Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications

Tingting Zhu (tingting.zhu@eng.ox.ac.uk)  
University of Oxford  https://orcid.org/0000-0002-1552-5630

Jin Li  
University of Oxford

Benjamin Cairns  
University of Oxford

Jingsong Li  
Zhejiang University

Article

Keywords:

Posted Date: October 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1258744/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications

Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu

Abstract—The recent availability of electronic health records (EHRs) have provided enormous opportunities to develop artificial intelligence (AI) algorithms. However, patient privacy has become a major concern that limits data sharing across hospital settings and subsequently hinders the advances in AI. Synthetic data, which benefits from the development and proliferation of generative models, has served as a promising substitute for real patient EHR data. However, the current generative models are limited as they only generate single type of clinical data for a synthetic patient, i.e., either continuous-valued or discrete-valued. In this paper, we propose a generative adversarial network (GAN) entitled EHR-M-GAN which simultaneously synthesizes mixed-type timeseries data. EHR-M-GAN is capable of capturing the multidimensional, heterogeneous, and correlated temporal dynamics in patient trajectories. We have validated EHR-M-GAN on three publicly-available intensive care unit databases with records from a total of 141,488 unique patients, and performed privacy risk evaluation of the proposed model. EHR-M-GAN has demonstrated its superiority over state-of-the-art benchmarks for synthesizing clinical timeseries with high fidelity, while addressing the limitations regarding data types and dimensionality in the current generative models. Notably, prediction models for outcomes of intensive care performed significantly better when training data was augmented with the addition of EHR-M-GAN-generated timeseries. EHR-M-GAN may have use in developing AI algorithms in resource-limited settings, lowering the barrier for data acquisition while preserving patient privacy.

The past decade has witnessed groundbreaking advancements made in computational health, owing to the explosion of medical data, such as electronic health records (EHRs) [1–3]. The secondary uses of EHRs give rise to research in a wide range of varieties, especially machine learning (ML)-based digital health solutions for improving the delivery of care [4–8]. However, in practice, the benefits of data-driven research are limited to healthcare organizations (HCOs) who possess the data [9, 10]. Due to concerns about patient privacy, HCO stakeholders are reluctant to share patient data [11–13]. Access to clinical data is often restricted, or can be prohibitively expensive to obtain, meaning that ML in biomedical research lags behind other areas in AI.

To accelerate the progress of developing AI methods in medicine, one promising alternative is for the data holder to create synthetic yet realistic data [14, 15]. By avoiding “one-to-one” mapping to the genuine data compared with data anonymization, synthetic data offers a solution to circumvent the issue of privacy, while the correlations in the original data distributions are preserved for downstream AI applications. There have been successes in the literature using synthetic data to improve AI models where otherwise not possible due to limited availability of resources [16–19]. For example, large-scale data sharing programs have been demanded for advancing studies related to COVID-19, such as in National COVID Cohort Collaborative (N3C) [20], and Clinical Practice Research Datalink (CPRD) database in the UK [21].

The availability of data synthesis facilitates the development of data-driven clinical models in research community for healthcare. However, some of the methods for generating synthetic patient EHRs rely heavily upon hand-crafted rules or clinical expertise [22, 23]. For example, McLachlan et al. [22] proposed to generate synthetic EHRs by formalizing the clinical practice guidelines into a state transition machine (STM), with the whole process being domain knowledge-intensive. In addition, since these methods can only handle data with low dimensionality (and are typically disease-specific), the general utility of these methods is severely limited. Recent advances in generative adversarial networks (GANs) [24] and their variants offer more efficacious means to generate multidimensional, high-fidelity data with complex correlations for many different applications [25–27]. Furthermore, owing to their flexible tuning techniques, GAN models can better generate samples of anomalous or sparse events, mitigating the issue of the rarity which potentially leads to bias during training downstream ML models [28, 29].

EHRs are sets of digital patient-centered records collected over time, thereby recording patient health status and care trajectories during the hospitalization. Recent research has aimed to use this information to generate clinical timeseries which capture the character of these trajectories [30–33]. In this study, we focus on generating timeseries in critical care, specifically in intensive care units (ICUs) where patient information are fully digitized. In the high-paced ICUs, various measurements and complex signals are recorded as mixed-type data for critically ill patients undergoing intensive monitoring [34, 35] (see Fig. 1 for data extraction). Typically, two types of data are observed: (1) the continuous-valued physiological signals, such as heart rate (HR), oxygen saturation (SpO2), and measurements from blood gas tests; (2) the discretized-valued medical intervention data that change over time, including the usage of therapeutic devices or intravenous medications. These mixed-type EHRs data can substantially capture a patient’s health status and care, therefore providing clinicians with a holistic perspective to make possibly more precise and complex analysis. These mixed-type clinical timeseries differ in dimensionalities and distributions, but are also highly correlated [36–38]. For example, the medications prescribed to patients rely heavily on measurements of their physiological status in the ICU. Meanwhile, the efficacy of the medical treatments, in
turn, can critically affect the patient’s physiological condition.

While mixed-type data can provide a rich environment for developing ML algorithms for assisting decision-making in critical care, they also impose complexity in creating generative models. When extending vanilla-GANs to mixed-type EHRs synthesis, major challenges are found in two aspects: (1) jointly modeling the underlying distribution of mixed-type data with heterogeneous nature, and (2) capturing the temporal correlations between them. As a result, most recent research has investigated how to synthesize the longitudinal data of different types separately, i.e., with a focus only on continuous or only on discrete timeseries [31, 39, 40] (please see Section S.1A and Fig. S1 in the Supplementary Materials for details). GANs have not previously been applied to address the issue in the mixed-data setting. Therefore for the first time, we propose a GAN for simultaneously synthesizing mixed-type longitudinal EHR data (denoted as EHR-M-GAN thereafter). Patient trajectories with high-dimensionality and heterogeneous data types (both continuous-valued and discrete-valued timeseries) are generated while the underlying temporal dependencies are captured. The main contributions of our work are as follows:

- A novel GAN model entitled EHR-M-GAN is proposed for simultaneously generating mixed-type multivariate EHR timeseries with high fidelity, and overcoming the challenges when extending GANs into the mixed-type data settings. First, to jointly model the underlying distributions of the heterogeneous features, EHR-M-GAN first maps data from different observational spaces into a reversible, lower-dimensional, shared latent space through a dual variational autoencoder (dual-VAE). Then, to capture the correlated temporal dynamics of the mixed-type
Table 1. Summary of the cohorts after preprocessing on three critical care databases. Number of patients and ICU admissions are provided for each dataset. Note that only the first ICU admission is selected for each patient. The dimension of the continuous- and discrete-valued data are provided. The conditional labels for training EHR-M-GAN\textsubscript{cond} and the corresponding counts for each class are also listed.

| Dataset  | Number of patients | Number of ICU admissions | Dimension of continuous-valued variables | Dimension of discrete-valued variables | Conditional labels | Counts |
|----------|--------------------|--------------------------|------------------------------------------|----------------------------------------|--------------------|--------|
| MIMIC-III | 28,344             | 28,344                   | 78                                       | 20                                    | ICU mortality      | 1,870  |
|          |                    |                          |                                          |                                        | Hospital mortality | 911    |
|          |                    |                          |                                          |                                        | 30-day readmission | 1,122  |
|          |                    |                          |                                          |                                        | No 30-day readmission | 24,441 |
| eICU     | 99,015             | 99,015                   | 55                                       | 19                                    | ICU mortality      | 4,589  |
|          |                    |                          |                                          |                                        | Hospital mortality | 3,291  |
|          |                    |                          |                                          |                                        | Hospital discharge | 91,224 |
| HiRID    | 14,129             | 14,129                   | 50                                       | 39                                    | ICU mortality      | 1,266  |
|          |                    |                          |                                          |                                        | ICU discharge      | 12,963 |

data, a sequentially coupled generator that is built upon a \textit{coupled recurrent network} (CRN) is employed. In addition, a conditional version of our model — EHR-M-GAN\textsubscript{cond} — is also implemented, which is capable of synthesizing condition-specific EHR patient data, such as those result in \textit{ICU mortality} or \textit{hospital readmission}. The code of our proposed work is publicly available on GitHub\textsuperscript{1}.  

- Evaluations are performed based on three publicly available ICU datasets: MIMIC-III\textsuperscript{[41]}, eICU\textsuperscript{[42]} and HiRID\textsuperscript{[43]} from a total of 141,488 patients. Standardized preprocessing pipelines are applied for the three ICU datasets to provide generalizable machine learning benchmarks. The code for the end-to-end preprocessing pipelines is also available on GitHub\textsuperscript{2}.

- Our EHR-M-GAN outperforms the state-of-the-art benchmarks on a diverse spectrum of evaluation metrics. When compared to real EHR data, both qualitative and quantitative metrics are used to assess the representativeness of the mixed-type data and their inter-dependencies. We further demonstrate the advantages offered by EHR-M-GAN in augmenting clinical timeseries for downstream tasks under various clinical scenarios.

- In the evaluation of privacy risks, we perform an empirical analysis on EHR-M-GAN based on membership inference attack\textsuperscript{[44]}. We then further evaluate the performance of EHR-M-GAN under the framework of differential privacy for its application in downstream task\textsuperscript{[45]}.

**Data and Evaluation**

**Dataset Description.** The following three publicly accessible ICU datasets are used for evaluating the performance of EHR-M-GAN in generating the longitudinal EHR data:

- **MIMIC-III** (Medical Information Mart for Intensive Care)\textsuperscript{[41]} — a freely accessible database that comprises de-identified EHRs associated with approximately 60,000 ICU admitted patients and 312 million observations to Beth Israel Deaconess Medical Center.

- **eICU** (eICU Collaborative Research Database)\textsuperscript{[42]} — a multi-center critical care database containing data for over 200,000 admissions and 827 million observations to ICUs from 208 hospitals located throughout the United States.

- **HiRID** (High time-resolution ICU dataset)\textsuperscript{[43]} — a high-resolution ICU dataset relating to more than 3 billion observations from almost 34,000 ICU patient admissions, monitored at the Department of Intensive Care Medicine, Bern University Hospital, Switzerland.

All these critical care databases include vital sign measurements, laboratory tests, treatment information, survival records, and other routinely collected data from hospital EHR systems. From these clinical observations, we feature the patient trajectories as the combination of continuous-valued physiological timeseries and discrete-valued medical intervention timeseries. Data are preprocessed following an open-source framework — MIMIC-Extract\textsuperscript{[46]}. Details on data curation, including the cohort selection criteria, full list of features, and imputation method, are explained in Supplementary Materials (see S.2 Datasets). Overall, the summarizing statistics of the finalised cohorts for three databases are shown in Table 1.

**Baseline models.** We compare the performance of EHR-M-GAN with eight state-of-the-art generative methods in literature. However, as these benchmarks can only synthesize single-type EHRs, we therefore compare using the corresponding partial component of our synthetic results, i.e., either the continuous-valued part or the discrete-valued part. For continuous-valued timeseries generation, benchmark GAN models include C-RNN-GAN\textsuperscript{[47]}, R(C)GAN\textsuperscript{[30]} and TimeGAN\textsuperscript{[39]}. For discrete-valued timeseries generation, classic medGAN\textsuperscript{[48]}, seqGAN\textsuperscript{[40]}, and two recently proposed work — synTEG\textsuperscript{[32]} and DualVAE\textsuperscript{[31]} are used for comparison. Apart from these GAN-based models, we also incorporate PrivBayes\textsuperscript{[49]} to synthesize discrete-valued timeseries, which falls in the class of non-GAN generative approaches using a Bayesian framework\textsuperscript{[18]}. As the original paper of PrivBayes focuses on data anonymization using differential privacy, we therefore implemented its ‘Non-Private’ version for a fair comparison with other baselines (see Section 4.1 Non-Private Methods in\textsuperscript{[49]}). For medGAN and PrivBayes, we feed the flattened temporal sequence as the input since the models cannot produce timeseries data.

Also, an \textit{ablation study} is performed to assess the effectiveness of our network architecture and its distinct modules. As EHR-M-GAN learns the joint representations from heterogeneous types of data using separate networks (while modeling their correlations using \textit{dual-VAE} and \textit{sequentially coupled generator}), we first compare it with a variant that allows the joint modeling in a unified network using a single VAE (denoted as GAN\textsubscript{Unified}). Then, two more model variants are considered, each using a main component of
Fig. 2. The network architectures in the ablation study. Three variants of EHR-M-GAN are implemented in the ablation study. Compared with the full model of EHR-M-GAN, \( \text{GAN}_{\text{unified}} \) learns the joint representations of heterogeneous types of data in a unified network; \( \text{GAN}_{\text{VAE}} \) maintains the basic architecture of EHR-M-GAN, but ignores the dependency learning (i.e., separate networks for two streams of inputs are trained in parallel); \( \text{GAN}_{\text{EL}} \) constructs the shared latent space using the dual-VAE module but omit the sequentially coupled generator for learning the temporal correlations in the mixed-type timeseries.

**Evaluation metrics.** Evaluating GAN models is a notoriously challenging task. Advantages and pitfalls of commonly used evaluation metrics for GANs are discussed in [51]. In this work, we first individually assess the representativeness of the synthetic continuous-valued and discrete-valued timeseries. This includes measuring the distance between underlying data distributions (such as Maximum mean discrepancy and Dimension-wise probability), comparing the indistinguishability of the synthetic data to the true data (i.e., Discriminative score). Secondly, we assess our model by using a set of qualitative and quantitative metrics (such as Embedding visualisation, Patient trajectory plot, Pearson pairwise correlations, and Autocorrelation function) for evaluating to which extent our model can reconstruct the interdependency between two types of data. Thirdly, we introduce data augmentation by incorporating synthetized EHR timeseries under various settings, and quantitatively assess the improvement provided by EHR-M-GAN in the **Downstream tasks** for medical intervention prediction in the ICU. Lastly, we measure the attribute of patient privacy-preserving of EHR-M-GAN under **Membership inference attack**. We also evaluate the performance of the same downstream tasks under **Differential privacy** guarantees (See Fig. 1 for the evaluation pipeline).

**Results**

**Maximum mean discrepancy.** To measure the similarity between the continuous-valued synthetic data and the real data,
Table 2. Maximum mean discrepancy (MMD) of continuous-valued synthetic data. Lower values of MMD indicate models which can better learn the distribution of the real data.

|                  | C-RNN-GAN | R(C)GAN | TimeGAN | GAN\text{cond} | GAN\text{VAE} | GAN\text{AE} | EHR-M-GAN | EHR-M-GAN\text{cond} |
|------------------|-----------|---------|---------|---------------|--------------|--------------|------------|----------------------|
| MIMIC-III        | 1.038 ± 0.013 | 0.971 ± 0.029 | 0.694 ± 0.025 | 0.893 ± 0.027 | 0.926 ± 0.038 | 0.745 ± 0.040 | 0.692 ± 0.034 | 0.804 ± 0.027 |
| eICU             | 1.139 ± 0.023 | 1.106 ± 0.042 | 0.672 ± 0.038 | 0.850 ± 0.032 | 0.842 ± 0.029 | 0.670 ± 0.034 | 0.651 ± 0.026 | 0.540 ± 0.018 |
| HRID             | 0.982 ± 0.017 | 0.865 ± 0.020 | 0.470 ± 0.024 | 0.518 ± 0.030 | 0.532 ± 0.035 | 0.508 ± 0.028 | 0.428 ± 0.015 | 0.389 ± 0.024 |

maximum mean discrepancy (MMD) is used. MMD can assess whether two samples, in our case, true data x and synthetic data x′, are from the same distributions. It can be expressed as the first-order moments, i.e., mean embeddings, of the two samples in a reproducing kernel Hilbert space (RKHS): 

\[ k(x, x') = \sum \text{exp} \left( \frac{|x - x'|^2}{\sigma^2} \right). \]

We follow the implementations in [52] by using a sum of Gaussian kernel sets.

As indicated in Table 2, EHR-M-GAN outperforms the state-of-the-art benchmarks among all three datasets in synthesizing continuous-valued timeseries. The conditional version — EHR-M-GAN\text{cond} further boosts the performance of the model by leveraging the information of the condition-specific inputs. Furthermore, as shown in the ablation study, EHR-M-GAN and EHR-M-GAN\text{cond} produce smaller MMD values when compared to their variants. Using MIMIC-III as an example, compared with the basic model GAN\text{VAE}, by integrating the shared latent space learning using dual-VAE under multiple loss constraints, the performance of GAN\text{AE} significantly improves (GAN\text{AE} vs. GAN\text{VAE}, 0.745 to 0.926, p-value < 0.05 from t-test\(^3\)). By further building the sequentially coupled generator based on BLSTMs and exploiting the information within mixed-type data, the MMD of EHR-M-GAN shows a nearly 24% improvement over GAN\text{VAE}. When synthesizing mixed-type timeseries using the unified network, the performance of GAN\text{cond} for generating continuous-valued timeseries lags behind the proposed EHR-M-GAN. It can therefore be inferred that, compared with EHR-M-GAN which extracts useful hierarchical representations for each data type using tailored encoding layers, it is quite challenging for GAN\text{cond} to learn marginal distributions from raw mixed-type timeseries with a unified architecture.

Dimension-wise probability. To evaluate the representativeness of the synthetic discrete-valued timeseries, the dimension-wise probability test is employed. As a sanity check, it investigates if the distribution across the spatial and temporal dimensions is matched between the real and synthetic data. Therefore, the Bernoulli success probability \( p \in [0, 1] \) is calculated over all feature dimensions for the discrete-valued timeseries, and is visualized through scatterplot. The correlation coefficients (CCs) and root-mean-square errors (RMSEs) are also adopted [53] based on the Bernoulli success probabilities to quantitatively measure the distribution divergence between real and synthetic data.

As shown in Fig. 3 (see Fig. S4 and S5 for more results on eICU and HRID datasets), the optimal results are provided by EHR-M-GAN and EHR-M-GAN\text{cond}. The close-to-real probability distributions that appear along the diagonal line indicate the remarkable similarity between the real data and the synthetic data provided by our models. The quantified CC and RMSE also correspond with the visualisation results, which are close to the highest mark (EHR-M-GAN: RMSE ≈ 0.0095, CC = 0.9973). Similar to the results in MMD, the dimension-wise distributions are better captured when modules such as dual-VAE and sequentially coupled generator are introduced in EHR-M-GAN. GAN\text{cond} suffers from mode collapse (the generator fails to produce outputs with sufficient diversity), and therefore shows poor performance compared with other variants when synthesizing discrete-valued timeseries. As the mixed-type features are treated as unimodal input without differentiating their heterogeneous nature, no marginal representations are explicitly learned.

Among all state-of-the-art benchmark models, DualAAE shows the best result but is slightly sub-optimal when compared to EHR-M-GAN. In contrast, both skewed distribution and low performance scores are observed in medGAN, as it lacks the ability to capture the temporal correlations within timeseries. SynTEG shows improved performance over medGAN, as it is capable of synthesizing discrete-valued features in EHRs with timestamps. The non-GAN generative method PrivBayes also shows good performance among all the benchmark synthesizers when modeling the underlying probability distribution of the discrete-valued EHR timeseries. On the other hand, despite the well-known performance of SeqGAN in natural language generation, it is not quite applicable in synthesizing sequential clinical EHRs.

Generating discrete-valued features are known to be problematic for traditional GANs. Due to their limitation in passing the gradients from the critic models, vanilla GANs cannot update their generators efficiently based on the adversarial loss [40, 48]. However, the result of EHR-M-GAN shows its superiority in explicitly capturing each dimension of the discrete-valued sequences. EHR-M-GAN mitigates this problem by learning the shared latent representations using dual-VAE. Discrete-valued timeseries are encoded into a gradient-differentiable space for further optimizing the generators and thus solving the problem.

Discriminative score. For both continuous-valued and discrete-valued data, the discriminative score is measured as the accuracy of a discriminator trained post-hoc to separate real from generated samples. Synthetic data are generated with the same amount of the hold-out test set from the original data, and are labeled as synthetic and real correspondingly to train the binary classifier. We opt to implement the classifier as a single-layered Bi-directional Long Short-Term Memory (Bi-LSTM) model (i.e., many-to-one), to characterize the temporal correlations within the patient EHR timeseries in this supervised task.

Results in Table 3 indicate that synthetic data that are highly indistinguishable from the original data are produced by EHR-M-GAN and EHR-M-GAN\text{cond}. Especially for EHR-M-GAN\text{cond}, it achieves the optimal discriminative scores consistently against other benchmarks for both continuous-valued and discrete-valued timeseries. For discrete-valued data generation, EHR-M-GAN-generated samples achieve the dis-

\(^3\)Unpaired (two-sample) t-test with a significance level of 0.05 is used throughout the paper unless specified otherwise.
The figure shows scatter plots of the dimension-wise probability test on MIMIC-III dataset. The x-axis and y-axis represent the probability distribution for the real data and synthetic data with the same sample size, respectively. Each dot represents a physiological measurement or treatment status at a particular time in the patient EHR data. Same color indicates the same measurement or status (but with varying timestamps). The optimal performance appears along the diagonal line. The corresponding CCs (0, 1) are calculated to quantify the probability distribution similarities between the real and synthetic EHRs. The dimension-wise probability plot for eICU and HiRID can be found in Supplementary materials (see S3A).

Table 3. Discriminative score of synthetic data. A discriminative model is trained post-hoc to discriminate between synthetic samples and real samples. The accuracy from the discriminative classifier is used as the discriminative score, where the lower value indicates better performance. The result is bounded by 0.5 when the classifier cannot distinguish between two distributions.

| Method                     | MIMIC-III | eICU | HiRID |
|----------------------------|-----------|------|-------|
| Continuous-valued          |           |      |       |
| synthetic data             |           |      |       |
| C-RNN-GAN                  | 0.825 ± 0.013 | 0.876 ± 0.019 | 0.74 ± 0.022 |
| R(Cond) GAN                | 0.833 ± 0.029 | 0.870 ± 0.021 | 0.724 ± 0.016 |
| TimeGAN                    | 0.763 ± 0.018 | 0.790 ± 0.013 | 0.716 ± 0.024 |
| GAN_{dist}                 | 0.809 ± 0.023 | 0.861 ± 0.027 | 0.749 ± 0.014 |
| GAN_{AE}                   | 0.842 ± 0.020 | 0.873 ± 0.014 | 0.802 ± 0.017 |
| GAN_{EL}                   | 0.796 ± 0.016 | 0.813 ± 0.021 | 0.752 ± 0.021 |
| EHR-M-GAN                  | 0.746 ± 0.018 | 0.776 ± 0.015 | 0.724 ± 0.015 |
| EHR-M-GAN_{cond}           | 0.729 ± 0.025 | 0.745 ± 0.017 | 0.693 ± 0.012 |
| Discrete-valued synthetic data |         |      |       |
| synthetic data             |           |      |       |
| medGAN                     | 0.903 ± 0.027 | 0.915 ± 0.034 | 0.896 ± 0.021 |
| seqGAN                     | 0.917 ± 0.025 | 0.924 ± 0.021 | 0.913 ± 0.027 |
| SynTEG                     | 0.879 ± 0.021 | 0.902 ± 0.030 | 0.878 ± 0.025 |
| DualAAE                    | 0.847 ± 0.029 | 0.860 ± 0.033 | 0.829 ± 0.024 |
| PrivBayes                  | 0.859 ± 0.036 | 0.863 ± 0.034 | 0.832 ± 0.017 |
| GAN_{dist}                 | 0.890 ± 0.022 | 0.907 ± 0.026 | 0.849 ± 0.015 |
| GAN_{AE}                   | 0.862 ± 0.024 | 0.881 ± 0.029 | 0.824 ± 0.018 |
| GAN_{EL}                   | 0.829 ± 0.032 | 0.844 ± 0.028 | 0.816 ± 0.025 |
| EHR-M-GAN                  | 0.813 ± 0.026 | 0.831 ± 0.024 | 0.802 ± 0.020 |
| EHR-M-GAN_{cond}           | 0.784 ± 0.024 | 0.803 ± 0.022 | 0.779 ± 0.019 |

(i.e., GAN_{EL}) have shown remarkable success as making the synthetic data more realistic than separately generating the latent embeddings based on VAEs (as in GAN_{AE}). The sequentially coupled generator further improves the model by capturing dynamics between mixed-type data and iterating over time, therefore enabling the synthetic timeseries to become more indistinguishable from the original. The proposed approach with GAN_{dist} that models the mixed-type data in a unified network, our proposed model enables effective learning for the marginal distributions and various types of data. More importantly, EHR-M-GAN can leverage its dependency learning components to explicitly capture the correlations between different variables.

**Interdependency characteristics.** In this section, we first employ Pearson pairwise correlation plot to qualitatively evaluate how closely the synthetic data can model the inter- and intra-correlations between continuous-valued and discrete-valued timeseries. And then, autocorrelation functions (ACF) and the corresponding root-mean-square errors (RMSEs) are calculated to show how EHR-M-GAN can capture the temporal correlations among the timeseries.

In Pearson pairwise correlation plot, the positive and negative correlation among each feature pair is measured by Pearson correlation coefficient (PCC), which ranges from -1 to 1. Five commonly measured vital sign and laboratory measurement features — *Oxygen Saturation, Systolic Blood Pressure, Respiratory Rate, Heart Rate, Temperature*, as well as two medical intervention features — *Mechanical Ventilation and Vasopressor* are considered and compared as an exemplar in Fig. 4. Timestamps are extracted with every 3 hours interval in a total of 24 hours ICU stay, to explore the temporal dependencies within different variables. As observed, correlation trends over distinct features are closely reflected by the synthetic data. For example, negative correlations are shown between *Vasopressor and Systolic Blood Pressure* in all three datasets. Similar correlations are also well-captured in the synthetic data provided by EHR-M-GAN. Such correlations can be interpreted as the time-delay effect of prescribing the
Fig. 4. Pearson pairwise correlation (PPC) between continuous-valued and discrete-valued timeseries. The plots contrast the PPC calculated within the real data (left column) and the synthetic data generated by EHR-M-GAN (right column). These visualisations indicate how well correlations in the synthetic data resemble the correlations observed in the real patient trajectories. As shown in this figure, SpO2, SBP, RR, HR, Temp represents Oxygen Saturation, Systolic Blood Pressure, Respiratory Rate, Heart Rate, Temperature, respectively. And Vent. and Vaso. corresponds to Vasopressor and Mechanical Ventilation. PPC is calculated every 3 hours over the total 24 hours of ICU stay (ticks of the timestamps are omitted).

treatment of vasopressors to elevate mean arterial pressure (MAP) once noticing patients’ abnormal blood pressure.

In addition, temporal dependencies revealed in the original EHR data are also preserved in the synthetic timeseries. For example, synchronized correlations across timestamps are observed between Respiratory Rate and Heart Rate in MIMIC-III dataset. This can be explained by the common regulation of these two features by the autonomic nervous system and their synchronized increase in cases of physiological stress, such as hypoxemia. Similar trends in the synthetic data suggest that EHR-M-GAN can successfully recover temporal dependencies with a high granularity from real patient trajectories. It is also worth noticing that these three ICU datasets exhibit similar yet not identical correlations over different feature pairs. This demonstrates the data heterogeneity originating from the distribution discrepancy of the underlying patient populations.

Autocorrelation measures the relationship between the timeseries and its lagged version. Fig. S6 - S8 in the Supplementary materials shows the ACF calculated for selected continuous-valued and discrete-valued variables (same as Pearson pairwise plot) on real and synthetic timeseries. The time lags are specified as the hourly intervals up to 24 hours before patients’ ICU endpoints (ICU discharge or death). Additionally, RMSEs are calculated to quantitatively evaluate the similarity between the corresponding two curves produced by real data and synthetic data.

Similar patterns are presented between the ACF calculated for real data and their synthetic counterparts, while the quantitative statistics also correspond with the observation. Moreover, overlapping confidence intervals indicate that the synthetic data is able to consistently capture the underlying temporal distributions within the real timeseries. The 24-hour timeseries data did not exhibit significant seasonal or periodic variation, mainly because these are short-timescale trajectories extracted from EHR data in the critical care setting. For variables such as Heart Rate, Oxygen Saturation, and Systolic Blood Pressure, the positive ACF coefficients rapidly decrease within the period of first few hours, followed by the growing trends of negative temporal correlation. The lag with the lowest correlation coefficient is identified at approximately 4 hours. Specifically, global peaks appear roughly at the 12-hour ticks of Temperature for both real and synthetic data on three critical databases. Meanwhile, the negative correlation strengthens as the time lag increase for Mechanical ventilation in the original timeseries. Since these behaviours can be reproduced by EHR-M-GAN, therefore they demonstrate that our model can effectively capture the temporal characteristics in the original timeseries.

Patient trajectory visualisation. Sample trajectories generated by EHR-M-GAN and EHR-M-GANcond are shown in Fig. 5 for MIMIC-III dataset (see Fig. S12, S13, S14 in Supplementary Materials for more results). By visual inspection, we can see that the proposed models can capture patterns within substantially different time-varying signals. Temporal dynamics in both rapidly fluctuating vital signs (e.g., Oxygen Saturation) and infrequently changing intervention (e.g., Mechanical Ventilation) can be well-preserved. The variability of the synthetic signals suggests that rather than just intelligently “memorizing” the training data, EHR-M-GAN is capable of modelling the underlying true data distributions in order to produce genuine samples.

Furthermore, EHR-M-GANcond shows even more superior performance as it can generate trajectories with predefined patient conditions. From the instance generated by EHR-M-GANcond under the conditional information of ICU mortality in Fig. 5, it is possible to observe the variations in physiological signals which suggest that the patient suffers from severe deterioration towards the end of the clinical endpoint. Meanwhile, temporal dependencies are observed within different physiological signals along with medical intervention. Synchronized correlations are shown between Oxygen Saturation, Systolic Blood Pressure, and Heart Rate from approximately 18 hr to 21 hr, followed by the presence of Mechanical Ventilation for providing the respiratory support for the patient at 22 hr.

Embedding visualisation. Conditional GAN helps to generate patient data with a set of specific medical conditions [50], therefore improving the diversity of the data when synthesizing clinical timeseries. We apply t-SNE to qualitatively visualise the latent embedding generated by EHR-M-GANcond on three critical care databases. The distributions of the latent embeddings are induced by the encoders in the dual-VAE during learning the shared latent space representations (See
Methods section, p11, for details).

As shown in Fig. S9 - S11, the result indicates that raw EHR timeseries do not have distinctly different modes in the lower dimensional space. Moreover, clusters can only be partially identified from the embedding results of EHR-M-GAN, which means that limited conditional information can be carried into the generative model. On the other hand, good separability is shown in the latent representations recovered from EHR-M-GAN\textsubscript{cond}. To learn contextual information from the patient trajectories, EHR-M-GAN\textsubscript{cond} leverages the conditional labels to build encoders and decoders in dual-VAE, and meanwhile places the semantic loss constraint to learn the shared latent representations. It can be inferred that the conditional extension of the proposed model can further yield benefits by synthesizing condition-specific EHR timeseries with respect to distinctive patient health status.

**Downstream tasks.** As previously discussed, one of the most prominent goals for GANs is to benefit the future downstream analyses in the real clinical application. A relevant question in the ICU is whether specialized medical treatments, such as therapeutic interventions or organ support, are required for critically ill patients during the admission. Accurate predictions on such tasks can help clinicians to provide actionable, in-time interventions in the resource-intensive ICU. Therefore in this section, **clinical intervention prediction** tasks are implemented to evaluate the potential of EHR-M-GAN and EHR-M-GAN\textsubscript{cond} in synthesizing high-fidelity synthetic data to further boost the performance of ML classifiers. In line with prior work [46, 55, 56], we establish LSTM-based classifiers to predict the status of mechanical ventilation and vasopressors using continuous-valued multivariate physiological signals as the predictors. A fixed duration of 12 hours is used for both observation window and prediction window (see Fig. 1). Four outcomes of medical intervention status are defined as: Stay on, Onset, Switch off, Stay off (detailed descriptions can be found in Fig. 1).

We partition the dataset as illustrated in Figure 6a, and the performances are assessed from two aspects (see Figure 6b): (i) **Traditional approach:** To explore whether the synthetic data can represent the real data accurately, we compare Train on Real, Test on Real (TRTR) with Train on Synthetic, Test on Real (TSRTR), to show whether the performance of a classifier trained on synthetic data from EHR-M-GAN or EHR-M-GAN\textsubscript{cond} can be generalized to real data. In addition to the proposed models, synthetic data produced by the baseline models are also used to train the downstream classifiers for comparison. Other than a measurement of data utility where the downstream task is to predict discrete-valued medical intervention (described as outcomes in this scenario) using continuous-valued physiological features (denoted as predictors), TSRTR can also be used to assess data synthesizers’ ability to capture the interdependencies between the mixed-type features. (ii) **Data augmentation approach:** As data augmentation is employed as a means of circumventing the issue caused by the under-resourced EHR data, here we explore whether synthetic data can used to improve the existing ML algorithms through data augmentation. Therefore, **Train on Synthetic and Real, Test on Real (TSRTR)** is compared with **TRTR** to measure the improvement of the classifier’s performance when trained on the augmented data [30, 33].

The augmentation ratio \( \alpha \) or \( \beta \) is applied on sub-train data \( D' \), or synthetic data \( B \), in two different scenarios of TSRTR, respectively. Details are explained as follows (also see Figure 6b for illustration).

Firstly, as the dearth of data potentially degrades the performance of downstream classifiers, given that the real data has a limited and fixed sample size, we investigate whether adding synthetic EHR data provided by EHR-M-GAN and EHR-M-GAN\textsubscript{cond} can improve the training of downstream classifiers. **Ratio \( \alpha \)** indicates the portion of synthetic data (see Figure 6b) being used to augment the real data to improve the quality and robustness of the downstream classifiers. \( \alpha \) is set to be 10%, 25%, and 50%, representing the availability of synthetic samples provided for augmentation.

Secondly, the acquisition of healthcare data is generally
time-consuming and expensive, therefore another overarching goal for the generative model is to minimize the efforts in collecting data. In this section, we investigate whether high-fidelity synthetic data can offer a viable solution for boosting the downstream classifiers’ performance when the availability of real data is limited. This allows us to understand if the sample size required for real data collection can be reduced while maintaining sufficient predictive power through the use of synthetic data. During the experiment, the synthetic data $B$ is given (to emulate the scenario where synthetic datasets are available for a particular clinical research purpose), which further is combined with limited real data (collected during clinical trial), to train the downstream classifiers (i.e., augment synthetic data with limited real data). Then by implementing EHR-M-GAN or EHR-M-GAN$_{cond}$ in TSRTR, we investigate the proportion of the real data $A'_{Tr}$ (ratio $\beta$) required to maintain the same performance as in TRTR based on the entire synthetic dataset $B$ (see Figure 6b).

**Fig. 6. Downstream intervention prediction experimental setup. a. Data splitting.**

During training stage, the real data is split into two sets with 70% training data $A$ and 30% test data $A'$. The test data $A'$ is further split into sub-train data $A'_{Tr}$, and sub-test data $A'_{Ts}$ with equal size. Then, the synthetic data $B$, with size equal to the sub-train data $A'_{Tr}$, is synthesized by EHR-M-GAN (or EHR-M-GAN$_{cond}$) trained on the real training data $A$. **b. Data augmentation scenarios.** Subsequent experiments are trained on set $A'_{Tr}$, or $B$, or $A'_{Tr}$ $\cup$ $B$ and then tested on $A'_{Ts}$. In traditional approach, results based on Train on Real, Test on Real (TRTR) and Train on Synthetic, Test on Real (TSRTR) are compared to assess the generalisability of the synthetic data. In data augmentation approach, i.e., Train on Synthetic and Real, Test on Real (TSRTR), we either augment real data $A'_{Tr}$ with $\alpha$ (augmentation ratio, 0 to 50%) of the synthetic samples $B$, or augment synthetic samples $B$ with $\beta$ (0 to 50%) of the real data $A'_{Tr}$.

**Traditional approach.** Table 4 compares the classification performances of predicting forthcoming medical interventions in the ICUs under the experimental setting of TRTR and TSRTR. It is expected that the optimal AUROCs are achieved by the classifiers that are trained on real data. In comparison, the classifiers trained on the synthetic data provided by proposed models can achieve similar performances. More specifically, synthetic data generated by EHR-M-GAN$_{cond}$ demonstrates better generalisability when compared with EHR-M-GAN in the downstream application, such as the task of predicting mechanical ventilation on the HiRID dataset (TRTR vs. TSRTR from EHR-M-GAN$_{cond}$: 0.867 to 0.856, with p=0.3906).

Compared with the baseline models, the proposed EHR-M-GAN shows improved performance in TSRTR, as it can model the distribution of mixed-type EHRs more accurately, while preserving the temporal correlations in the heterogeneous time series through the dependency learning components. The results indicate that interdependency between the mixed-type EHRs is weakly captured by GAN$_{AE}$, as the two streams of inputs are trained in parallel and separately. GAN$_{Gated}$ attempts to capture the temporal correlations of mixed-type EHRs through jointly modeling their underlying distribution in a unified network. However, its unified architecture limits the model’s capacity to learn the marginal distribution of each data type, the resulted quality of the synthetic EHRs is impaired and so is its performance in TSRTR.

**Data augmentation approach (with ratio $\alpha$).** The results in Table 5a demonstrate that classifiers boosted by EHR-M-GAN can consistently outperform TRTR (see Table 4) at the augmentation ratio of 50%. In comparison, only 25% of augmentation ratio is needed to achieve improved results for EHR-M-GAN$_{cond}$. For example, the classifier trained on MIMIC-III to predict the status of Vasopressor with augmentation ratio $\alpha$ set as 50%, significantly increase the AUROC by 6% when compared to the classifier trained using only the real data (EHR-M-GAN$_{cond}$ vs. TRTR: 0.896 to 0.841, p < 0.05). Our experiment results have demonstrated that the proposed models can be used for data augmentation to overcome the issue of data scarcity and subsequently improve the classifiers’ performance.

**Data augmentation approach (with ratio $\beta$).** On the other hand, as shown in Table 5b, by augmenting with the synthetic data provided by EHR-M-GAN, only approximately 50% of the real data is required to keep the classification AUROCs on par with, or even significantly better than fully exploiting the real data under TRTR. For EHR-M-GAN$_{cond}$, the ratio needed for real data to maintain the comparable predictive power is further reduced to 25%, which equivalently indicates a 75% reduction of sample size required in real data collection. Overall, results presented in Table 5b demonstrate that by exploiting only a limited ratio of the real data, EHR-M-GAN and EHR-M-GAN$_{cond}$ can robustly maintain the level of prediction performance, therefore alleviating the necessity for acquiring clinical data at scale.

**Privacy risk evaluation.** Patient privacy is a major concern with regards to sharing electronic health records in any means. Although generative models overcome the explicit one-to-one mapping towards the underlying original data (in contrast to data anonymisation), GAN could potentially raise privacy concerns of information leakage if they simply “memorise” the training data. In that case, sensitive medical information (e.g. national insurance number) belonging to a specific patient used in training GANs can be retrieved during the generative stage, thus posing challenges for preserving privacy in downstream applications.

In this section, we first quantify the vulnerability of EHR-M-GAN to adversary’s membership inference attacks, also known as presence disclosure [57, 58]. The threat model is
Table 4. Downstream task evaluation. Downstream tasks are evaluated under the training scenarios of Train on Real, Test on Real (TRTR) and Train on Synthetic, Test on Real (TSRTR). Prediction of two outcomes of interest—intervention by Mechanical ventilation (Vent.) and Vasopressors (Vaso.) are selected as exemplary tasks. Macro-AUROC is used to score the performance of the LSTM-based classifiers on the multi-class prediction tasks (labeled as Stay on, Onset, Switch off, Stay off).

| Dataset  | Treatments | Real data | $GAN_{	ext{train}}$ | $GAN_{	ext{val}}$ | $GAN_{	ext{test}}$ | EHR-M-GAN | EHR-M-GAN$_{\text{test}}$ |
|----------|------------|-----------|----------------------|------------------|------------------|-----------|---------------------|
| MIMIC-III| Vent.      | 0.894 ± 0.016 | 0.724 ± 0.015 | 0.701 ± 0.018 | 0.728 ± 0.010 | 0.740 ± 0.009 | 0.823 ± 0.020 |
|          | Vaso.      | 0.841 ± 0.009 | 0.694 ± 0.012 | 0.651 ± 0.015 | 0.679 ± 0.009 | 0.725 ± 0.015 | 0.810 ± 0.019 |
| eICU     | Vent.      | 0.666 ± 0.015 | 0.697 ± 0.014 | 0.702 ± 0.009 | 0.718 ± 0.012 | 0.745 ± 0.008 | 0.795 ± 0.015 |
|          | Vaso.      | 0.813 ± 0.018 | 0.648 ± 0.011 | 0.657 ± 0.012 | 0.665 ± 0.014 | 0.706 ± 0.014 | 0.748 ± 0.017 |
| HiRID    | Vent.      | 0.867 ± 0.012 | 0.765 ± 0.014 | 0.747 ± 0.013 | 0.803 ± 0.008 | 0.825 ± 0.019 | 0.856 ± 0.033 |
|          | Vaso.      | 0.978 ± 0.010 | 0.754 ± 0.018 | 0.752 ± 0.020 | 0.779 ± 0.013 | 0.814 ± 0.015 | 0.844 ± 0.018 |

Table 5. Downstream task evaluation with data augmentation. Downstream tasks are evaluated under the training scenarios of Train on Synthetic and Real, Test on Real (TSRTR). The predictive tasks and evaluation metrics are in accordance with Table 4. The upper (↑) indicates that the AUROC value under TSRTR is higher than TRTR in Table 4 for the corresponding task, while the bold arrow (†) indicates that the value is significantly improved using t-test ($p<0.05$).

(a) Performance of the downstream LSTM-based classifier under TSRTR with data augmentation ratio $\alpha$. All data from sub-train data $A_{\text{Tr}}$ concatenated with $\alpha$ of the synthetic data $B$ (augmentation ratio $\alpha = 10\%$, $25\%$ or $50\%$) is used as the training set.

| Dataset  | Treatments | $\alpha = 10\%$ | $\alpha = 25\%$ | $\alpha = 50\%$ |
|----------|------------|------------------|-----------------|------------------|
| MIMIC-III| Vent.      | 0.828 ± 0.013 | 0.877 ± 0.014 | 0.912 ± 0.015 (↑) |
|          | Vaso.      | 0.816 ± 0.015 | 0.834 ± 0.023 | 0.859 ± 0.014 (↑) |
| eICU     | Vent.      | 0.858 ± 0.008 | 0.862 ± 0.012 | 0.873 ± 0.014 (↑) |
|          | Vaso.      | 0.798 ± 0.015 | 0.805 ± 0.020 | 0.821 ± 0.028 (↑) |
| HiRID    | Vent.      | 0.871 ± 0.025 (↑) | 0.882 ± 0.021 (↑) | 0.913 ± 0.019 (↑) |
|          | Vaso.      | 0.850 ± 0.016 | 0.874 ± 0.022 | 0.894 ± 0.018 (↑) |

(b) Performance of the downstream LSTM-based classifier under TSRTR with data augmentation ratio $\beta$. All data from synthetic data $B$ concatenated with $\beta$ of the sub-train data $A_{\text{Tr}}$ (augmentation ratio $\beta$ = $10\%$, $25\%$ or $50\%$) is used as the training set.

| Dataset  | Treatments | $\beta = 10\%$ | $\beta = 25\%$ | $\beta = 50\%$ |
|----------|------------|-----------------|-----------------|-----------------|
| MIMIC-III| Vent.      | 0.777 ± 0.016 | 0.824 ± 0.010 | 0.885 ± 0.009 |
|          | Vaso.      | 0.786 ± 0.019 | 0.810 ± 0.020 | 0.849 ± 0.017 (↑) |
| eICU     | Vent.      | 0.761 ± 0.011 | 0.822 ± 0.012 | 0.870 ± 0.019 (↑) |
|          | Vaso.      | 0.742 ± 0.014 | 0.797 ± 0.013 | 0.846 ± 0.018 (↑) |
| HiRID    | Vent.      | 0.856 ± 0.012 | 0.879 ± 0.019 (↑) | 0.895 ± 0.021 (↑) |
|          | Vaso.      | 0.826 ± 0.024 | 0.859 ± 0.013 | 0.893 ± 0.018 (↑) |

implemented under the membership inference for GANs in the black-box settings [57]. The attacker is assumed to possess complete knowledge of all the patient records set $P$, where a subset from $P$ further is used to train GANs. During the experiment, the number of samples in the subset for training EHR-M-GAN are varied to investigate the impact of the availability of training data on the success of the attacker (see Figure 7a). By observing the synthetic patient records from EHR-M-GAN, the adversary’s goal is to determine whether a single known record $x$ in the patient record set $P$ is from the data used in training EHR-M-GAN. Determined by whether the attacker can correctly infer a given record is in or not in GAN’s training, the accuracy and recall can be calculated.

As shown in Figure 7a, when 90% of the training data is used for developing EHR-M-GAN, the attacker had a recall of 0.533 and accuracy of 0.527 to recover which training data are considered. This is eminently close to flipping a coin with random guess (i.e., 0.5), indicating EHR-M-GAN is sufficiently robust against the membership inference attack. On the other hand, as the percentage of the training data reduces, both accuracy and recall for membership inference attacks rise. An accuracy of 0.624 and recall of 0.732 are reached with 20% of training data. This offers a guideline for future application in developing GANs that incorporating more training data can make the generator less susceptible to such attack. This is also consistent with the conclusion derived from the experiment on membership inference attacks in the prior research [59].

The concept differential privacy (DP) [60], which is a rigorous mathematical definition of privacy, has emerged to be the prevailing notion in the context of statistically analyzing data privacy. The $(\epsilon, \delta)$-differential privacy is guaranteed for model $M$, if given any pair of adjacent datasets $D$ and $D'$ (differing on a single patient record), it holds: $P[M(D) \in S] \leq e^\epsilon P[M(D') \in S] + \delta$. In our case, $M(\cdot)$ is the GAN model trained based on $D$ or $D'$, and $S$ is the subset of any possible outcomes of the generative process. By perturbing the underlying data distribution, DP bounds the maximum variations of the algorithm when any single individual is included or excluded from the dataset. In practice, recent works on developing differentially private deep learning models has benefited from differential private stochastic gradient descent (DP-SGD) algorithm. DP-SGD operates DP by gradient clipping and noise adding during SGD, thereby ensuring that the impact of single record in the training dataset on algorithm parameters is limited within DP’s extend. In this section, $(\epsilon, \delta)$-differential privacy is implemented in EHR-M-GAN using TensorFlow Privacy$^4$. We then perform the same

$^4$https://github.com/tensorflow/privacy
downstream tasks on medical intervention prediction using synthetic data generated from DP-guaranteed EHR-M-GAN, and compare its performance with TSTR (as shown in Table 4).

Figure 7b shows the TSTR performance of EHR-M-GAN under differential privacy guarantee with varying budgets $\epsilon$ ($\delta$ fixed at $\leq 0.001$). The value $\epsilon$ determines how strict the privacy is, where the smaller value indicates a stronger privacy restriction. As suggested in Figure 7b, the performance of the downstream tasks operated based on the synthetic data generated by EHR-M-GAN improves as the DP budget increases ($\epsilon$ increases). We observe that the AUROC of DP-bounded EHR-M-GAN can maintain at an acceptable level even under strict privacy setting. For example, the AUROC for predicting the treatment of Vasopressor can maintain at 0.714 (AUROC = 0.725 under $\epsilon \leq 0.1$) even when the $\epsilon$ decrease to 4, which is an empirically reasonable value for implementing DP in practice [61]. Future work that focuses on privacy-preserving GAN under DP-guarantee is expected, where the fidelity of the synthetic data can be restored without compromising its privacy.

Discussion and conclusions

In this study, we propose a generative adversarial network entitled EHR-M-GAN, aiming at mitigating the challenge of synthesizing longitudinal EHR with mixed data types. To better capture the correlations between the continuous-valued and discrete-valued timeseries, shared latent representations are learnt based on the proposed dual-VAE, where the dimensionality for the subsequent adversarial learning is reduced. Then the proposed sequentially coupled generator built based on the architecture of RNNs enables EHR-M-GAN to model the temporal dependencies between heterogeneous data. During the quantitative and qualitative evaluations of the proposed model, both EHR-M-GAN and its conditional version, EHR-M-GAN\_cond, demonstrate consistent improvements against the state-of-the-art benchmark GANs in synthesizing timeseries data with high-fidelity. Notably, as opposed to previous models which were confined to synthesizing only one specific type of data, EHR-M-GAN can produce mixed-type timeseries and successfully capture the temporal dynamics within. During downstream task evaluation, given the prediction of medical interventions in fast-paced critical care environments as an exemplar, the results demonstrate the broad applicability of our model in developing ML algorithm-based decision support tools through data augmentation.

There are limitations in the current work. First, our model is built on smoothed and imputed clinical timeseries, therefore neither the irregular time intervals between signals nor missing values within the timeseries are modeled. As during the data preprocessing, we hourly aggregate patients’ physiological and intervention signals based on their mean statistics, and then complete the missing value in the timeseries through the “Simple Imputation” approach [46, 62] that is commonly used in clinical research (see S.2.B in Supplementary materials for details). However, dealing with irregularity of the timestamps when synthesizing clinical events in EHRs could be useful for predicting outcomes that are time-aware in the downstream tasks [32]. Modeling such time intervals could be non-trivial as the determinative perspectives sometimes go beyond the scope of inferring patients’ physiological status such as resource allocations within hospitals. Furthermore, synthesizing timeseries while incorporating the missing values could also be beneficial in the real-world application scenarios. As ML models are sometimes sensitive to the data missingness, imputing the incomplete data in EHRs using generative approaches could improve the performance of ML models, and has become an area of active research [63].

Second, although our models successfully extend the scope of synthetic data generation into mixed-type timeseries, they are still limited in single modality (i.e., structured EHR data). However, the adoption of digitized healthcare information systems provide a vast range of other — heterogeneous databases with mixed modalities [64]. Open-access data with various modalities are provided across the computational health community, including medical image data (such as magnetic resonance imaging (MRI) scans from BraTS [65]), physiological waveform signals (such as ECG and PPG data from MIMIC-
III Waveform [66]), as well as unstructured natural language information (such as free-text clinical notes from 1232 platform [67]). Future work can be investigated in the topic of generating synthetic data with multi-modalities, as integrating such data will increase the value of data-driven ML models in the real world clinical applications.

Finally, the conditional aspect of our model is currently limited as it can not generate patient-specific EHRs conditioning on information at a more granular level. Even though the proposed conditional GANs can synthesize a subgroup of patients with target outcomes or statuses that clinicians specify, it is still limited in incorporating personalised information during the conditional generation. Future work for developing GANs in healthcare data can be extended to patient-level EHRs generation, such as synthesizing counterfactual information of a target patient for treatment effect estimation [68, 69].

Ultimately, by constructing the “synthetic twin” of patients using GANs, the synthesis tool can become more generalisable for precision medicine and support the clinical decision making in delivering personalized healthcare.

Synthetic data provides an alternative to sharing real patient data while preserving patient privacy. Results in our study demonstrate that the proposed EHR-M-GAN and EHR-M-GANcons can generate realistic longitudinal EHR timeseries with mixed data types. By providing synthetic EHR data with higher fidelity and more variety, the proposed model can therefore enable faster development in AI-driven clinical tools with increased robustness and adaptability. In addition to the improved performance against the existing state-of-the-art benchmark models, augmentation provided by synthetic data during training boosts the predictive performance in downstream clinical tasks. EHR-M-GAN can help eliminate the barriers to data acquisition for healthcare studies, therefore overcoming the challenges posed by the paucity of medical data available and approved for research use. Despite the novelty of this study in filling the research gap for synthesizing longitudinal EHRs in mixed-type settings, we acknowledge that there is still a gap between the real EHRs data and its synthetic counterparts produced by current generative methods. Therefore developing advanced EHR synthesizers especially in mixed-type settings still requires active research in the future study.

Methods

In this section, we first formulate the problem based on the mixed-type temporal EHR data and its corresponding mathematical notation. Then the proposed EHR-M-GAN model is introduced in details.

Problem formulation. The longitudinal patient EHR dataset is denoted as $D = \{ (x_{i,T}) \}_{i=1}^{N}$, with each record (e.g., individual patient) being indexed by $i \in \{1, 2, \ldots, N \}$. Here we consider the $i$-th instance tuple $x_{i,T} = \{ x_{i,1:T_i}, x_{i,1:T_i} \}$ consists of two components (i.e., two types of data). Let $x_{i,1:T_i} \in \mathbb{R}^{J_i}$ denote the $J_i$-dimensional continuous-valued timeseries, such as physiological signals from real-time bedside monitors. And $x_{i,1:T_i} \in \mathbb{R}^{K_i}$ denotes the $K_i$-dimensional discrete-valued timeseries, such as life-support interventions with the categorical value indicate its status (presence or absence).

Proposed model. The preliminary for GANs and conditional GANs can be found in Supplementary materials (See Section S.1.B). We factorize EHR-M-GAN into two key components (see Network architecture in Fig. 1): (1) a dual-VAE framework for learning the shared latent space representations; (2) an RNN-based sequentially coupled generator and its corresponding sequence discriminators. As shown in Fig. 1, during the pretrain stage, both continuous-valued and discrete-valued temporal trajectories are first jointly mapped into a shared latent space (first component). Then, the sequentially coupled generator in EHR-M-GAN (second component) produces the synthetic latent representations, which further can be recovered into features in the observational space by the pretrained decoders in the dual-VAE. Finally, the adversarial loss is provided based on discriminative results and backpropagated to update the network. The following sections discuss them in turn.

**Dual-VAE pretraining for shared latent space representations.** One premise of successfully training EHR-M-GAN to generate reversible latent codes is to meet the assumption that for the same patient indexed with $i$, both $x_{i,1:T_i}$ and $\tilde{x}_{i,1:T_i}$ can be encoded into the same latent space $\mathcal{H}^{\theta}$ $\subset \mathbb{R}^{\theta}$, where $\mathcal{S}$ denotes its spatial dimension. For the sake of simplicity, the subscripts $i$ are omitted throughout most of the paper. To achieve this, we propose to use a dual-VAE framework to encode both continuous and discrete multivariate timeseries into dense representations within $\mathcal{H}^{\theta}$ based on multiple constraints. We found VAE preferable to vanilla autoencoder in our case, considering that (1) the KL regularization in VAE strengthens the learning of the compressed latent representations, which further narrow the domain gap for mixed-type features [70]; (2) VAE can be easily extended to the conditional learning scenario in EHR-M-GANcons.

Fig. S2 (see Section S.1.C in Supplementary materials) diagrams the details of the proposed dual-VAE framework for learning the shared latent representations. We start with training two encoders, i.e., $Enc^C : \phi_{T \times X} \rightarrow \phi_{T \times \mathcal{H}^C}$ and $Enc^D : \phi_{T \times X} \rightarrow \phi_{T \times \mathcal{H}^D}$, with the embedding functions:

$$
\tilde{z}_{1:T_i}^C = Enc^C(x_{1:T_i}) \quad \tilde{z}_{1:T_i}^D = Enc^D(x_{1:T_i}) \quad \text{[1]}
$$

After passing data from $\mathcal{X}^C$ and $\mathcal{X}^D$ through two encoders, a pair of embedding vectors $(\tilde{z}_{1:T_i}^C, \tilde{z}_{1:T_i}^D)$ in the shared latent space $\mathcal{H}^C$ can be obtained. Then the decoders for both domains $Dec^C : \psi_{T \times \mathcal{H}^C} \rightarrow \psi_{T \times \mathcal{X}^C}$ and $Dec^D : \psi_{T \times \mathcal{H}^D} \rightarrow \psi_{T \times \mathcal{X}^D}$ further reconstruct features based on the latent embeddings using mapping functions that operate in the opposite direction:

$$
\tilde{x}_{1:T_i}^C = Dec^C(\tilde{z}_{1:T_i}^C) \quad \tilde{x}_{1:T_i}^D = Dec^D(\tilde{z}_{1:T_i}^D) \quad \text{[2]}
$$

For the conditional extension in EHR-M-GANcons, both encoders and decoders in the dual-VAE condition on the auxiliary (one-hot) labels from $\mathcal{L}$, to make the model better adapted to particular contexts. Also, to incentivize dual-VAE to better bridge the gap between domains of continuous-valued and discrete-valued timeseries, we enforce a weight-sharing constraint [71, 72] within specific layers of both the encoders pairs ($Enc^C, Enc^D$), and the decoders pairs ($Dec^C, Dec^D$). The weight-sharing constraint can extract and broadcast high-level concepts of the input features across domains of mixed-type data. The implementation details can be found in Supplementary materials (See Section S.1.B).

In the following subsections, we define multiple loss constraints for the optimization of dual-VAE, including ELBO.
loss, Matching loss, Contrastive loss, as well as Semantic loss
for EHR-M-GAN$_{cond}$. Intuitions and descriptions behind the
objectives are discussed in turn.

Evidence Lower Bound (ELBO). We first incorporate the
standard VAE loss, with the optimization objective as the
evidence lower bound (ELBO). VAE holds the assumption of
spherical Gaussian prior for the distribution of latent embed-
dings, where features can then be reconstructed by sampling
from that space. We leverage the re-parameterization tricks
to enable differentiable stochastic sampling and network op-
timization. For encoder and decoder in the dual-VAE for
domain $d \in \{C, D\}$, the objective function is defined as:

$$
\mathcal{L}^\text{ELBO}_{d} = -\mathbb{E}_{q_{\phi}(x | z)} \left[ \log p_{\theta}(x | z) \right] + \
\beta_{\text{KL}} D_{\text{KL}}(q_{\phi}(z | x) \| p_{\psi}(z))
$$

[3]

where $z \sim \text{Enc}(x) \triangleq q_{\phi}(z | x), \hat{x} \sim \text{Dec}(z) \triangleq p_{\psi}(x | z)$, and
$D_{\text{KL}}$ is the Kullback-Leibler divergence. The first term in
Eq. (3) is the expected log likelihood term that penalizes the
deviations in reconstructing the inputs, while the second term
of KL-divergence is the regularization imposed over the latent
distribution from its Gaussian prior (normally chosen to be
$\mathcal{N}(0, I)$). $\beta_{\text{KL}}$ is the hyperparameter for balancing the weights
between two terms.

Matching loss. Typically, representations derived from the
same patient are assumed to capture the shared context. There-
fore, embedding vectors $(z^c_{i1 : T_i}, z^d_{i1 : T_i})$ projected from the
same patient $i$, are supposed to be positioned closely in the
shared latent space (See Fig S2 in Supplementary materi-
als). The matching loss ensures that low-dimensional latent
space can be shared between heterogeneous features. Hence,
the pairwise matching loss is incorporated to motivate the
encoders to minimize the distance within the corresponding
representation pairs. In the low-dimensional Euclidean space,
we optimize the network by using the following objective:

$$
\mathcal{L}^\text{Match} = \mathbb{E}_{d \sim p_{\mathcal{D}}} \sum_{i \in T} \left\| z^c_i - z^d_i \right\|^2
$$

[4]

The pairwise matching loss achieve its optimal when the dis-
tance proxy $\mathcal{L}^\text{Match}$ becomes zero. However, as minimizing the
matching loss alone might lead to a trivial function [73], we
further add contrastive loss and semantic loss (for the conditional
extension of EHR-M-GAN$_{cond}$) to balance the optimization for
the shared latent space for dual-VAE.

Contrastive loss. On the flip side, pairwise reconstruction er-
ror (i.e., intra-correlations within one instance) measured by
matching loss neglects the commonalities present across pa-
tients (inter-correlations of data) [74]. In order to guarantee
sufficient bound for representation learning, we incorporate
contrastive loss as another distance metric to capture the
inter-correlations among the population.

Contrastive learning is a concept that has recently been pop-
ularized in self-supervised learning (SSL) and representation
learning, which aims to capture intrinsic patterns from input
data without human annotations. The core of contrastive
learning is to encourage networks to attract positive pairs
closer and repulse negative pairs apart in the latent space. In
this study, we instantiate the contrastive loss via NT-Xent, which is proposed by Chen et al. in their work SimCLR [75].
We define the corresponding auxiliary tasks for calculating
the contrastive loss as — whether a set of representations transformed from the observational space belong to the same
patient. And this leads to the corresponding pairs (true) and negative pairs (false).

For patient data of $N$ records, we can obtain $N$ pairs of
talent representations from the encoders in dual-VAE. For
patient indexed with $i$, $h^c_i$ and $h^d_i$ denotes the embeddings derived from the continuous-valued and discrete-valued observ-
Tional space, respectively. Due to the symmetric architecture of
dual-VAE, here we use $d$ and $d'$ to represent one of each differ-
ent domain, i.e., $d, d' \in \{C, \bar{D}\}$ and $d \neq d'$. Therefore,
the positive pairs for patient $i$ can be referred as $(i^c, i^d)$, while the other $(2N - 1)$ samples are regarded as negative pairs.

Then the contrastive loss for a positive pair $(i^c, i^d)$ is defined as:

$$
\mathcal{L}^\text{Contra}_{i^c, i^d} = - \log \frac{\exp \left( \frac{\text{sim}(h^c_i, h^d_i)}{\tau} \right)}{\sum_{d'^{\neq i}} \sum_{d'^{\neq i'}} \exp \left( \frac{\text{sim}(h^c_i, h^{d'\prime}_{i'})}{\tau} \right)}
$$

[5]

where $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denotes the cosine similarity
between two vectors. $\tau > 0$ denotes a temperature hyperpara-
meter. $\{i, i' \mid i \neq i' \} \in \{0, 1\}$ is an indicator evaluating to 1 iff
$n \neq m$. And $\{d, d' \mid d \neq d' \}$ represents the index of latent embeddings from both data types. The final contrastive loss is
computed across the total number of $|d^c - i^d| = N$ positive
pairs for both $(i^c, i^d)$ and $(i^d, i'^c)$, and is defined as:

$$
\mathcal{L}^\text{Contra} = \frac{1}{2N} \sum_{i^c=1}^{N} \sum_{i^d=1}^{N} \left[ \mathcal{L}^\text{Contra}_{i^c, i^d} + \mathcal{L}^\text{Contra}_{i^d, i'^c} \right]
$$

[6]

Semantic loss. In EHR-M-GAN$_{cond}$, semantic loss is imposed to
better align patients with same labels (conditions) into the
same latent space clusters. For example, if the label of severe
clinical deterioration in the ICU is given for conditional data
generation, the corresponding synthetic continuous-valued
timeseries (e.g., severely deranged vital signs) is supposed to
be strongly associated with the discrete-valued timeseries (e.g.,
intensive medical interventions) under the same label. For both
domains, additional linear classifiers are trained to classify
the latent embeddings based on their corresponding conditions
in the observational space. We implement logistic regression
as the linear classifiers and calculate the cross entropy as the
semantic losses for both domains. For $d \in \{C, \bar{D}\}$, given the
latent embedding vector $z^d$ and the conditional information
vector $y$:

$$
\mathcal{L}^\text{Class}_d = \mathbb{E}_{z^d \in \mathcal{H}^d} \mathbb{C}E \left( f^d_{\text{linear}}(z^d), y \right)
$$

[7]

where $f^d_{\text{linear}}$ denotes the linear classifier for the correspond-
ing domain. And $CE = - \sum y_j \log (y_j \hat{y}_j) \ (j = 1, 2, \ldots, |L|)$
denotes the cross entropy loss, where $y_j$ is the output of the
linear classifier, and $\hat{y}_j$ is the ground truth label for class $j$ in
condition vector $y$.

In summary, to train the dual-VAE for learning the shared
latent space representation, the total objective function for
d $\in \{C, \bar{D}\}$ is:

$$
\mathcal{L}_d = \beta_0 \mathcal{L}^\text{ELBO} + \beta_1 \mathcal{L}^\text{Match} + \beta_2 \mathcal{L}^\text{Contra}
$$

[8]

Under the conditional learning scenario of EHR-M-GAN$_{cond}$,
the total loss becomes:

$$
\mathcal{L}_d = \beta_0 \mathcal{L}^\text{ELBO} + \beta_1 \mathcal{L}^\text{Match} + \beta_2 \mathcal{L}^\text{Contra} + \beta_3 \mathcal{L}^\text{Class}_d
$$

[9]
where $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are scalar loss weights used to balance the loss terms. To validate the effectiveness of multiple losses and the weight-sharing constraint in the proposed dual-VAE network, we perform the corresponding ablation study using MIMIC-III dataset as an example. The results can be found in S.3.E in the Supplementary materials. As shown in Table S7, all the components in the proposed dual-VAE network contribute to the improvement of EHR-M-GAN’s performance when generating mixed-type timeseries data.

**Sequentially coupled generator based on CRN.** The objective of EHR-M-GAN is to capture the dynamics in the mixed-type timeseries, which can be regarded as the composition of the spatial and temporal correlations. To this end, we propose the sequentially coupled generator, which is built based on the coupled recurrent network (CRN). Specifically, a CRN comprises two single recurrent networks (SRN) (one for each type of the multivariate spatial timeseries, i.e., continuous or discrete-valued). In each SRN, LSTM network is introduced as the recurrent layer to preserve the temporal dependencies in the sequences. The description for the basic structure of SRN can be found in the Supplementary materials (see Section S.1.D). In the following section, we first discuss the structure of BLSTM in detail as an essential part of CRN, and then build the sequentially coupled generator based on CRN.

**Bilateral long short-term memory.** As shown in Fig. 1, two streams of BLSTM entangles with each other in a CRN, and are jointly trained to sequentially exploit the correlations between two types of trajectories. Given $d, d' \in \{C, D\}$ in CRN, since two branches of BLSTM associated with the corresponding domains are symmetric, we present the mathematical expressions from the perspective of branch $d$ as an example to illustrate the algorithm of BLSTM in details. Similar as the formulations in SRN (see Section S.1.D for the detailed description), $\psi^d_t$ and $h^d_T$ denotes the input vector and hidden state vector for domain $d$ at time step $t$, respectively. An additional set of weights for introducing hidden states representations $h^d_T$ from domain $d'$ is included when updating the input gate $i^d_t$, forget gate $f^d_t$, output gate $o^d_t$, and cell memory $c^d_T$. The state transition functions for BLSTM are:

\[
\begin{align*}
    h^d_{t+1} &= \sigma \left(W_{\text{adv}} \psi^d_t + W_{\text{adv}} h^d_t + W_{\text{adv}} d h^d_{t-1} + b_{\text{adv}}\right) \\
    f^d_{t+1} &= \sigma \left(W_{\text{adv}} \psi^d_t + W_{\text{adv}} h^d_t + W_{\text{adv}} d h^d_{t-1} + b_{\text{adv}}\right) \\
    o^d_t &= \sigma \left(W_{\text{adv}} \psi^d_t + W_{\text{adv}} h^d_t + W_{\text{adv}} d h^d_{t-1} + b_{\text{adv}}\right) \\
    c^d_t &= \tanh \left(W_{\text{adv}} h^d_t + h^d_{t-1} + b_{\text{adv}}\right) \\
    h^d_t &= o^d_t \cdot \tanh (c^d_t) \\
\end{align*}
\]

The additional set of weights in BLSTM facilitates the model to intrinsically decide how much information it should pass through the gates from its counterpart. An illustration of the BLSTM cell can be found in the Supplementary materials (see Fig. S3).

**Coupled recurrent network.** The architecture of CRN consists of three layers: the input layers, the recurrent layers, and the fully connected layers. First, the random noise vectors $\psi^C_t$ and $\psi^D_t$ for two domains are separately fed into the input layers. Then the recurrent layers $f_{\text{rec}}$, which are built based on BLSTM as suggested in Eq. (10), are used to recursively iterate hidden states from both branches. Finally, the fully connected layers $f^C_{\text{con}}$ and $f^D_{\text{con}}$ produce the generated latent vectors $\mathbf{z}^C_t$ and $\mathbf{z}^D_t$ for the decoding stage in dual-VAE. At time step $t$, CRN can be formulated as:

\[
\begin{align*}
    (h^C_t, h^D_t) &= f_{\text{rec}}((\psi^C_t, \psi^D_t), (h^C_{t-1}, h^D_{t-1})) \\
    \mathbf{z}^C_t &= f^C_{\text{con}}(h^C_t) \\
    \mathbf{z}^D_t &= f^D_{\text{con}}(h^D_t)
\end{align*}
\]

Note that the parameters are initialized independently within the two branches of the recurrent layers $f_{\text{rec}}$ regardless of the coupled architecture of the network. With respect to EHR-M-GAN$_{\text{cond}}$, the generator $G^{CRN}$ generates synthetic latent embeddings conditioned on the auxiliary information $y$ (conditional labels $y$ are also incorporated during the pretrain stage for dual-VAE). In summary, heterogeneous timeseries that exhibits mutual influence are integrated into the parallelly coupled network architecture to leverage the shared information. By introducing BLST, two streams of the inputs in the generator are encouraged to “communicate” with each other. CRN is therefore capable of exploiting the interplay between mixed-type data that correlates over time.

**Joint training and optimization.** The overall architecture of EHR-M-GAN is shown in Fig. 1. In this section, we give a detailed description of the entire network’s structure and the optimization objective of the model. The steps for the training and optimization of EHR-M-GAN are as follows:

- **The pretraining of dual-VAE:** First, a dual-VAE network which consists of a pair of encoders ($\text{Enc}^C, \text{Enc}^D$) and decoders ($\text{Dec}^C, \text{Dec}^D$) is pretrained with both continuous and discrete data. Based on multiple objective constraints in Eq. 8 (for EHR-M-GAN$_{\text{cond}}$ the objective function can be referred in Eq. 9), a shared latent space is learnt using dual-VAE, where the gap between the embedding representations $(\mathbf{z}^C_{1:T}, \mathbf{z}^D_{1:T})$ from both domains is minimized.

- **The generation of latent representations based on CRN:** Then, during the joint training stage, the sequentially coupled generator is built based on CRN with the internal transition functions iterating across the timesteps $t \in \{1, 2, ..., T\}$. Therefore, the synthetic latent embedding representations $(\mathbf{z}^C_{1:T}, \mathbf{z}^D_{1:T})$ for both continuous and discrete data can be obtained.

- **The decoding for the mixed-type timeseries:** Next, the generated latent embeddings $(\mathbf{z}^C_{1:T}, \mathbf{z}^D_{1:T})$ are further fed into the pretrained decoders ($\text{Dec}^C, \text{Dec}^D$) and decoded into the corresponding synthetic patient trajectories $(\mathbf{x}^C_{1:T}, \mathbf{x}^D_{1:T})$ in the observational space.

- **The adversarial loss update based on the discriminators:** Finally, the adversarial loss can be calculated from the LSTM network-based discriminators $D^C$ and $D^D$ by distinguishing between the real samples and synthetic timeseries for both data types.

The mathematical expression for the min-max objectives in
EHR-M-GAN is provided as follows:

$$\min_G \max_D V_{\text{EHR-M-GAN}} =$$

$$\mathbb{E}_{x \sim p_{\text{data}}} [\log D^C(x^C)] + \mathbb{E}_{x \sim p_{\text{data}}} [\log D^P(x^P)] + \mathbb{E}_{x \sim p_{\text{data}}} [\log(1 - D^C(x^C))] + \mathbb{E}_{x \sim p_{\text{data}}} [\log(1 - D^P(x^P))].$$

Note that in the conditional version of EHR-M-GAN$_{\text{cond}}$, as we mentioned before, the auxiliary label information is used during the pretraining step of Dual-VAE. Both the encoders and decoders condition on the contextual information, while the additional semantic loss is also incorporated during the optimization for the shared latent space (see Eq. 9). Meanwhile, the same conditional labels are also applied in the sequentially coupled generator and discriminators, where the classes are fed as additional inputs through concatenation, as in the original CGAN architecture proposed by Mirza et al [50]. The adversarial loss for EHR-M-GAN$_{\text{cond}}$ can be denoted as follows:

$$\min_G \max_D V_{\text{EHR-M-GAN}_{\text{cond}}} =$$

$$\mathbb{E}_{x \sim p_{\text{data}}} [\log D^C(x^C|x^P)] + \mathbb{E}_{x \sim p_{\text{data}}} [\log D^P(x^P|x^P)] + \mathbb{E}_{x \sim p_{\text{data}}} [\log(1 - D^C(x^C|x^P))] + \mathbb{E}_{x \sim p_{\text{data}}} [\log(1 - D^P(x^P|x^P))].$$

The pseudocodes for dual-VAE and EHR-M-GAN are provided in the Supplementary materials (see Section S.1.E).

Acknowledgement

We owe thanks to PhysioNet for the availability of the critical care databases. We also would like to express our sincere thanks to Lara Prisco (Nuffield Department of Clinical Neurosciences, University of Oxford) and Baptiste Vasey (Nuffield Department of Surgical Sciences, University of Oxford) for the clinical input in the paper. JL was funded by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. TZ was supported by the RAEng Engineering for Development Research Fellowship and the National Institute for Health Research Oxford Biomedical Research Centre.

Author contributions

J.L., T.Z., and B.J.C. conceived the idea. J.L. and T.Z. contributed to the model implementations and experiment designs. J.L., T.Z., and B.J.C. interpreted, analyzed and presented the results. J.L., T.Z., and B.J.C. contributed to the writing and revising of the manuscript. J.L., J.S.L., and T.Z. contributed to the data acquisition and resource allocation.

Competing interests

The authors declare no competing interests.

Data availability

All datasets are freely accessible. The MIMIC-III dataset can be accessed by: https://physionet.org/content/mimiciii/1.4/. The eICU-CRD database can be accessed by https://physionet.org/content/eicu-crd/2.0/. The HiRID dataset can be accessed by: https://physionet.org/content/hirid/1.1.1/.

Code availability

Algorithms are developed using Python with the deep learning networks implemented with Tensorflow. Code for preprocessing three ICU datasets can be found in: https://github.com/jil0117/preprocessing_physionet. Code for proposed EHR-M-GAN model can be found in: https://github.com/jil0117/ehrmogan.

References

[1] Niranjan Shalom Arzi, Smadar Shilo, Eran Hadar, Hagai Rossmann, Shiri Barbash-Hazan, Avi Ben-Harsush, Ran D Balicer, Becca Feldman, Aman Winizter, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine, 26(1): 71–76, 2020.
[2] Lars Lau Raketa, Jörn Jaskolkowski, Bruce J Kinon, Jens Christian Brasen, Linus Jonsson, Allan Wehrent, and Paolo Fusi-Pol. Dynamic electronic health record detection (detect) of individuals at risk of a first episode of psychosis: a case-control development and validation study. The Lancet Digital Health, 2(3):e229–e239, 2020.
[3] Vincent Menger, Marco Spruit, Roel Van Est, Eline Nap, and Floor Schepers. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA network open, 2(7):e196709–e196709, 2019.
[4] Jack Wilkinson, Kelyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, Andrew Beam, Stefan Koryvzogi, Christopher Lippert, et al. Time to reality check the promises of machine learning powered precision medicine. The Lancet Digital Health, 2020.
[5] David S Watson, Jenny Krutinina, Ian N Bruce, Christopher EM Griffiths, Iain M Bokmøes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. BMJ, 364, 2019.
[6] Joseph Futoma, Morgan Simonsen, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Cole. The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health, 2(9):e489–e492, 2020.
[7] Andre Estova, Katherine Chou, Serena Yeung, Nikiel Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. NPJ digital medicine, 4(1):1–9, 2021.
[8] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. New England Journal of Medicine, 380(14):1347–1358, 2019.
[9] Felix Nicolaus Wirth, Thierry Meurers, Marco Johns, and Fabian Prasser. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Medical Informatics and Decision Making, 21(1):1–13, 2021.
[10] Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. Gigascience, 5(1):s13742–016, 2016.
[11] Ricardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 19(6):1236–1246, 2018.
[12] Jihoon Kim, Larissa Neumann, Paulina Paul, Michele E Day, Michael Aratow, Douglas S Bell, Jason N D’Orso, Ludwig C Hincke, Xiaoqian Jiang, Katherine K Kim, et al. Privacy-protecting, reliable response data discovery using covid-19 patient observations. Journal of the American Medical Informatics Association, 28(1):1765–1776, 2021.
[13] Gregory E Simon, Susan M Shortreed, R Yates Coley, Robert P Bentolida, Rebecca C Rossom, Beth E Wiatfelder, Katherine Sanchez, and Frances L Lynch. Assessing and minimizing re-identification risk in research data derived from health care records. eGEMs, 7(1), 2019.
[14] James Jordson, Jinsuig Yoon, and Mihaela Van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In International conference on learning representations, 2018.
[15] Maayan Frid-Aidar, Iddi Diamant, Eyal Klang, Michal Amitai, Jacob Goldberg, and Hayit Ben-Haroush, Ran D Balicer, Becca Feldman, Aman Winizter, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine, 26(1): 71–76, 2020.
[16] James Jordson, Jinsuig Yoon, and Mihaela Van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In International conference on learning representations, 2018.
[17] James Jordson, Jinsuig Yoon, and Mihaela Van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In International conference on learning representations, 2018.
[18] Maayan Frid-Aidar, Iddi Diamant, Eyal Klang, Michal Amitai, Jacob Goldberg, and Hayit Ben-Haroush, Ran D Balicer, Becca Feldman, Aman Winizter, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine, 26(1): 71–76, 2020.
[19] James Jordson, Jinsuig Yoon, and Mihaela Van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In International conference on learning representations, 2018.
[20] Maayan Frid-Aidar, Iddi Diamant, Eyal Klang, Michal Amitai, Jacob Goldberg, and Hayit Ben-Haroush, Ran D Balicer, Becca Feldman, Aman Winizter, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine, 26(1): 71–76, 2020.
[25] Vasant Kesavay, Jason W Chan, Tianqi Wang, Alan Perry, Martina Descovich, Olivier Morin, Sue Y Song, and Timothy D Sollberg. Dosegan: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. Scientific reports, 10(1):1–8, 2020.

[26] Chengyang Yang, Pingyan Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xiaquan Miu, Manudeep K Karia, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE transactions on medical imaging, 37(8):1348–1357, 2018.

[27] Mohamed Maroul, Robert Macfarlane, Wei Wang, Christoph Kilian, Daniel S Magruder, Chris T Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. Nature communications, 11(1):1–12, 2020.

[28] Kang Zhou, Shenghua Gao, Jun Cheng, Zaixiong Gu, Huazhu Fu, Zhi Tu, Jionglang Yang, Yitian Zhao, and Jiang Liu. Sparse-gan: Sparsity-constrained generative adversarial network with wasserstein distance and perceptual loss. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pages 1227–1231, IEEE, 2020.

[29] Thomas Schlegel, Philipp Seeböck, Sebastian Stadlober, Georg Lang, and Ursula Schmidt-Erfurth. F-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis, 54:30–44, 2019.

[30] Cristobal Estevez, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. arXiv preprint arXiv:1706.02633, 2017.

[31] Dongha Lee, Hwanjo Jeon, Dooeak Roghi, Meghana Guadala, Mubeen Tejarj, Quichen Zhang, and Li Xiong. Generating sequential electronic health records using dual adversarial autocoder. Journal of the American Medical Informatics Association, 27(9):1411–1419, 2020.

[32] Zip Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. Synteg: a framework for temporal structured electronic health data simulation. Journal of the American Medical Informatics Association, 28(3):596–604, 2021.

[33] Dan Kvasnica, Girish Abate Tadesse, Louise Thwaites, Tizting Zhu, David Clifton, et al. Plheaugment: Gan-based ppg augmentation for medical diagnosis in low-resource settings. IEEE journal of biomedical and health informatics, 24(1):3226–3235, 2020.

[34] L Nelson Sanchez-Pinto, Yuan Luo, and Matthew M Churpek. Big data and data science in critical care. Critical care, 21(1):1–15, 2019.

[35] Duncan Shillan, Jonathan AC Stemer, Alan Champneys, and Ben Gibson. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Critical care, 23(11):1–15, 2019.

[36] Hao Yu, Jining Liu, and Hongyi Zhao. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC medical informatics and decision making, 19(2):111–120, 2019.

[37] Marzyeh Ghassemi, Mike Wu, Michael S Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. In 34th Conference on Neural Information Processing Systems, pages 1255–1265. 2020.

[38] Lu Wang, Wei Zhang, and Xiaofeng He. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In International Conference on Database Systems for Advanced Applications, pages 36–52. Springer, 2019.

[39] Jinseung Yoon, Daniel Jarrett, and Mihaela Van Der Schaar. Time-series generative adversarial networks. 2019.

[40] Lantao Yu, Weian Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

[41] Alistair E W Johnson, Tom J Pollard, Lu Shen, H Lehman Li Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, and Finale Doshi-Velez. Data augmentation intervention onset in the ICU with switching state space models. AMIA Summits on Translational Science Proceedings, 2017:82, 2017.

[42] Lu Wang, Wei Zhang, and Xiaofeng He. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In International Conference on Database Systems for Advanced Applications, pages 36–52. Springer, 2019.

[43] Jinseung Yoon, Daniel Jarrett, and Mihaela Van Der Schaar. Time-series generative adversarial networks. 2019.

[44] Lantao Yu, Weian Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

[45] Cynthia Dwork. Differential privacy: A survey of results. In International conference on theory and applications of models of computation, pages 1–19. Springer, 2008.

[46] Apple Differential Privacy Team. Learning with privacy at scale. 2017.

[47] Cheng Ren, Sanjay Purushotham, Kyoung Ho Chung, David Song, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. Scientific reports, 8(1):1–12, 2018.

[48] Jinseung Yoon, James Jordan, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In International conference on machine learning, pages 5689–5698. PMLR, 2018.

[49] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Iman Benanie, and Matthew P Luggren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine, 3(1):1–9, 2020.

[50] Bjoern H Merse, Andres Jakab, Stefan Bauer, Jayashree Kabthapramer, Keyvan Farahani, Justin Kirby, Yulina Burrow, Nicole Porz, Johannes Stobrott, Robin West, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging, 34(10):1993–2004, 2015.

[51] B Moody, G Moody, M Villarcel, G Clifford, and I Silva III. Mimir-iii waveform database (version 1.0), 2020.

[52] National NHL Clinical Challenges (n2cl2). 2022 n2cl2 Shared Task and Workshop, Boston, MA: Harvard Medical School, howpublished = https://n2cl2.dmi.hms.harvard.edu/2022-challenge. Accessed: 2021-12-08.

[53] Jinseung Yoon, James Jordan, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In International Conference on Learning Representations, 2018.

[54] Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. Syncite: Treatment effect estimation with longitudinal outcomes. Advances in Neural Information Processing Systems, 34, 2021.

[55] Ziyu Wan, Bo Zheng, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Photo restoration via deep latent space translation. arXiv preprint arXiv:2009.07047, 2020.

[56] Jingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.

[57] Mingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.

[58] Mingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.

[59] Mingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.

[60] Mingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.

[61] Mingyu Liu, Thomas Bruezel, and Jan Kautz. Unsupervised image-to-image translation with cyclic adversarial networks. In Advances in neural information processing systems, pages 700–708, 2017.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.pdf