Evaluating Temporal Fidelity in Synthetic Time-series Electronic Health Records

1st Emmanuella Budu
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
emmanuella.budu@hh.se

2nd Amira Soliman
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
amira.soliman@hh.se

3rd Thorsteinn Rögnvaldsson
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
thorsteinn.rognvaldsson@hh.se

4th Farzaneh Etminani
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
farzaneh.etminani@hh.se

***Abstract*—Synthetic data generation has been proposed as a potential solution to accessing Electronic Health Records (EHRs) while minimizing the privacy risks associated with real EHRs. Nevertheless, the practical use of synthetic EHRs rests on their ability to resemble the quality of real EHRs. Existing evaluations of synthetic EHRs often focus on assessing them as static snapshots frozen in time, neglecting temporal dependencies and varying temporal patterns. Moreover, some of these methods rely on subjective judgments, are limited to segmentable time-series, and employ methods that adopt a one-to-one approach. This study employs a comprehensive approach to evaluating fidelity in synthetic time-series EHRs to address these challenges. We extend the functionality of time-series analysis methods such as temporal clustering, time-series similarity measures, Sample Entropy, and trend analysis, to evaluate varying temporal patterns in synthetic time-series EHRs. Our findings provide valuable insights into how synthetic EHRs align with real EHRs in the temporal context, considering aspects such as patient groupings, temporal dynamics, predictability, and directional change. We empirically demonstrate the feasibility of assessing temporal fidelity with these methods, offering an understanding of the quality of synthetic EHRs in capturing the varying temporal patterns inherent in EHRs.**

***Index Terms*—synthetic data, Electronic Health Records (EHRs), times-series, fidelity, similarity**

I. Introduction

Electronic Health Records (EHRs), as illustrated in Fig. 1, comprise a comprehensive health history of a patient recorded at different points in time. In recent years, synthetic EHR generation has emerged as an alternative to obtaining real EHRs, potentially addressing the privacy concerns of using real medical data [1]. Synthetic EHRs, like their real counterparts, are a valuable resource for a wide range of purposes, including research studies, testing new algorithms, medical education, developing healthcare systems, and facilitating the public release of data [3]. For instance, a study [4] investigated the use of synthetic EHRs as a proxy for real EHRs sourced from the New York State Department of Health to predict the length of stay (LOS) of patients in hospitals. Another study [5] explored using synthetic data to investigate healthcare policies and make decisions on resource allocations.

Synthetic EHRs are commonly represented in two formats: static snapshots frozen in time or time-series that portray the patient's health status changing over time. Generating static EHRs involves learning the static features' statistical properties to generate new data points that correspond with these properties. In contrast, generating time-series EHRs is more challenging, comprising modelling the temporal dependencies in features within a time frame and across time [6]. To this end, some recent works [6, 7, 8] have proposed generative models that capture these temporal correlations. This involves mapping the time-series data to a latent space that captures these temporal correlations, thus enhancing the generation process. Furthermore, these studies employ autoregressive architectures such as Recurrent Neural Networks (RNNs) to model sequences of medical events. These models learn the temporal dependencies within time-series data and use this learned representation to generate new data. Consequently, the quality of the synthetic data rests on how well the temporal dependencies are modelled.

Once generated, synthetic EHRs are evaluated to ascertain whether the synthetic data preserves the real data's statistical and structural properties. This is formally known as fidelity. To this end, methods that evaluate variable distributions [1], low dimensional representations, correlations, and statistics [1] at a single time-point have been employed to assess fidelity. A drawback of this is that these methods overlook the temporal dependencies and patterns inherent in time-series EHRs [9] as they only assess single time points.

A handful of studies have attempted to assess the fidelity of synthetic EHRs while considering the temporal dependencies in time-series data. For instance, Li et al. [7] employed visual inspection, comparing patient trajectories in the real and synthetic EHRs to evaluate fidelity. However, the subjective nature of this approach creates inaccurate assessments [9]. An

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00107

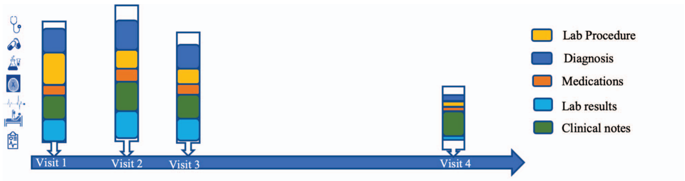


Fig. 1: An Electronic Health Record showing a series of clinical events with lab procedures, diagnosis, prescribed medications, lab results, and clinical notes for a patient over different visits [2]

alternative approach employed by Li et al. [7] employed the autocorrelation function (ACF) to examine the preservation of temporal correlations for selected variables in synthetic data. Additionally, another study [8] employed an approach that compares the length of sequences and event distributions in real and synthetic cases.

In assessing temporal fidelity, Dash et al. [10] employed a methodology to segment the time-series data into meaningful intervals to calculate and compare the summary statistics in real and synthetic cases. However, this approach is only effective when the time-series can be partitioned into intervals. Pearson's Correlation, employed in some studies [7] and [9] revealed limitations in evaluating non-linearly correlated data, as Pearson's method assumes independence in all observations [11].

Furthermore, Bhanot et al. [9] introduced one-to-one metrics such as Root Mean Squared Error (RMSE), Directional Symmetry (DS), and Short Time-series Distance (STS) to assess fidelity. Like the Euclidean Distance measure, these employ a lock-step approach that does not account for temporal misalignments [12].

Given these challenges, our work proposes a comprehensive approach to evaluating temporal fidelity in synthetic time-series EHRs. The proposed approach effectively evaluates varying temporal patterns in synthetic time-series EHRs, overcoming limitations associated with individual judgments, one-to-one approaches, and segmentable time-series. Consequently, our work aims to make the following contributions:

1) We extend the functionality of time-series analysis methods such as temporal clustering, time-series similarity measures, Sample Entropy, and trend analysis for assessments on fidelity. These methods account for varying temporal dynamics and patterns.
2) We experimentally demonstrate the feasibility of these methods for assessing fidelity in three synthetically generated time-series EHRs. We showcase the effectiveness of these methods for evaluating synthetic EHRs by employing two state-of-the-art generators, TimeGAN [6] and EHR-M-GAN [7], alongside a randomly generated dataset.
3) Our experiments illustrate the application and interpretation of these methods in the context of the fidelity of synthetic time-series EHRs. This includes insights into how these methods effectively discern temporal patterns and similarities.

The remainder of the paper is organized as follows: Section II provides a background of the time-series methods employed. Section III details the data, synthetic data generators, and methods for assessing temporal fidelity in synthetic EHRs. Section IV presents the experimental results. Section V provides a discussion of the findings, and Section VI presents conclusions and future research directions.

II. Background

A. Temporal Clustering

Temporal clustering groups time-series data using a predefined similarity measure [13]. Three main categories of temporal clustering methods exist hierarchical, model-based, and partitioning. However, we focus on the commonly used partitioning methods. As such, two key considerations for partition-based clustering are the optimal number of clusters and the cluster quality.

Methods for determining optimal cluster numbers include statistical (e.g., gap statistic) and direct methods (e.g., Elbow and Silhouette methods) [14, 15]. To assess cluster quality, measures like the Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index, and Adjusted Mutual Information (AMI) [16] can be employed, but most require known class labels except for the Silhouette Coefficient.

B. Time-series Similarity Measures

Time-series similarity measures quantify the degree of similarity between time-series sequences. Different types of time-series similarity measures [12] exist, however, we focus on the commonly used Dynamic Time Warping (DTW) [17] in this paper.

DTW in (1) aligns temporal sequences by accommodating local shifts and differences in speed. Given two sequences $A = (a_1, a_2, \ldots, a_m)$ and $B = (b_1, b_2, \ldots, b_n)$ of length m and n, DTW finds a warping path, W that maps the elements in the two sequences such that their distance is minimized [17]. A warping path is determined by constructing a distance matrix, d of size m by n, whose entries are the distances between elements in A and B. The warping path, of length p is the optimal path between the distance matrix that minimizes the cost of aligning the two sequences.

$$\mathrm{DTW}(A, B) = \min_{W} \left(\sum_{k=1}^{p} d(a_k, b_k) \right) \tag{1}$$

C. Sample Entropy (SampEn)

SampEn [18], described in (2) is a measure used to quantify the regularity or predictability of data in a time-series. It determines whether similar patterns persist throughout the time-series. SampEn is characterized by three parameters: N, r and m. N is the length of the time-series. r is the tolerance level for accepting a match between elements in a time-series. m is the length of sequences to be compared, while C is a count of patterns of length, m within the tolerance level r. The resulting entropy is computed by calculating the negative log probability with which sequences of length m from the time-series remain similar as the time interval varies. An important characteristic of SampEn is that it is robust because it is independent of the length of the time-series, making it suitable for time-series data of varying lengths [18]. Additionally, SampEn is not highly insensitive to its hyperparameters as demonstrated in some studies [19]. Lower values of SampEn close to 0 indicate low entropy, hence regularity and vice-versa.

$$SampEn(m, r) = -\ln\left(\frac{C_{m+1}(r)}{C_m(r)}\right) \tag{2}$$

D. Trend Analysis

Trend analysis, according to Rae [20], is a method used to assess how things change over time. This approach involves identifying patterns or trends to offer insights into the temporal changes and facilitate predictions about behaviours. As emphasized by Ely et al. [21] trend analysis is valuable in cases where the trend is not apparent, with limited data, or when there is a large variability in rates from consecutive periods. This analysis sheds light on whether there is an increasing or decreasing trend in data.

III. Methodology

This section presents an overview of the methodology employed in this paper as illustrated in Figure 2.

We employed temporal clustering, time-series similarity measures, Sample Entropy, and trend analysis to measure the similarity between real and synthetic EHR time-series while paying attention to the inherent temporal patterns.

A. Dataset

We utilized EHRs from the Medical Information Mart for Intensive Care (MIMIC-IV) data repository. The MIMIC-IV repository is a publicly accessible database of de-identified patient records from the Beth Israel Deaconess Medical Centre from 2001 to 2012 [22]. We specifically extracted EHRs containing vital signs and corresponding lab measurements from the emergency department (ED) module. The module contains measurements of vital signs documented for patients during their stay in the ED. This corresponds to real data. Table I provides an overview of the variables utilized in our study. We define our dataset, S as a collection of m patients. Each patient, denoted as a_i is characterized by a tuple of n vital sign recordings taken over several observations:

$$S = \{a_1, a_2, \ldots, a_m\}$$

Where each a_i is defined as:

$$a_i = (v_1, v_2, \ldots, v_n)$$

We imputed missing values for the initial preprocessing stage using the mean value in a day's recording. Subsequently, we aggregated visits on the same day and retained the initial ten visits for each patient. Lastly, we normalized the values to the same scale.

B. Assessment of Fidelity

In assessing fidelity, we first employ k-means clustering with DTW to group the real EHRs, examining whether the synthetic EHRs exhibit similar clustering patterns. The optimal number of clusters for the real EHRs is determined using the Silhouette Coefficient. We combine the real dataset with each synthetic dataset, assigning labels to distinguish real from synthetic records. Subsequently, each combined dataset is clustered, and fidelity is assessed by determining the proportion of real and synthetic records in each cluster, with the algorithm executed multiple times for result consistency.

Secondly, DTW distances are computed over patient sequences for selected variables from a cohort from the real and synthetic sets. Multiple separate DTW comparisons are conducted, each between a sequence in one set and all other sequences in the other set. A heatmap derived from the distance matrix illustrates the similarity between the patient sequences. To compare the heatmaps, we adopt a similar approach in this study [7] and compare summary statistics (mean, maximum, and minimum) over the distance matrices. This approach aims to determine whether synthetic datasets accurately capture patterns from real records and provide insights into the overall characteristics of the dataset.

Thirdly, we employ SampEn to assess the predictability of the real data and compare it to the synthetic data. We computed SampEn with the parameters for m and r as 2 and 0.3, respectively. For each patient sequence in each variable in the real and synthetic cohorts, we individually compute SampEn values, quantifying the entropy of each patient's time-series. Subsequently, we aggregate the SampEn values for all patients, visualizing the distribution of SampEn in the different cohorts from the synthetic data generators. To assess fidelity, we compare the distributions of SampEn in the real data against the synthetic sets. Additionally, we calculate the mean of the distributions per variable. The comparison of SampEn distributions and the mean values provides insights into whether the predictability in the real data is accurately replicated in the synthetic data.

Lastly, we assess fidelity by extracting the trend and examining how the patient sequences change over time. For each patient sequence in each variable, we fit a polynomial regression model of degree two for each variable. Subsequently, we aggregate the calculated slope for each patient and compute the mean of the slopes across the real and synthetic cohorts.

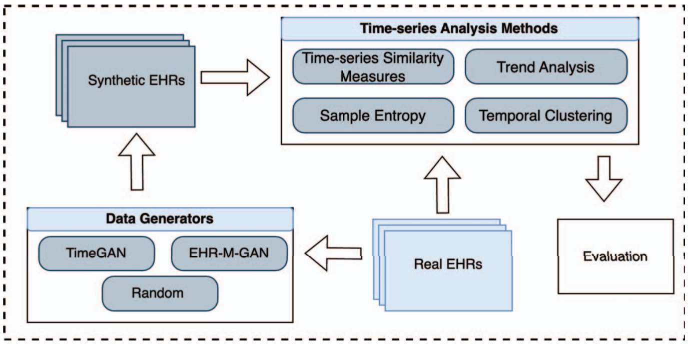


Fig. 2: Overview of the methodology: Synthetic EHRs are generated from TimeGAN, EHR-M-GAN, Randomness and evaluated using time-series analysis methods

TABLE I: Description of variables used in the experiments

Variable	Description	Range	Type
heartrate	Heart rate (beats per minute)	40–140	Discrete
sbp	Systolic Blood Pressure (mmHg)	90–200	Continuous
dbp	Diastolic Blood Pressure (mmHg)	60–200	Continuous
o2sat	Oxygen saturation (%)	75–120	Continuous
resprate	Respiration rate (breaths per minute)	9–40	Discrete
sodium	Sodium levels (mEq/L)	130.0–155.0	Continuous
potassium	Potassium levels (mEq/L)	3.0–6.0	Continuous

We assess fidelity by comparing the direction and magnitude of the slopes for the real and synthetic sets.

C. Synthetic Data Generators

1) TimeGAN: TimeGAN [6] is a state-of-the-art time-series synthetic data generator comprising a GAN with four network components: an embedding function, recovery function, sequence generator, and sequence discriminator. According to Yoon et al., [6], this architecture enables the generation of realistic time series and static data that preserve temporal dynamics. We employ the default parameters for TimeGAN.

2) EHR-M-GAN: EHR-M-GAN [7] comprises a GAN framework capable of generating high-fidelity multivariate synthetic EHRs. This is accomplished by mapping the data into a shared latent through a dual Variational Autoencoder (dual-VAE). Consequently, a sequentially coupled generator built upon a coupled recurrent network (CRN) captures the temporal correlations and generates synthetic data. We employ the default parameters for EHR-M-GAN.

3) Random Data: We introduce a third synthetic EHR dataset by deriving random values between the minimum and maximum values of the different variables in the real EHR dataset.

IV. Experimental Results

A. Temporal Clustering

We determined the optimal number of clusters for the real EHRs as two, as indicated in Table II. The silhouette score 0.45 obtained for two clusters suggests two true patient groups in the real data.

TABLE II: Silhouette Scores for different numbers of clusters

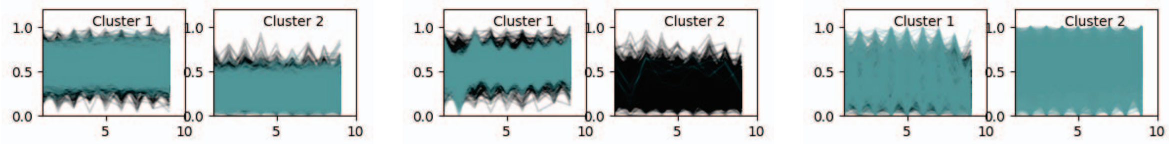
Clusters	2	3	4	5	6	7	8
Silhouette Score	**0.45**	0.26	0.17	0.13	0.12	0.1	0.1

Fig. 3 shows a visual representation of the presence of two main clusters based on the systolic blood pressure variable. In the first cluster, we observe values between 0.2 and 0.9. In contrast, in the second cluster, we observe average to lower systolic blood pressure values. Fig. 3 also suggests that the variance of the synthetic data from TimeGAN and EHR-M-GAN closely matches that of the real data in the two clusters. This contrasts the findings for data generated by randomness, where the variance differs significantly from the real data.

Table III shows the assignment of patient records to the different clusters. C1 and C2 refer to clusters 1 and 2 respectively. In the ideal case, where the fidelity of the synthetic EHRs matches the real data, there should be a balanced proportion of the number of records from the real and synthetic sets in the different clusters. From the table, in the TimeGAN case, we observe, that clusters 1 and 2 have a more balanced proportion of real and synthetic records. In the EHR-M-GAN and Random case, we see that across the two clusters, there is a highly uneven distribution of records from the real and synthetic cases. We also observed an extreme case where the synthetic records make up only 0.35 percent of the records in cluster 2 for the EHR-M-GAN data.

B. Time-series Similarity Measures

Fig. 4 illustrates a heatmap for DTW distance matrices for patients in the real and synthetic cohorts for the systolic



(a) TimeGAN

(b) EHR-M-GAN

(c) Random

Fig. 3: Results showing cluster assignments (k=2) for real and synthetic EHRs for systolic blood pressure. Real data is represented in black, synthetic data is depicted in blue

TABLE III: Percentage of records assigned to the different clusters in the real and synthetic cases

		C1	C2
TimeGAN	**Real**	44.73	53.92
	Synthetic	55.27	46.08
EHR-M-GAN	**Real**	22.67	99.65
	Synthetic	77.33	0.35
Random	**Real**	79.38	20.22
	Synthetic	20.62	79.78

blood pressure variable. We selected a cohort of patients with consecutively high blood pressure readings, defined as a systolic blood pressure exceeding 140 mm/HG, within both the real and synthetic datasets. In Fig. 4a, we illustrate the similarity in the real vs. real case, representing the ideal scenario where records in the real data are compared against each other, resulting in a symmetric matrix. Figures 4b, 4c, 4d show the similarity between the records from the real compared against each of the synthetic data. Each coloured element represents the distance or cost of aligning two patients' sequences concerning their systolic blood pressure. The darker shades represent high dissimilarity, while the lighter shades represent similarity.

From Fig. 4, we see that the regions of similarity and dissimilarity in the data from the synthetic cohorts vary from that of the real cohort. Figures 4c and 4b depict slightly similar patterns to the real data. To compare the overall characteristics of the datasets for preserving temporal dynamics, we compute the maximum, minimum, and mean values over the DTW matrix for patients in both the real and synthetic sets in Table IV. TimeGAN and EHR-M-GAN generated sequences with similar temporal dynamics to the real EHRs. This is evident in the similar values for the maximum, minimum, and mean values. The random data exhibits the highest dissimilarity, as expected.

TABLE IV: Statistics from the DTW matrix in the real and synthetic sets over systolic blood pressure

	Real	**TimeGAN**	**EHR-M-GAN**	**Random**
Max	1.52	1.74	1.16	2.39
Min	0.00	0.01	0.02	0.04
Mean	0.30	0.25	0.26	0.67

C. SampEn

Fig. 5 shows the distribution of SampEn for heartrate in the real and synthetic EHRs. The expected distribution is seen in the real case where we have a peak between SampEn values of 0. Notably, in the EHR-M-GAN case, the data closely resembles the distribution of real data. For TimeGAN, the generated data has more patients with regular heart rates than real data, as evident in the longer peak around SampEn values of 0. Conversely, the randomly generated data displays a slightly different pattern, with fewer patients with regular heartrates.

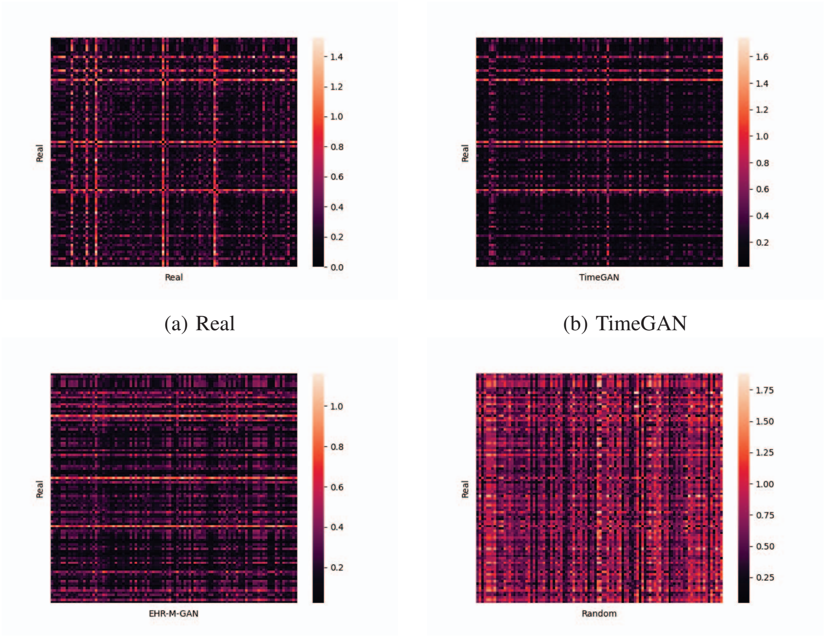
We further compute the SampEn for the remaining variables for each patient record in the real and synthetic EHRs and compare the mean of the distributions of SampEn in Table V. The closest match is seen in heartrate from EHR-M-GAN and systolic blood pressure in the case of the random data. For the rest of the variables under consideration, we observe significant discrepancies in the amount of predictability based on the mean of the distribution of SampEn between the real and synthetic data.

TABLE V: Mean of SampEn recorded over all the variables in the real and synthetic sets

Variable	**Real**	**TimeGAN**	**EHR-M-GAN**	**Random**
heartrate	0.09	0.02	0.08	0.13
sbp	0.14	0.04	0.11	0.13
dbp	0.06	0.10	0.09	0.09
o2sat	0.03	0.22	0.09	0.17
resprate	0.03	0.06	0.05	0.10
sodium	0.02	0.08	0.11	0.11
potassium	0.02	0.11	0.05	0.17

D. Trend Analysis

Table VI presents the mean of the slopes of the different variables for the real and synthetic cohorts. We observe several discrepancies across all the slopes of all the variables presented. First, for the heartrate, the real and synthetic cohorts, except the Random data show positive trends. Secondly, for the blood pressure (sbp and dbp), the patients in the real cohorts have a negative trend, while the synthetic cohort either has a positive trend or a negative trend with varying magnitudes. For the other variables, oxygen saturation, resprate, sodium, and potassium, the real data generally shows positive trends for o2sat and potassium, with negative trends for resprate and sodium. In contrast, the synthetic cohorts display diverse



(a) Real

(b) TimeGAN

(c) EHR-M-GAN

(d) Random

Fig. 4: DTW distance matrix for the real and synthetic sets over systolic blood pressure

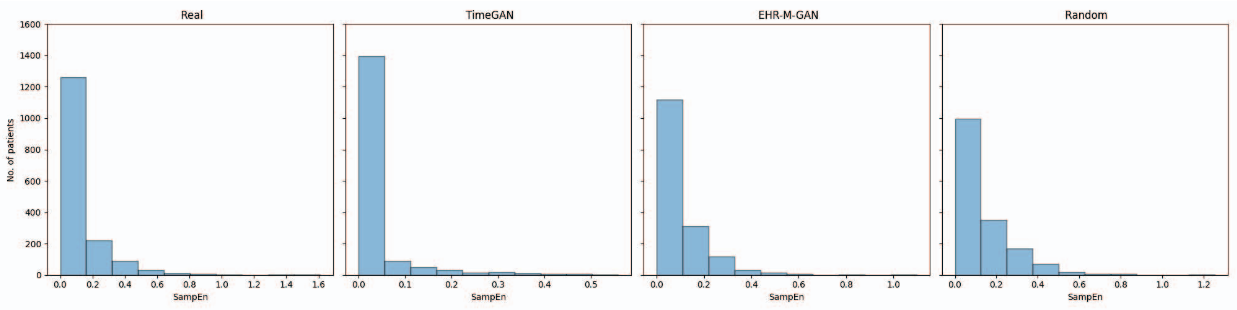


Fig. 5: Distribution of SampEn for the heartrate of patients in real and synthetic EHRs

patterns, capturing some trends but deviating notably in the data from EHR-M-GAN.

TABLE VI: Mean of the distribution of slope over all the variables in the real and synthetic sets

Variable	Real	TimeGAN	EHR-M-GAN	Random
heartrate	0.0026	0.0012	0.0759	-0.0020
sbp	-0.0028	0.0031	-0.0227	-0.0042
dbp	-0.0020	-0.0003	0.0189	0.0002
o2sat	0.0014	0.0008	-0.0401	-0.0007
resprate	-0.0004	0.0022	0.0592	0.0037
sodium	-0.0006	0.0009	0.0206	-0.0050
potassium	0.0021	0.0071	0.0834	-0.0009

V. Discussions

In this study, we investigated the application of time-series analysis methods to evaluate temporal fidelity in time-series synthetic EHRs. These methods aimed at evaluating fidelity by identifying temporal patterns and dependencies. The findings from this study have several notable implications on the fidelity of synthetic EHRs from the aspects of patient groupings, temporal dynamics, predictability, and directional change. We discuss these in the subsequent sections.

A. Patient Groupings

Temporal clustering with DTW identified distinct patient groupings in the real data. Notably, in the case of the systolic blood pressure variable, we observe patients that share common characteristics, with some demonstrating higher values, and others featuring average to lower values. Despite the clear patient groups in the real data, the synthetic cohorts present a few discrepancies in their groupings. These findings indicate that the temporal patterns in the real data are not accurately captured in the synthetic datasets. Various factors could contribute to these discrepancies, including limitations in the generator's inability to generate data that captures the temporal characteristics in the real data.

B. Temporal Dynamics

The DTW distance matrix revealed similarities and dissimilarities in patient records between the real and synthetic cohorts concerning capturing temporal dynamics. In contrast to one-to-one methods [9] which are sensitive to temporal variations, DTW aligns temporal sequences by accommodating local shifts and differences in speed through a one-to-many approach. We observe that the synthetic data captures the temporal variation suggesting a similarity to real EHRs, except for randomly generated data. This suggests a degree of effectiveness in capturing the diverse temporal variations inherent in real EHR data.

C. Predictability

Predictability focuses on the consistency of patterns in the time-series records. SampEn is a measure used to capture the regularity or predictability of the time series. The findings reveal that the predictability of heartrate from the real data is not accurately replicated in the synthetic data. We have a case, where the synthetic datasets have more or fewer patients with regular vital signs. A plausible reason for this can be attributed to the challenges in synthesizing realistic temporal patterns in synthetic data generators.

D. Directional Change

Trend analysis identified the direction and strength of the different variables in the real and synthetic cases. Positive slopes indicate increasing trends over time, while negative trends indicate the opposite. Likewise, the higher the magnitude of the slope, the more variations there are in data and vice-versa. From the findings, the synthetic data did not accurately capture the direction and change observed in the real data. It occasionally indicated negative trends as positive trends, or it indicated lower magnitudes as high magnitudes. This indicates that the generators have trouble accurately replicating how the data changes over time, including the magnitude and direction of such changes.

E. Influence of the architecture of the synthetic data generator

The fidelity of the generated EHRs may also be influenced by the architecture of the data generators, particularly in the case of TimeGAN and EHR-M-GAN. The choice of architecture can significantly impact the generators' ability to capture temporal patterns present in real EHRs. For instance, GANs often face challenges in modelling mixed data types [7, 23] such as continuous and discrete variables. The real data contained both discrete and continuous values. This heterogeneity in the data can contribute to the observed discrepancies in fidelity.

Furthermore, using a latent space-based generator can effectively contribute to the fidelity of the generated EHRs. Latent spaces are a lower-dimensional representation of the original data that captures the underlying structure of the data. TimeGAN and EHR-M-GAN employed latent space representations to capture the temporal correlations in the real data to enhance the generation process. However, as noted by Fonseca and Bacao [24], defining the architecture of a model that learns the appropriate latent space representation is not an intuitive task. The method employed to derive latent representation and the structure or quality of the latent space representation in these models can potentially influence the quality of synthetic data and impact fidelity.

F. Criteria for choosing the optimal synthetic data generator

The selection of the optimal synthetic data generation method depends on the specific use case for which the synthetic EHRs were generated. The ideal data generator essentially captures the directionality, patient groups, predictability and temporal dynamics inherent in real EHRs.

For instance, consider a use case to study patients whose vital signs deteriorate over time. In this context, the optimal generator should generate synthetic EHRs that accurately reflect the temporal dynamics and directional changes of the real EHRs. Similarly, in the context of using synthetic EHRs for resource planning, which involves identifying patients who require prioritized care, it is crucial that the groupings of patients in the synthetic EHRs align with those in real EHRs. Alternatively, some studies [25] have explored assigning weights and using ranking mechanisms based on specific use cases to determine the optimal synthetic data generator.

However, generating synthetic EHRs that satisfy these criteria is challenging in the current state of synthetic EHR generation due to the inherent complexities of EHRs. Addressing these challenges requires continued research efforts in generating high-fidelity synthetic EHRs.

While this study offers insights into the fidelity of synthetic EHRs, some limitations exist. Firstly, the methods employed do not comprehensively evaluate the multivariate fidelity of synthetic EHRs. Additionally, the analysis focuses on structured EHRs.

VI. Conclusion and Future Directions

Synthetic EHRs offer a promising alternative to real EHRs, potentially alleviating the risks associated with using real EHRs. The practical utilization of synthetic EHRs rests on their ability to resemble the statistical and structural properties of real EHRs. Assessing the temporal fidelity of synthetic EHRs is crucial to ensure that they accurately resemble real EHRs in all aspects. In this paper, we employed time-series analysis methods to evaluate the fidelity of synthetic time-series EHRs. Our findings highlight the potential of these methods for assessing fidelity and reveal how the generated EHRs align with real EHRs concerning patient groups, temporal dynamics, predictability, and directional change. Future research in this domain will explore methodologies for selecting the optimal synthetic data generation methods. Additionally, an unexplored avenue exists in evaluating multivariate fidelity in synthetic time-series EHRs.

References

[1] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic data generation: State of the art in health care domain," *Computer Science Review*, vol. 48, p. 100546, 2023.

[2] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on EHR trajectories: A systematic review," *Journal of Biomedical Informatics*, vol. 144, p. 104430, Aug. 2023.

[3] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit Health*, vol. 2, p. e0000082, Jan. 2023.

[4] D. Bietsch, R. Stahlbock, and S. Voß, "Synthetic Data as a Proxy for Real-World Electronic Health Records in the Patient Length of Stay Prediction," *Sustainability*, vol. 15, p. 13690, Jan. 2023. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[5] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *npj Digit. Med.*, vol. 6, pp. 1–8, Oct. 2023. Publisher: Nature Publishing Group.

[6] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series Generative Adversarial Networks," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2021.

[7] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," *npj Digit. Med.*, vol. 6, pp. 1–18, May 2023. Number: 1 Publisher: Nature Publishing Group.

[8] L. Mosquera, K. El Emam, L. Ding, V. Sharma, X. H. Zhang, S. E. Kababji, C. Carvalho, B. Hamilton, D. Palfrey, L. Kong, B. Jiang, and D. T. Eurich, "A method for generating synthetic longitudinal health data," *BMC Medical Research Methodology*, vol. 23, p. 67, Mar. 2023.

[9] K. Bhanot, J. Pedersen, I. Guyon, and K. P. Bennett, "Investigating synthetic medical time-series resemblance," *Neurocomputing*, vol. 494, pp. 368–378, 2022.

[10] S. Dash, A. Yale, I. Guyon, and K. P. Bennett, "Medical Time-Series Data Generation Using Generative Adversarial Networks," in *Artificial Intelligence in Medicine* (M. Michalowski and R. Moskovitch, eds.), vol. 12299, pp. 382–391, Cham: Springer International Publishing, 2020. Series Title: Lecture Notes in Computer Science.

[11] R. Aggarwal and P. Ranganathan, "Common pitfalls in statistical analysis: The use of correlation techniques," *Perspect Clin Res*, vol. 7, no. 4, pp. 187–190, 2016.

[12] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, SIGMOD '05, (New York, NY, USA), pp. 491–502, Association for Computing Machinery, June 2005.

[13] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015.

[14] P. Patel, B. Sivaiah, and R. Patel, "Approaches for finding optimal number of clusters using k-means and agglomerative hierarchical clustering techniques," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, pp. 1–6, 2022.

[15] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, p. 31, Feb. 2021.

[16] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, pp. 243–256, Jan. 2013.

[17] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, p. 359–370, AAAI Press, 1994.

[18] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, pp. H2039–H2049, June 2000. Publisher: American Physiological Society.

[19] A. Delgado-Bonal and A. Marshak, "Approximate Entropy and Sample Entropy: A Comprehensive Tutorial," *Entropy*, vol. 21, p. 541, May 2019.

[20] A. Rae, "Trend Analysis," in *Encyclopedia of Quality of Life and Well-Being Research* (A. C. Michalos, ed.), pp. 6736–6736, Dordrecht: Springer Netherlands, 2014.

[21] J. W. Ely, J. D. Dawson, J. H. Lemke, and J. Rosenberg, "An introduction to time-trend analysis," *Infect Control Hosp Epidemiol*, vol. 18, pp. 267–274, Apr. 1997.

[22] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database," 2015. Version Number: 1.4 Type: dataset.

[23] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[24] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: a literature review," *Journal of Big Data*, vol. 10, p. 115, July 2023.

[25] C. Yan, Y. Yan, Z. Wan, Z. Zhang, L. Omberg, J. Guinney, S. D. Mooney, and B. A. Malin, "A multifaceted benchmarking of synthetic electronic health record generation models," *Nature Communications*, vol. 13, dec 2022.