measures how well the synthetic data captures the temporal dynamics in the real patient trajectories. The similarity score ranges from 0 to 1, where 0 indicates dissimilarity and 1 indicates similarity.

The table shows that TempEHR consistently achieves high DTW scores across the MIMIC-CHF and MIMIC cohorts, with scores of 0.649 (±0.002) and 0.633 (±0.001), respectively. TimeGAN also demonstrates strong capabilities in modelling the temporal dynamics of patient trajectories. On the other hand, EHR-M-GAN has the lowest scores among the models, which suggests it has challenges in accurately replicating the underlying temporal movements and shifts in the real patient trajectories.

A crucial aspect of EHR synthesis is modelling the temporal nature of real EHRs. In real EHRs, the order and timing of events are crucial as they influence clinical decisions. As such, the time between clinical events plays an important role in modelling. Most studies on EHR modelling tend to ignore this. Thus producing outputs that may not fully capture the temporal patterns present in real patient trajectories [9]. TempEHR attempts to address this by incorporating time information between patient visits in the generative model, enabling it to better replicate shifts in values better over time.

5.5 Predictive Errors

We evaluate the utility of the generated data on a downstream sequence prediction task using the TSTR framework. Specifically, we train a T-LSTM-based model to predict different patient outcomes. For the MIMIC-CHF cohort, we predict whether a patient visit was planned or not. In contrast, for the MIMIC cohort, we predict the hospital expire flag, which indicates whether a patient died during hospitalisation or survived.

We compare the Mean Absolute Error (MAE) of the predictions from the synthetic data with the predictions from the real data as a baseline. Table 2

Table 3: Trend similarity on continuous variables in the MIMIC-CHF and MIMIC-RSP cohorts. Lower scores (bolded) are optimal. SBP: Systolic Blood Pressure, DBP: Dystolic Blood Pressure, O2sat: Oxygen Saturation, Resprate: Respiration Rate.

Dataset	Variable	TimeGAN	EHR-M-GAN	TempEHR
MIMIC-CHF	Age	0.661 ± 0.132	0.415 ± 0.019	**0.199 ± 0.009**
	Potassium	0.711 ± 0.167	**0.387 ± 0.080**	0.513 ± 0.056
	Urea Nitrogen	0.129 ± 0.072	**0.067 ± 0.009**	0.143 ± 0.016
	Sodium	0.424 ± 0.006	0.465 ± 0.027	**0.377 ± 0.134**
	Creatinine	0.324 ± 0.057	0.173 ± 0.165	**0.140 ± 0.060**
	Chloride	0.368 ± 0.067	0.412 ± 0.058	**0.267 ± 0.022**
	Hematocrit	0.144 ± 0.115	**0.076 ± 0.021**	0.143 ± 0.021
	Hemoglobin	0.105 ± 0.085	**0.028 ± 0.016**	0.147 ± 0.034
MIMIC	Age	0.831 ± 0.001	0.832 ± 0.001	**0.545 ± 0.100**
	Heartrate	**0.069 ± 0.001**	0.125 ± 0.052	0.124 ± 0.099
	Resprate	0.189 ± 0.001	0.194 ± 0.044	**0.116 ± 0.019**
	O2sat	0.206 ± 0.001	**0.072 ± 0.036**	0.262 ± 0.109
	SBP	**0.102 ± 0.001**	0.152 ± 0.016	0.103 ± 0.052
	DBP	0.289 ± 0.001	**0.175 ± 0.016**	0.260 ± 0.063
	Hemoglobin	0.118 ± 0.066	**0.057 ± 0.003**	0.135 ± 0.106
	Glucose	0.138 ± 0.001	0.175 ± 0.010	**0.127 ± 0.022**
	Temperature	0.439 ± 0.001	**0.396 ± 0.005**	0.468 ± 0.069

shows the MAE for the CHF and MIMIC datasets. The table shows that the TempEHR-generated data performs comparably to the real data for predictive purposes, obtaining an MAE of 0.262 (±0.002) in the MIMIC-CHF dataset and 0.242 (±0.009) in the MIMIC dataset. This demonstrates that machine learning models trained on TempEHR-generated data can generalise well to real-world scenarios, maintaining low prediction errors. Notably, TimeGAN and EHR-M-GAN yield relatively low predictive errors as well.

Evaluating the utility of EHR generation models allows us to determine whether these generated EHRs can be utilised for medical or clinical purposes, such as analysing healthcare policies and planning resources [23].

5.6 Privacy Evaluation

We assess the privacy preservation of the generated data using membership inference attack as in previous studies [1, 6], and report the accuracy of these attacks. An optimal attack accuracy is approximately 0.5, indicating that the attacker's performance is no better than random guessing.

As illustrated in Figure 7, TempEHR achieves a relatively lower attack accuracy of 0.612 (±0.038) for the MIMIC-CHF cohort and 0.616 (±0.072) for the MIMIC cohort. Likewise, TimeGAN and EHR-M-GAN obtain attack accuracies of 0.624 (±0.072) and 0.632 (±0.112) for the MIMIC-CHF cohort, respectively.

Table 4: Average pairwise DTW similarity scores between patient trajectories from the real and synthetic data. Higher scores (bolded) indicate the closest match to real data.

	MIMIC-CHF	**MIMIC**
TimeGAN	0.535 ± 0.001	$\mathbf{0.669 \pm 0.001}$
EHR-M-GAN	0.546 ± 0.002	0.304 ± 0.002
TempEHR	$\mathbf{0.649 \pm 0.002}$	0.633 ± 0.001

For the MIMIC cohort, TimeGAN and EHR-M-GAN report accuracies of 0.609 (±0.181) and 0.628 (±0.093), respectively.

Notably, all models yield attack accuracies close to 0.6, with TempEHR yielding consistently low accuracies across both the MIMIC-CHF and MIMIC datasets. This indicates similar levels of privacy preservation and limited leakage of patient identities. The near-random attack accuracy for TempEHR, TimeGAN, and EHR-M-GAN demonstrates their effectiveness in maintaining patient identities while generating plausible synthetic EHRs.

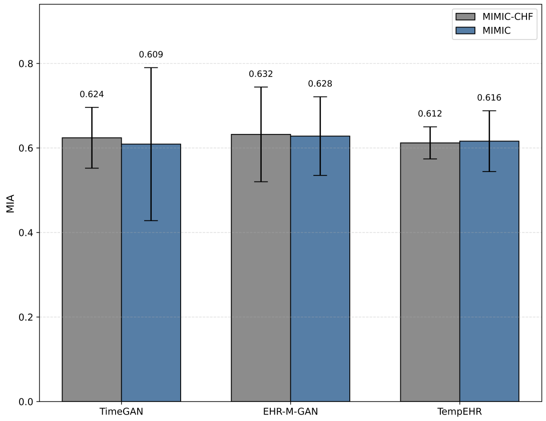


Fig. 7: Accuracy on MIA

5.7 Ablation Studies

To evaluate the effectiveness of the components of TempEHR, we conduct an ablation study to assess the contribution of each component. We report the discriminative scores and predictive errors from these in Table 5. The experiments are described as follows: S-1: TempEHR without the TrendFinder, S-2: TempEHR without T-LSTM layers, and S-3: TempEHR without time loss.

Table 5: Discriminative scores and predictive errors from ablation studies. S-1:TempEHR without the TrendFinder, S-2: TempEHR without T-LSTM layers, S-3: TempEHR without time loss.

	Model	**MIMIC-CHF**	**MIMIC**
Discriminative Accuracy	S-1	0.806 ± 0.032	0.537 ± 0.019
	S-2	0.793 ± 0.018	0.574 ± 0.080
	S-3	0.791 ± 0.039	0.573 ± 0.043
Predictive Score	S-1	0.291 ± 0.012	0.258 ± 0.004
	S-2	0.267 ± 0.017	0.277 ± 0.013
	S-3	0.289 ± 0.018	0.254 ± 0.009

From Table 5, we observe that T-LSTM layers, TrendFinder, and time loss components all contribute to the overall performance of TempEHR in learning an enriched representation of the data and generating synthetic EHRs.

6 Conclusions

In this work, we proposed TempEHR, a generative framework for synthesising irregularly-timed EHRs. TempEHR enhances the generation of synthetic EHRs by incorporating an enriched latent state representation that captures temporal dynamics, enabling more consistent replication of temporal dependencies in real EHRs than previous generators. In modelling healthcare data, capturing the order and timing of clinical events is essential for evaluating a patient's health status at any point in time. Our results demonstrate that TempEHR can preserve the global and temporal relationships of real EHRs, achieving low predictive errors and a more consistent performance across most evaluation measures. These findings highlight its potential for improving synthetic EHRs for healthcare research and innovation.

Although TempEHR performs well on most temporal patterns, some remain challenging, as seen in the trend similarity scores. In future work, we aim to refine the architecture to better replicate complex clinical patterns, enhancing the quality of synthetic EHRs for healthcare applications.

References

1. Yoon, J., Mizrahi, M., Ghalaty, N.F., Jarvinen, T., Ravi, A.S., Brune, P., Kong, F., Anderson, D., Lee, G., Meir, A., Bandukwala, F., Kanal, E., Arık, S.Ö., Pfister, T.: EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. npj Digit. Med. 6(1), 1–11 (Aug 2023), https://www.nature.com/articles/s41746-023-00888-7, number: 1 Publisher: Nature Publishing Group
2. Lee, J.M., Hauskrecht, M.: Personalized Event Prediction for Electronic Health Records. Artificial Intelligence in Medicine 143, 102620 (Sep 2023), http://arxiv.org/abs/2308.11013, arXiv:2308.11013 [cs, stat]

3. Yoon, J., Jarrett, D., van der Schaar, M.: Time-series Generative Adversarial Networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf
4. Hashemi, A.S., Etminani, K., Soliman, A., Hamed, O., Lundström, J.: Time-series Anonymization of Tabular Health Data using Generative Adversarial Network. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (Jun 2023), https://ieeexplore.ieee.org/document/10191367, iSSN: 2161-4407
5. Biswal, S., Ghosh, S.: EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders. In: Proceedings of Machine Learning Research. vol. 149, p. 22 (2021)
6. Li, J., Cairns, B.J., Li, J., Zhu, T.: Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. npj Digit. Med. 6(1), 1–18 (May 2023), https://www.nature.com/articles/s41746-023-00834-7, number: 1 Publisher: Nature Publishing Group
7. Lee, D., Yu, H., Jiang, X., Rogith, D., Gudala, M., Tejani, M., Zhang, Q., Xiong, L.: Generating sequential electronic health records using dual adversarial autoencoder. Journal of the American Medical Informatics Association 27(9), 1411–1419 (Sep 2020), https://academic.oup.com/jamia/article/27/9/1411/5912632
8. Nikolentzos, G., Vazirgiannis, M., Xypolopoulos, C., Lingman, M., Brandt, E.G.: Synthetic electronic health records generated with variational graph autoencoders. npj Digit. Med. 6(1), 1–12 (Apr 2023), https://www.nature.com/articles/s41746-023-00822-x, number: 1 Publisher: Nature Publishing Group
9. Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient Subtyping via Time-Aware LSTM Networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 65–74. KDD '17, Association for Computing Machinery, New York, NY, USA (Aug 2017), https://dl.acm.org/doi/10.1145/3097983.3097997
10. Shukla, S.N., Marlin, B.M.: A Survey on Principles, Models and Methods for Learning from Irregularly Sampled Time Series (Jan 2021), http://arxiv.org/abs/2012.00168, arXiv:2012.00168 [cs, stat]
11. Zhang, Y., Yang, X., Ivy, J., Chi, M.: ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling pp. 4369–4375 (2019), https://www.ijcai.org/proceedings/2019/607
12. Wu, Y., Meng, X., Zhang, J., He, Y., Romo, J.A., Dong, Y., Lu, D.: Effective LSTMs with seasonal-trend decomposition and adaptive learning and niching-based backtracking search algorithm for time series forecasting. Expert Systems with Applications 236, 121202 (Feb 2024), https://www.sciencedirect.com/science/article/pii/S0957417423017049
13. Wang, W., Tang, P., Lou, J., Shao, Y., Waller, L., Ko, Y.a., Xiong, L.: IGAMT: Privacy-Preserving Electronic Health Record Synthesization with Heterogeneity and Irregularity. Proceedings of the AAAI Conference on Artificial Intelligence 38(14), 15634–15643 (Mar 2024), https://ojs.aaai.org/index.php/AAAI/article/view/29491, number: 14
14. Zhong, Y., Wang, X., Wang, J., Zhang, X., Wang, Y., Huai, M., Xiao, C., Ma, F.: Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models (Jun 2024), http://arxiv.org/abs/2406.13942, arXiv:2406.13942 [cs]
15. Zhao, S., Song, J., Ermon, S.: Infovae: Balancing learning and inference in variational autoencoders. In: Proceedings of the aaai conference on artificial intelligence. vol. 33, pp. 5885–5892 (2019)

16. Lin, T., Guo, T., Aberer, K.: Hybrid neural networks for learning the trend in time series. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 2273–2279. IJCAI'17, AAAI Press, Melbourne, Australia (Aug 2017)
17. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR Workshop Conf Proc 56, 301–318 (Aug 2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5341604/
18. Hyndman, R.J.: Moving Averages, pp. 866–869. Springer Berlin Heidelberg, Berlin, Heidelberg (2011), https://doi.org/10.1007/978-3-642-04898-2_380
19. Menth, M., Hauser, F.: On Moving Averages, Histograms and Time-DependentRates for Online Measurement. In: Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering. pp. 103–114. ICPE '17, Association for Computing Machinery, New York, NY, USA (Apr 2017), https://dl.acm.org/doi/10.1145/3030207.3030212
20. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data 3(1), 160035 (May 2016), https://www.nature.com/articles/sdata201635, number: 1 Publisher: Nature Publishing Group
21. Patki, N., Wedge, R., Veeramachaneni, K.: The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 399–410. IEEE, Montreal, QC, Canada (Oct 2016), http://ieeexplore.ieee.org/document/7796926/
22. Budu, E., Etminani, K., Soliman, A., Rögnvaldsson, T.: Evaluation of synthetic electronic health records: A systematic review and experimental assessment. Neurocomputing 603, 128253 (2024), https://www.sciencedirect.com/science/article/pii/S0925231224010245
23. Budu, E., Soliman, A., Rögnvaldsson, T., Etminani, F.: Evaluating Temporal Fidelity in Synthetic Time-series Electronic Health Records. In: 2024 IEEE Conference on Artificial Intelligence (CAI). pp. 541–548 (Jun 2024), https://ieeexplore.ieee.org/abstract/document/10605528
24. Liu, Y., Wijewickrema, S., Li, A., Bester, C., O'Leary, S., Bailey, J.: Time-Transformer: Integrating Local and Global Features for Better Time Series Generation (Jan 2024), http://arxiv.org/abs/2312.11714, arXiv:2312.11714 [cs]
25. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. p. 359-370. AAAIWS'94, AAAI Press (1994)
26. Wang, N., Wang, M., Zhou, Y., Liu, H., Wei, L., Fei, X., Chen, H.: Sequential Data-Based Patient Similarity Framework for Patient Outcome Prediction: Algorithm Development. J Med Internet Res 24(1), e30720 (Jan 2022), http://www.ncbi.nlm.nih.gov/pubmed/34989682