

## ARTICLE OPEN



# Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications

Jin Li<sup>1,2</sup>, Benjamin J. Cairns<sup>3</sup>, Jingsong Li<sup>1,4</sup>✉ and Tingting Zhu<sup>1,2</sup>✉

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. To mimic the nature of clinical decision-making which encompasses various data types/sources, in this study, we propose a generative adversarial network (GAN) entitled EHR-M-GAN that simultaneously synthesizes *mixed-type* timeseries EHR 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.

npj Digital Medicine (2023)6:98; <https://doi.org/10.1038/s41746-023-00834-7>

## INTRODUCTION

The past decade has witnessed ground-breaking advancements been made in computational health, owing to the explosion of medical data, such as electronic health records (EHRs)<sup>1–3</sup>. 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<sup>4–8</sup>. However, in practice, the benefits of data-driven research are limited to healthcare organizations (HCOs) who possess the data<sup>9,10</sup>. Due to concerns about patient privacy, HCO stakeholders are reluctant to share patient data<sup>11–13</sup>. 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.

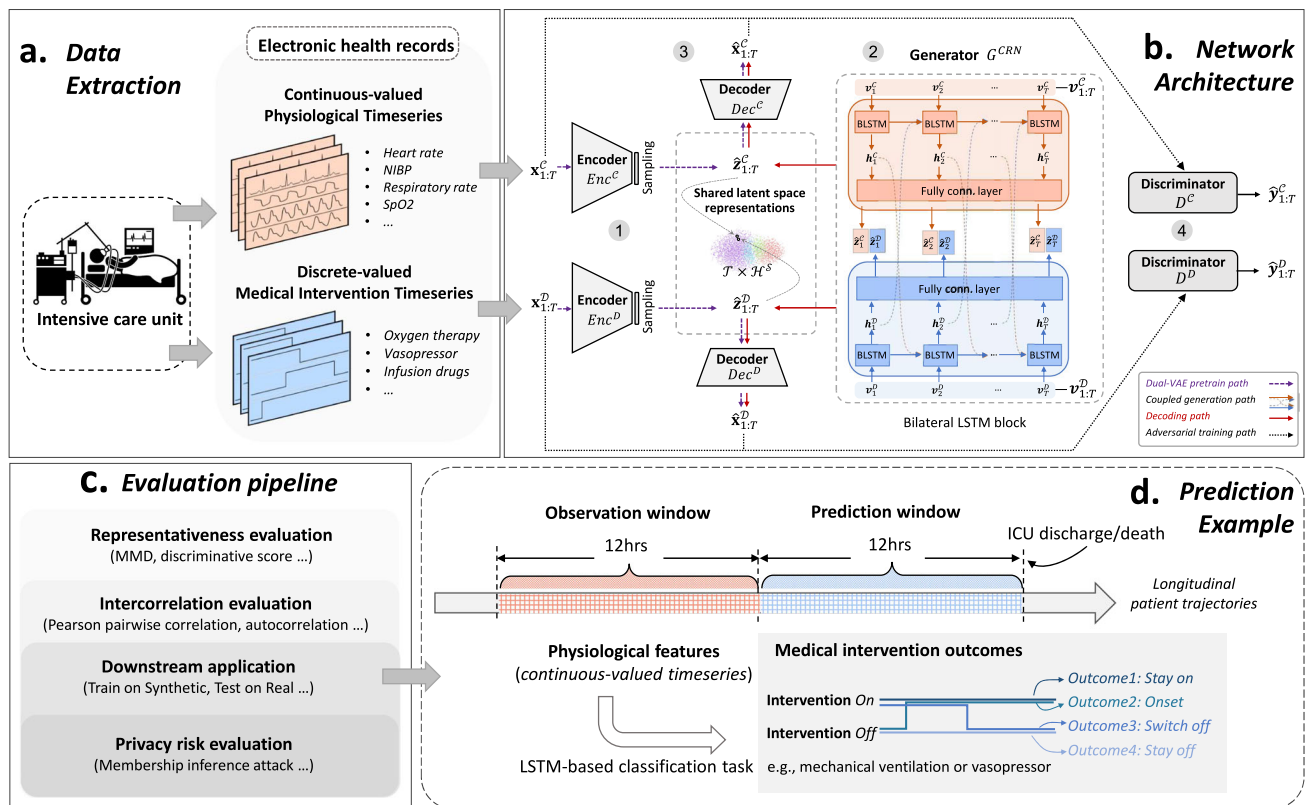
To accelerate the progress of developing AI methods in medicine, one promising alternative is for the data holder to create *synthetic* yet realistic data<sup>14,15</sup>. By avoiding “one-to-one” mapping to the genuine data compared with data anonymization, synthetic data offers a solution to circumvent 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<sup>16–18</sup>. 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)<sup>19</sup>, and Clinical Practice Research Datalink (CPRD) database in the UK<sup>20</sup>.

Recent advances in generative adversarial networks (GANs)<sup>21</sup> and their variants offer efficacious means to generate EHRs for a wide range of clinical applications<sup>22–24</sup>. Over the past years, EHR synthesizers have evolved from generating static patient

information to producing longitudinal EHR timeseries<sup>25–27</sup>. As longitudinal EHRs contain patient trajectories for describing the underlying health condition, synthesizing such EHR timeseries, therefore, enables new clinical applications related to the status of disease progression<sup>28</sup>, such as dynamic forecasting of risks, predicting the onset of diseases, and survival analysis based on the time-to-event data. However, existing studies focus on synthesizing the longitudinal EHRs of a single data type<sup>25,26,29</sup>, whereas the clinical decision-making in real practice includes a variety of information sources in the form of *mixed-type* timeseries. For example, patient physiological signals and laboratory test results are collected in the EHR as *continuous-valued* timeseries, while the medication and diagnostic information are recorded as *discretized-valued* data as binary indicators or categorical ICD codes. Information provided in these mixed-type longitudinal EHRs offer opportunities for more precise and complex clinical analysis. Furthermore, the predictive power and robustness of the ML models can be boosted by utilizing longitudinal EHR timeseries with various types/sources.

Existing GANs are limited in simulating mixed-type EHRs due to two reasons. Firstly, it is intrinsically difficult to model the underlying joint distribution of mixed data type timeseries using a single unified framework. Since GANs require the network architectures of the generator and discriminator to be fully differentiable<sup>30</sup>, its success is typically limited to generating real-valued, continuous data while facing obstacles for directly generating sequences of discrete tokens, such as ICD codes, that also commonly appear in EHRs. Previous methods<sup>31,32</sup> circumvent this problem by learning representations from the original data which further enables backpropagation in discrete settings, but

<sup>1</sup>Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China. <sup>2</sup>Department of Engineering Science, University of Oxford, Oxford, UK. <sup>3</sup>Clinical Trial Service Unit and Epidemiological Studies, Nuffield Department of Population Health, University of Oxford, Oxford, UK. <sup>4</sup>Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, China. ✉email: [ljs@zju.edu.cn](mailto:ljs@zju.edu.cn); [tingting.zhu@eng.ox.ac.uk](mailto:tingting.zhu@eng.ox.ac.uk)



**Fig. 1 Overall schematics.** **a** Data extraction. Electronic health records (EHRs) data are routinely collected for patients in intensive care units (ICUs). Intensively monitored vital signals and laboratory measurements are recorded as continuous-valued timeseries, while the presence or absence of medical interventions is collected as discrete-valued timeseries during the ICU admission. These mixed-type EHR data are correlated but distributed differently, and they change over time depending on the diagnoses provided by clinicians. **b** Network architecture. EHR-M-GAN contains two key components—Dual-VAE and Coupled Recurrent Network (CRN). Step 1: Dual-VAE is first pretrained for mapping heterogeneous data ( $x^c, x^d$ ) into shared latent representations ( $z^c, z^d$ ). Multiple objective loss constraints are used to bridge the domain/distribution gap, including ELBO loss, matching loss, contrastive loss, and semantic loss (for EHR-M-GAN<sub>cond</sub> only). Both encoders and decoders in the Dual-VAE are implemented with LSTMs. The training process for Step 1 is indicated in the Dual-VAE pretrain path (dashed purple line). Step 2: Then, a CRN is established as the generator based on the parallel bilateral LSTM block, which takes the random noise vectors ( $u^c, u^d$ ) as inputs (see the Coupled generation path). Step 3: The synthetic latent representations ( $\hat{z}^c, \hat{z}^d$ ) provided by CRN are then decoded into synthetic samples ( $\hat{x}^c, \hat{x}^d$ ) using the pretrained decoder in Dual-VAE, which is indicated in the Decoding path (solid red line). Step 4: Finally, the adversarial loss is derived from the LSTM-based discriminators and backpropagated to update the network, which is indicated in the Adversarial training path (dotted black line). **c** Evaluation pipeline. The pipeline includes metrics for evaluating the fidelity of both continuous-valued and discrete-valued timeseries, and the correlations within the mixed-type data. Also, downstream task (in **d**) is performed for evaluating the application of synthetic data in realistic clinical use case. Finally, membership inference attack and differential privacy are used to evaluate our model's privacy risk empirically. **d** Prediction example. Data within 24-hours prior to the patient's endpoints in the ICU (discharge or mortality) is extracted. Both the observation window and prediction window are fixed as 12 h. The classification task is to use patients' continuous-valued physiological measurements within the observation window as input, to predict the forthcoming discrete-valued medical intervention status in the prediction window. The four outcomes of the intervention status can be categorized as follows: Stay On: The intervention begins with on and stays on within the prediction window; Onset: The intervention begins with off and is turned on within the prediction window; Switch off: The intervention begins with on and is stopped within the prediction window; Stay Off: The intervention begins with off and stays off within the prediction window.

there is still a lack of a generative approach for joint modeling of the mixed-type timeseries with heterogeneous nature. Second, although mixed-type clinical timeseries differ in syntax and distributions, they are highly correlated and inform one another of the underlying health of an individual<sup>33–35</sup>. It is therefore important to capture the temporal correlations between them when generating the synthetic EHR data. For example, the medications (documented in the form of discrete data) prescribed to patients are based on measurements of patients' physiological status (presented as continuous-valued signals). Concurrently, the efficacy of the medical treatments, affect the patient's physiological condition directly. It is therefore critical to accurately capture the temporal correlation between the mixed-type patient trajectories simultaneously to improve clinical decision support.

To address the aforementioned limitations, for the first time, we propose a GAN framework for simultaneously synthesizing mixed-type longitudinal EHR data (denoted as EHR-M-GAN thereafter). Specifically, we focus on generating timeseries in the critical care setting, where the intensive care units (ICU) patients are continuously and closely monitored (see Fig. 1a). 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 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 (see Fig. 1b). 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 data, a sequentially coupled generator that is built upon a *coupled recurrent network* (CRN) is employed. In addition, a conditional version of our model—EHR-M-GAN<sub>cond</sub>—is also implemented, which is capable of synthesizing condition-specific EHR patient data, such as those result in *ICU mortality* or *hospital readmission*. The code of our proposed work is publicly available on GitHub.

- Evaluations are performed based on three publicly available ICU datasets: MIMIC-III<sup>36</sup>, eICU<sup>37</sup>, and HiRID<sup>38</sup> 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.
- 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<sup>39</sup>. We then further evaluate the performance of EHR-M-GAN under the framework of differential privacy for its application in downstream task<sup>40</sup>.

## RESULTS

### Evaluation metrics

Evaluating GAN models is a notoriously challenging task. Advantages and pitfalls of commonly used evaluation metrics for GANs are discussed in ref. <sup>41</sup>. In this work, a systematic evaluation framework is adopted to assess the quality of synthetic patient EHRs with respect to its *fidelity*, *correlation*, *utility*, and *privacy* (see Table 1). 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 feature-level statistics between the real and synthetic data (*Patient trajectories*), and assessing the indistinguishability of the synthetic data to the true data (i.e., *Discriminative score*). Secondly, we evaluate to which extent our model can reconstruct the interdependency between different features (*Pearson pairwise correlations*), and the temporal dynamics in the patient trajectories (*Autocorrelation function*), by using a set of qualitative and quantitative metrics. Thirdly, we introduce data augmentation by incorporating synthesized 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 (i.e., the utility of the synthetic data). 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. 1c and Table 1 for the evaluation pipeline).

### Maximum mean discrepancy

To measure the similarity between the continuous-valued synthetic data and the real data, maximum mean discrepancy (MMD) is used. MMD can assess whether two sets of samples are from the same distributions, and in our case, one from the true

data  $x$  and one from synthetic data  $x'$  generated by GANs. To calculate the statistics, a kernel function  $K: X \times X' \rightarrow \mathbb{R}$  is used to quantify the similarity between the two distributions. In this study, a sum of Gaussian kernel sets is adopted following the implementations in ref. <sup>42</sup>, which can be expressed as:

$$K(\mathbf{x}, \mathbf{x}') = \sum_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_F^2}{\sigma_i^2}\right) \quad (1)$$

where  $\sigma_i$  is the value of the  $i$ -th selected bandwidth for calculating MMD. As in our study, the real and synthetic samples are multivariate timeseries aligned along the fixed time axis (i.e., 24 data points per patient), we therefore handle these multivariate timeseries as matrices and use the kernel function to calculate the Frobenius norm ( $\|\cdot\|_F$ ) between them<sup>25</sup>.

Finally, given samples  $\{\mathbf{x}_i\}_{i=1}^N$  from real distributions, and samples  $\{\mathbf{x}'_j\}_{j=1}^M$  from the synthetic distributions (with  $N$  and  $M$  denoting the corresponding sample sizes), the estimation of MMD can be defined as:

$$\widehat{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n K(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{j=1}^m K(\mathbf{x}_i, \mathbf{x}'_j) + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m K(\mathbf{x}'_i, \mathbf{x}'_j) \quad (2)$$

It can be inferred from the Eq. (2) that higher similarity between the two distributions leads to the lower MMD value, with the lower bound value zero indicating that the two distributions are identical.

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<sub>cond</sub>—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<sub>cond</sub> produce smaller MMD values when compared to their variants. Using MIMIC-III as an example, compared with the basic model GAN<sub>VAE</sub>, by integrating the shared latent space learning using *dual-VAE* under multiple loss constraints, the performance of GAN<sub>SL</sub> significantly improves ( $p$ -value < 0.05; Unpaired  $t$ -test with a significance level of 0.05 is used throughout the paper unless specified otherwise). By further building the sequentially coupled generator and exploiting the information within mixed-type data, the MMD of EHR-M-GAN shows a nearly 24% improvement over GAN<sub>VAE</sub>. When synthesizing mixed-type timeseries using the unified network, the performance of GAN<sub>Unified</sub> for generating continuous-valued timeseries lags behind the proposed EHR-M-GAN. It therefore can 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<sub>Unified</sub> 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. To test the probability distributions between the real and synthetic binary features, the Bernoulli success probability  $p \in [0, 1]$  is calculated for the discrete-valued timeseries, and is visualized through scatterplot. As a sanity check, it investigates if the probability of the medical intervention being active at the given timestamps is matched between the real data ( $x$ -axis) and synthetic data ( $y$ -axis). The correlation coefficients (CCs) and root-mean-square errors (RMSEs) are also adopted<sup>43</sup> based on the Bernoulli success probabilities to quantitatively measure the distribution divergence between real and synthetic data.

**Table 1.** Summary of the evaluation protocol in this study.

Evaluation metric	Data type	Definition	Better performance
Fidelity	Continuous	A kernel-based statistic is calculated to determine whether the real and synthetic data are from the same distribution	Lower value
	Discrete	The Bernoulli success probability of each feature dimension (i.e., the probability of whether the treatment is active) at the given timestamp is calculated. Probabilities from real and synthetic data are represented on the x and y axis in a single plot to compare the consistency	Scatters closer to the diagonal line (lower RMSE and CC)
Correlation	Continuous and discrete	A classifier is trained post hoc to tell the difference between the real and synthetic data with its accuracy calculated	Lower accuracy
	Continuous and discrete	The mean and standard deviation per time point of real and synthetic patient trajectories are compared and visualized	Similar distributions between the real and synthetic data
	Continuous and discrete	The correlation between different features is calculated and visualized in a heatmap for both real and synthetic data	Heatmaps corresponding to real and synthetic data more similar (lower CorAcc and $\mu_{abs}$ )
	Continuous and discrete	The correlation between the timeseries and its lagged version is calculated and visualized as an ACF curve for both real and synthetic data	ACF curves corresponding to real and synthetic data more similar
Utility	Continuous and discrete	The downstream classifier is trained which uses synthetic data as training set, and (hold-out) real data as test set. The result is compared with TRTR to see whether it can maintain the same	Higher AUROCs (with TRTR as baselines)
	Continuous and discrete	The downstream classifier is trained which uses real data and synthetic data as training set, and (hold-out) real data as test set. The result is compared with TRTR to see whether the performance can be improved	Higher AUROCs (with TRTR as baselines)
Privacy	Continuous and discrete	A threat model is trained under the black-box setting to determine whether a record is used for training GANs. This quantifies the risk of sensitive information from real data being revealed by synthetic data	Lower accuracy or recall
	Continuous and discrete	The downstream classifier is trained with differential privacy guarantee. The result is compared with TRTR to see whether it can maintain the same	Higher AUROCs (with TRTR as baselines)

A comprehensive set of evaluation metrics are used to test the *Fidelity*, *Correlation*, *Utility*, and *Privacy* of the synthetic EHR data. Definitions of evaluation metrics for corresponding data types are explained. The last column illustrates when the corresponding evaluation metric indicates better performance.

**Table 2.** Maximum mean discrepancy (MMD) of continuous-valued synthetic data.

	C-RNN-GAN	R(C)GAN	TimeGAN	GAN <sub>Unified</sub>	GAN <sub>VAE</sub>	GAN <sub>SL</sub>	EHR-M-GAN	EHR-M-GAN <sub>cond</sub>
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	<b>0.692 ± 0.034</b>	<b>0.604 ± 0.027</b>
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	<b>0.651 ± 0.026</b>	<b>0.540 ± 0.018</b>
HiRID	0.982 ± 0.017	0.865 ± 0.020	0.470 ± 0.024	0.518 ± 0.030	0.532 ± 0.035	0.508 ± 0.028	<b>0.428 ± 0.015</b>	<b>0.389 ± 0.024</b>

Lower values of MMD indicate models that can better learn the distribution of the real data. Bolded values denote best scores.

As shown in Fig. 2 (see Supplementary Figs. 4 and 5 for results on eICU and HiRID datasets), the optimal results are provided by EHR-M-GAN and EHR-M-GAN<sub>cond</sub>. 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 visualization results, which are close to the highest mark (EHR-M-GAN: RMSE = 0.0095, CC = 0.9973). Similar to the results in MMD, the dimensional-wise distributions are better captured when modules such as *dual-VAE* and *sequentially coupled generator* are introduced in EHR-M-GAN. GAN<sub>Unified</sub> 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. The result of EHR-M-GAN shows its superiority in explicitly capturing each dimension of the discrete-valued sequences. This indicates that the proposed EHR-M-GAN mitigates the challenge of generating discrete-valued features in traditional GANs by learning the shared latent representations using *dual-VAE*.

### Patient trajectories

We compare the distribution of patient trajectories per timepoint between the real data and synthetic data generated by EHR-M-GAN for the MIMIC-III dataset. Five commonly measured vital sign and laboratory 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. 3. It can be inferred that the proposed model can accurately capture the statistical distribution (mean and standard deviation) of both continuous-valued and discrete-valued features. The temporal dynamics are well-preserved in the synthetic timeseries. For example, the variance of *Oxygen Saturation* gradually increases towards the ICU endpoints in the real data, and is closely reflected in the synthetic timeseries. Furthermore, EHR-M-GAN<sub>cond</sub> shows superior performance as it can generate correct trajectories with specific patient conditions (see Supplementary Note 4).

### Discriminative score

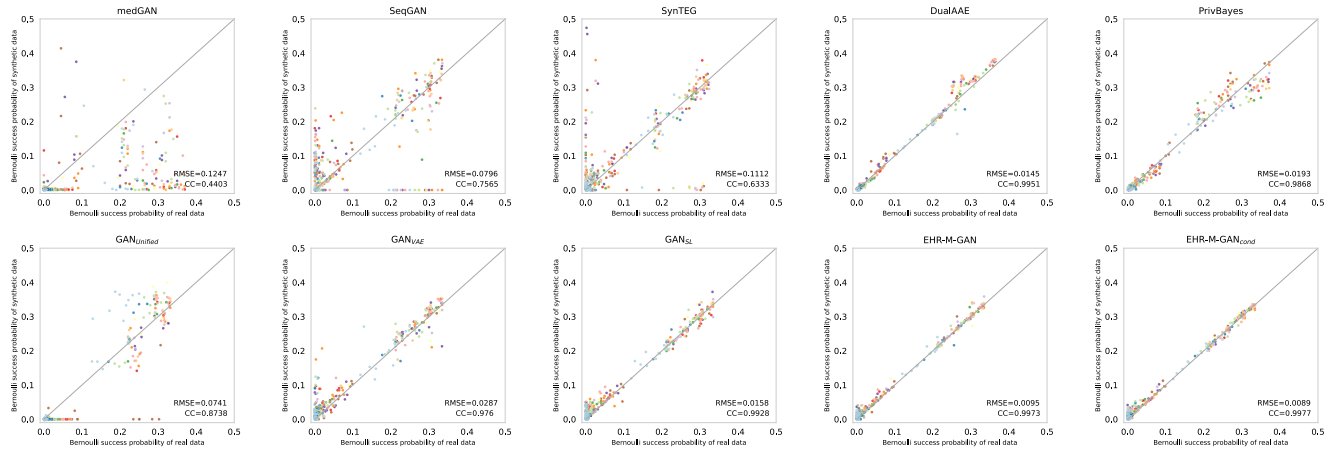
For both continuous-valued and discrete-valued data, the discriminative score is measured as the accuracy of a discriminator trained post-hoc to separate real from generated samples. Synthetic data are generated with the same amount of the hold-out test set from the original data, and are labeled as *synthetic* and *real* correspondingly to train the binary classifier. In this study, the classifier (critic) is implemented with a single-layered Bi-directional Long Short-Term Memory (Bi-LSTM) model, with its parameters randomly initialized (as opposed to critic built upon representations from the trained generative model<sup>28</sup>). The critic trained from the supervised learning task can be used to characterize the temporal correlations across the patient EHR timeseries.

As indicated from the results in Table 3, it appears that EHR-M-GAN and EHR-M-GAN<sub>cond</sub> can produce synthetic data that are less distinguishable from real data than the benchmarked models. Especially for EHR-M-GAN<sub>cond</sub>, 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 discriminative score of 0.813 on the MIMIC-III dataset, which has a 4% statistically significant improvement over the best performing benchmark ( $p < 0.05$ ). The overall discriminative scores produced by PrivBayes on three ICU databases are comparable with the GAN models such as SynTEG and DualAAE. For continuous-valued timeseries generation, the discriminative score of TimeGAN on HiRID dataset outperforms the other models as well as EHR-M-GAN, though not statistically significant ( $p = 0.4374$ ). By leveraging the additional information from the conditional inputs, EHR-M-GAN<sub>cond</sub> can provide significantly better result than TimeGAN ( $p < 0.05$ ).

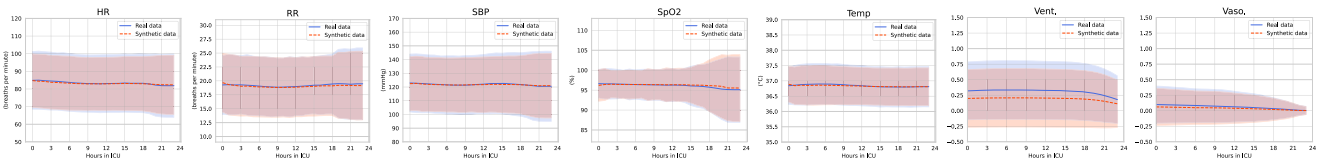
The ablation study has proved the effectiveness of EHR-M-GAN for generating high quality EHR timeseries. The shared latent space representation learning in the *dual-VAE* (i.e., GAN<sub>SL</sub>) have shown remarkable success compared with GAN<sub>VAE</sub>, which generates the latent embeddings based on separate VAEs. The *sequentially coupled generator* further improves the model by capturing the dynamics between mixed data types. Further compared with GAN<sub>Unified</sub> which models the mixed-type data in a unified network, our proposed model enables effective learning for the marginal distributions from each data type. In addition, as shown in Supplementary Table 9, EHR-M-GAN can provide more realistic synthetic samples than the *dual-VAE* module alone (see Supplementary Note 3 for details).

### Interdependency characteristics

In this section, we first employ Pearson pairwise correlation (PPC), which ranges from  $-1$  to  $1$ , to evaluate how closely the synthetic data can model the correlations between continuous-valued and discrete-valued timeseries. Timestamps of the patient trajectories are extracted with every 3 h interval in a total of 24 h ICU stay, to explore the temporal dependencies within different variables. To further quantitatively measure the difference between heatmaps generated from real and synthetic samples, we calculate the mean value of the absolute difference between the two PCC matrices



**Fig. 2 Scatterplot of the dimension-wise probability test on MIMIC-III dataset.** Dimension-wise probability calculates the Bernoulli success probability of each dimension, i.e., the probability of the treatment being active at a particular time. The x-axis and y-axis represent dimension-wise probability for the real data and synthetic data generated from different models, respectively. The same color indicates the same treatment (but with varying timestamps). The optimal performance appears along the diagonal line. The corresponding CCs ( $[0, 1]$ ), the higher the better) and RMSEs ( $[0, +\infty)$ ), the lower the better) are also calculated to quantify the probability distribution similarities between the real and synthetic EHRs timeseries. Dimension-wise probability plot for eICU and HiRID dataset can be found in Supplementary Note 4.



**Fig. 3 Comparison of the distribution of values at each timepoint (mean and standard deviation) between real and synthetic patient trajectory produced by EHR-M-GAN.** Multivariate timeseries 24 h before patients' ICU endpoints are generated, including *Heart Rate*, *Respiratory Rate*, *Systolic Blood Pressure*, *Oxygen Saturation*, *Temperature*, *Mechanical Ventilation* and *Vasopressor*. The mean value of the real/synthetic feature at each timepoint is plotted by the solid/dotted line, with the shaded area indicating  $\pm 1$  standard deviation. For *Mechanical Ventilation* and *Vasopressor*, the y-axis indicates the probability distribution of such intervention being applied ("On") at a given time. The synthetic patient trajectories generated by EHR-M-GAN<sub>cond</sub> under different conditions can be found in Supplementary Note 4.

( $\mu_{abs}$ ). We also adopted correlation accuracy (*CorAcc*)<sup>44</sup>, which quantifies the similarity of two heatmaps within the range of 0 to 1. We discretize the correlation coefficients into 6 correlation levels: *strong negative* ( $[-1, -0.5)$ ), *middle negative* ( $[-0.5, -0.3)$ ), *low negative* ( $[-0.3, -0.1)$ ), *no correlation* ( $[-0.1, 0.1)$ ), *low positive* ( $[0.1, 0.3)$ ), *middle positive* ( $[0.3, 0.5)$ ), and *strong positive* ( $[0.5, 1)$ ). Then, *CorAcc* can be calculated as the percentage of pairs where the real and synthetic data is assigned to the same correlation level.

As observed (see Fig. 4), correlation trends over distinctive features are closely reflected by the synthetic data, with the quantitative measure *CorAcc* consistently exceed 0.8 on three critical care databases. It is also worth noticing that EHR-M-GAN can successfully recover temporal dependencies with a high granularity from real patient trajectories. For example, synchronized correlations across timestamps are observed between *Respiratory Rate* and *Heart Rate* in the MIMIC-III dataset. Such trends are preserved in synthetic data. 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. In summary, the proposed EHR-M-GAN can reconstruct the temporal dynamics and correlations between features in the real data, which is valuable for downstream ML-based classification and prediction applications.

Then, autocorrelation functions (ACF)<sup>45</sup> and the corresponding root-mean-square errors (RMSEs) are calculated to show how EHR-M-GAN can capture the temporal correlations among the timeseries. ACF measures the relationship between the timeseries and its lagged version. Supplementary Figs. 6–8 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 h 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. 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 behaviors can be reproduced by EHR-M-GAN, therefore they demonstrate that our model can effectively capture the temporal characteristics in the original timeseries.

### 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

**Table 3.** Discriminative score of synthetic data.

	Method	MIMIC-III	eICU	HiRID
Continuous-valued synthetic data	C-RNN-GAN	0.825 ± 0.013	0.876 ± 0.010	0.774 ± 0.022
	R(C)GAN	0.833 ± 0.028	0.850 ± 0.021	0.742 ± 0.016
	TimeGAN	0.763 ± 0.018	0.790 ± 0.013	<b>0.716 ± 0.024</b>
	GAN <sub>Unified</sub>	0.809 ± 0.023	0.863 ± 0.027	0.749 ± 0.014
	GAN <sub>VAE</sub>	0.842 ± 0.020	0.871 ± 0.014	0.802 ± 0.017
	GAN <sub>SL</sub>	0.786 ± 0.016	0.813 ± 0.023	0.752 ± 0.021
	EHR-M-GAN	<b>0.746 ± 0.018</b>	<b>0.776 ± 0.015</b>	0.724 ± 0.015
	EHR-M-GAN <sub>cond</sub>	<b>0.729 ± 0.025</b>	<b>0.745 ± 0.017</b>	<b>0.693 ± 0.012</b>
	Discrete-valued synthetic data	medGAN	0.903 ± 0.027	0.915 ± 0.034
seqGAN		0.937 ± 0.025	0.924 ± 0.023	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	0.883 ± 0.034	0.832 ± 0.017
GAN <sub>Unified</sub>		0.890 ± 0.022	0.907 ± 0.026	0.849 ± 0.015
GAN <sub>VAE</sub>		0.862 ± 0.024	0.881 ± 0.029	0.824 ± 0.018
GAN <sub>SL</sub>		0.829 ± 0.032	0.844 ± 0.028	0.816 ± 0.025
EHR-M-GAN		<b>0.813 ± 0.026</b>	<b>0.831 ± 0.024</b>	<b>0.802 ± 0.020</b>
EHR-M-GAN <sub>cond</sub>		<b>0.784 ± 0.024</b>	<b>0.803 ± 0.022</b>	<b>0.778 ± 0.019</b>

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. Bolded values denote best scores.

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<sub>cond</sub> in synthesizing high-fidelity synthetic data to further boost the performance of ML classifiers. In line with prior work<sup>46–48</sup>, 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 h 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 Fig. 5a, and the performance is assessed from two aspects (see Fig. 5b: (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* (TSTR), to show whether the performance of a classifier trained on synthetic data from EHR-M-GAN or EHR-M-GAN<sub>cond</sub> 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

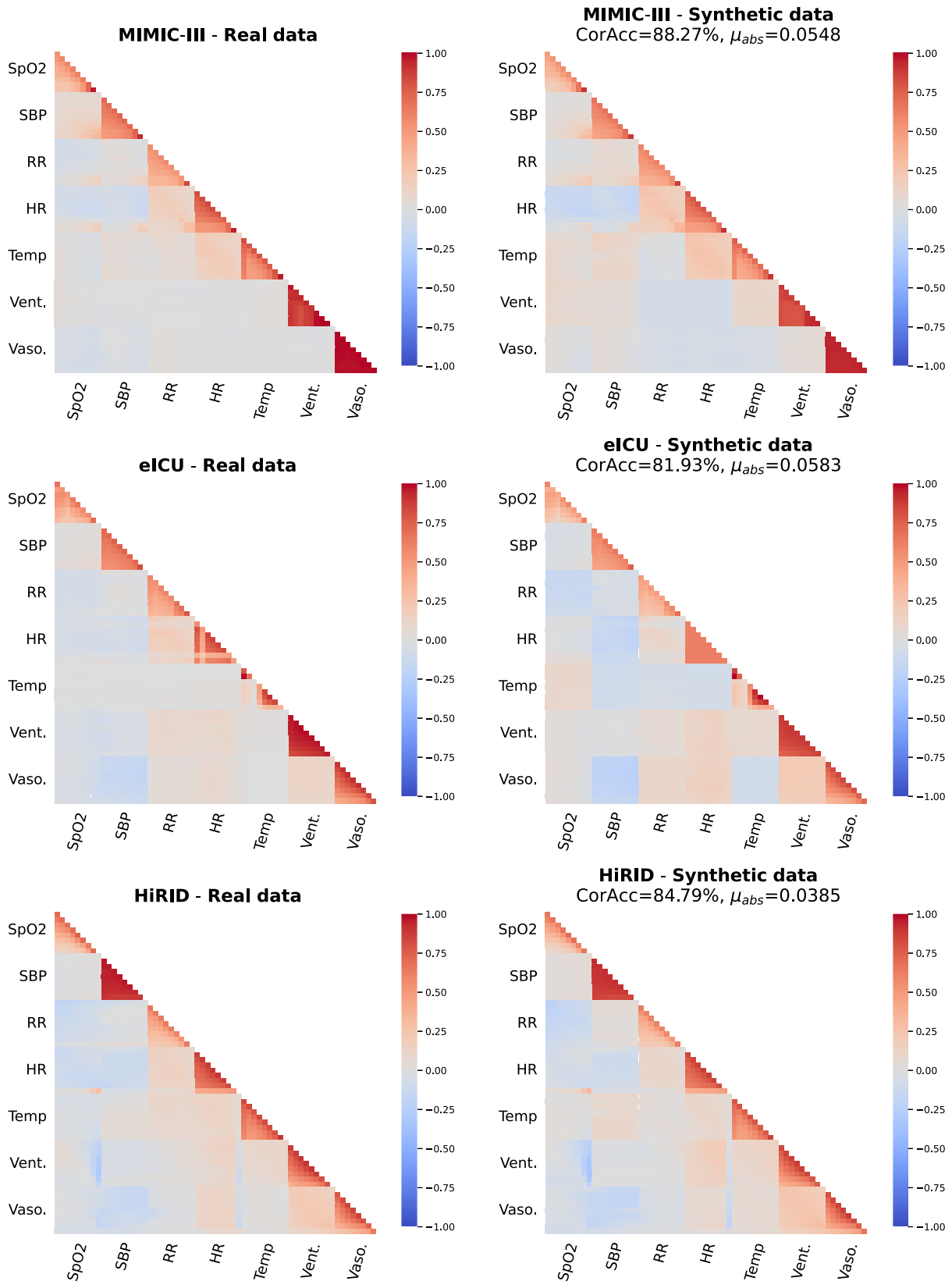
medical intervention (described as outcomes in this scenario) using continuous-valued physiological features (denoted as predictors), TSTR can also be used to assess data synthesizers' ability to capture the interdependencies between the mixed-type features. (ii) Data augmentation approach: As data augmentation is employed as a means of circumventing the issue caused by the under-resourced EHR data, here we explore whether synthetic data can be 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<sup>25,49</sup>. The augmentation ratio  $\alpha$  or  $\beta$  is applied on sub-train data  $A'_T$  or synthetic data  $B$ , in two different scenarios of TSRTR, respectively. Details are explained as follows (also see Fig. 5b 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<sub>cond</sub> can improve the training of downstream classifiers. **Ratio  $\alpha$**  indicates the portion of synthetic data (see Fig. 5b 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<sub>cond</sub> in TSRTR, we investigate the proportion of the real data  $A'_T$  (**ratio  $\beta$** ) required to maintain the same performance as in TRTR based on the entire synthetic dataset  $B$  (see Fig. 5b).

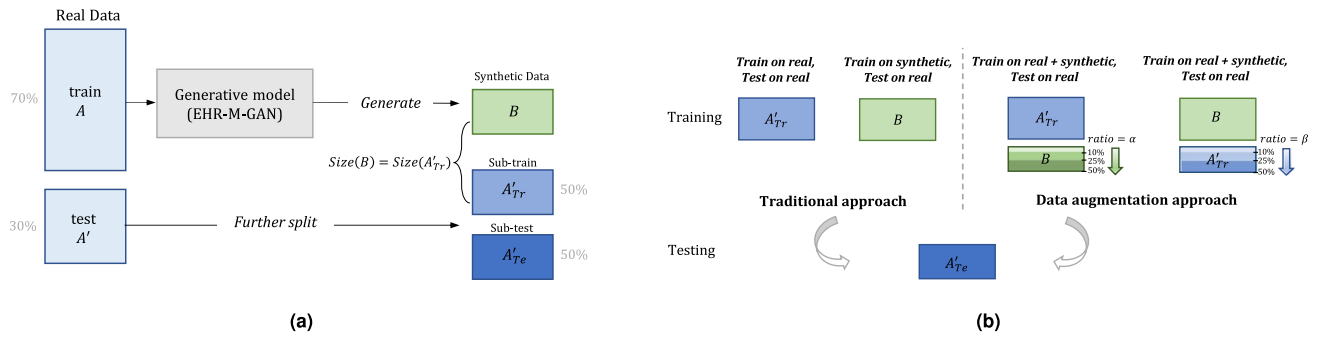
**Traditional approach.** Table 4 compares the classification performances of predicting forthcoming medical interventions in the ICUs under the experimental setting of TRTR and TSTR. It is expected that the optimal AUROCs are achieved by the classifiers that are trained on real data. In comparison, the classifiers trained on the synthetic data provided by proposed models can achieve similar performances. More specifically, synthetic data generated by EHR-M-GAN<sub>cond</sub> demonstrates better generalizability when compared with EHR-M-GAN in the downstream application, such as the task of predicting *mechanical ventilation* on the HiRID dataset.

Compared with the baseline models, the proposed EHR-M-GAN shows improved performance in TSTR, as it can model the distribution of mixed-type EHRs more accurately, while preserving the temporal correlations in the heterogeneous timeseries through the dependency learning components. The results indicate that interdependency between the mixed-type EHRs is weakly captured by GAN<sub>VAE</sub>, as the two streams of inputs are trained in parallel and separately. GAN<sub>Unified</sub> 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 TSTR.



**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). Besides the visual inspection, the similarity between two heatmaps are quantified by  $CorAcc$  and  $\mu_{abs}$ . These metrics indicate how well the synthetic data reconstruct 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 h over the total 24 h of ICU stay (ticks of the timestamps are omitted).





**Fig. 5 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  $A'_{Te}$  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<sub>cond</sub>) 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  $A'_{Te}$ . In traditional approach, results based on *Train on Real, Test on Real (TRTR)* and *Train on Synthetic, Test on Real (TSTR)* are compared to assess the generalizability 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  $A'_{Tr}$ .

**Table 4.** Downstream task evaluation.

Dataset	Treatments	Real data	GAN <sub>Unified</sub>	GAN <sub>VAE</sub>	GAN <sub>SL</sub>	EHR-M-GAN	EHR-M-GAN <sub>cond</sub>
MIMIC-III	Vent.	<b>0.894 ± 0.016</b>	0.724 ± 0.015	0.701 ± 0.018	0.728 ± 0.010	0.740 ± 0.009	0.823 ± 0.020
	Vaso.	<b>0.841 ± 0.009</b>	0.694 ± 0.012	0.651 ± 0.015	0.679 ± 0.009	0.725 ± 0.015	0.810 ± 0.019
eICU	Vent.	<b>0.868 ± 0.015</b>	0.697 ± 0.014	0.702 ± 0.009	0.718 ± 0.012	0.745 ± 0.008	0.795 ± 0.015
	Vaso.	<b>0.813 ± 0.018</b>	0.648 ± 0.011	0.657 ± 0.012	0.665 ± 0.014	0.706 ± 0.014	0.748 ± 0.017
HiRID	Vent.	<b>0.867 ± 0.012</b>	0.765 ± 0.014	0.747 ± 0.013	0.803 ± 0.008	0.825 ± 0.019	0.856 ± 0.033
	Vaso.	<b>0.878 ± 0.010</b>	0.754 ± 0.018	0.752 ± 0.020	0.779 ± 0.013	0.814 ± 0.015	0.844 ± 0.018

Downstream tasks are evaluated under the training scenarios of *Train on Real, Test on Real (TRTR)* and *Train on Synthetic, Test on Real (TSTR)*. 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*). Bolded values denote best scores.

**Table 5.** Downstream task evaluation with data augmentation ratio  $\alpha$ .

Dataset	Treatments	EHR-M-GAN			EHR-M-GAN <sub>cond</sub>		
		$\alpha = 10\%$	$\alpha = 25\%$	$\alpha = 50\%$	$\alpha = 10\%$	$\alpha = 25\%$	$\alpha = 50\%$
MIMIC-III	Vent.	0.828 ± 0.013	0.877 ± 0.014	0.912 ± 0.015 (↑)	0.845 ± 0.022	0.896 ± 0.013 (↑)	<b>0.923 ± 0.018</b> (↑)
	Vaso.	0.816 ± 0.015	0.834 ± 0.023	0.859 ± 0.013 (↑)	0.848 ± 0.012 (↑)	0.876 ± 0.017 (↑)	<b>0.896 ± 0.015</b> (↑)
eICU	Vent.	0.858 ± 0.008	0.862 ± 0.012	0.873 ± 0.014 (↑)	0.865 ± 0.009	0.879 ± 0.014 (↑)	<b>0.883 ± 0.016</b> (↑)
	Vaso.	0.798 ± 0.015	0.805 ± 0.020	0.821 ± 0.028 (↑)	0.813 ± 0.016 (↑)	0.834 ± 0.019 (↑)	<b>0.839 ± 0.014</b> (↑)
HiRID	Vent.	0.871 ± 0.025 (↑)	0.882 ± 0.021 (↑)	0.913 ± 0.019 (↑)	0.894 ± 0.015 (↑)	0.906 ± 0.018 (↑)	<b>0.923 ± 0.021</b> (↑)
	Vaso.	0.850 ± 0.016	0.874 ± 0.022	0.894 ± 0.018 (↑)	0.883 ± 0.017 (↑)	0.908 ± 0.024 (↑)	<b>0.913 ± 0.019</b> (↑)

Downstream tasks are evaluated under the training scenarios of *Train on Synthetic and Real, Test on Real (TSRTR)*. All data from sub-train data  $A'_{Tr}$  concatenated with  $\alpha$  of the synthetic data  $B$  (augmentation ratio  $\alpha = 10\%$ ,  $25\%$  or  $50\%$ ) is used as the training set. The upper arrow (↑) 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 ( $p \leq 0.05$ ). Bolded values denote best scores.

**Data augmentation approach (with ratio  $\alpha$ ).** The results in Table 5 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<sub>cond</sub>. 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 ( $p < 0.05$ ). Our experiment results have demonstrated that the proposed models can be used for data augmentation to

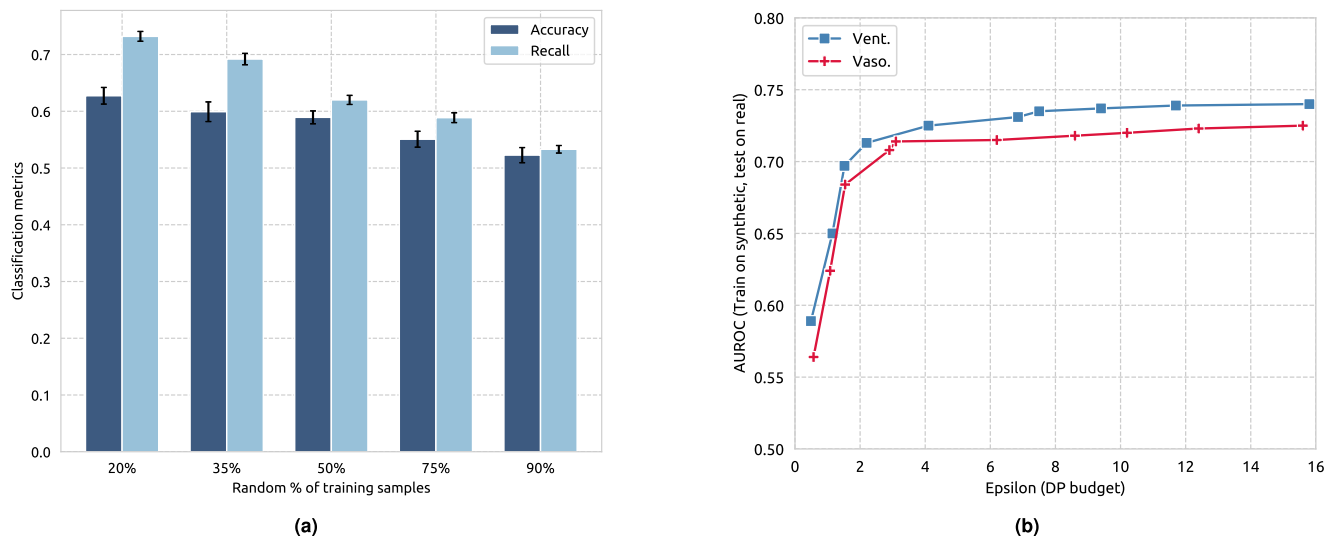
overcome the issue of data scarcity and subsequently improve the classifiers' performance.

**Data augmentation approach (with ratio  $\beta$ ).** On the other hand, as shown in Table 6, 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<sub>cond</sub>, the ratio needed for real data to maintain the comparable predictive power is further reduced to 25%, which

**Table 6.** Downstream task evaluation with data augmentation ratio  $\beta$ .

Dataset	Treatments	EHR-M-GAN			EHR-M-GAN <sub>cond</sub>		
		$\beta = 10\%$	$\beta = 25\%$	$\beta = 50\%$	$\beta = 10\%$	$\beta = 25\%$	$\beta = 50\%$
MIMIC-III	Vent.	0.757 ± 0.016	0.824 ± 0.010	0.885 ± 0.009	0.847 ± 0.017	0.903 ± 0.014 (↑)	<b>0.915 ± 0.009 (↑)</b>
	Vaso.	0.786 ± 0.019	0.810 ± 0.020	0.849 ± 0.017 (↑)	0.823 ± 0.014	0.851 ± 0.019 (↑)	<b>0.873 ± 0.017 (↑)</b>
eICU	Vent.	0.761 ± 0.011	0.822 ± 0.012	0.870 ± 0.019 (↑)	0.816 ± 0.016	0.845 ± 0.018	<b>0.872 ± 0.020 (↑)</b>
	Vaso.	0.742 ± 0.014	0.797 ± 0.013	0.846 ± 0.018 (↑)	0.785 ± 0.022	0.819 ± 0.021 (↑)	<b>0.834 ± 0.013 (↑)</b>
HiRID	Vent.	0.856 ± 0.012	0.879 ± 0.019 (↑)	0.895 ± 0.021 (↑)	0.874 ± 0.016 (↑)	0.896 ± 0.018 (↑)	<b>0.904 ± 0.012 (↑)</b>
	Vaso.	0.826 ± 0.024	0.859 ± 0.013	0.893 ± 0.018 (↑)	0.865 ± 0.025	0.897 ± 0.021 (↑)	<b>0.914 ± 0.018 (↑)</b>

Downstream tasks are evaluated under the training scenarios of *Train on Synthetic and Real, Test on Real (TSRTR)*. All data from synthetic data  $B$  concatenated with  $\beta$  of the sub-train data  $A'_T$  (augmentation ratio  $\beta = 10\%$ ,  $25\%$  or  $50\%$ ) is used as the training set. The upper arrow (↑) 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 ( $p \leq 0.05$ ). Bolded values denote best scores.



**Fig. 6 Privacy risk evaluation of EHR-M-GAN on MIMIC-III dataset. a** Membership inference attack against EHR-M-GAN vs. the percentage of the training data. Accuracy and recall are used to evaluate the success rate of such attacks. Lower accuracy or recall indicates less privacy information disclosed by the attacker from the generative model (0.5 can be seen as the *random guess* baseline where strong privacy guarantees are provided by GANs). Recall indicates the ratio of samples that are successfully claimed by the attacker among all the real data that are used in training GAN models. Error bars represent the standard error. **b** Differential privacy. Performance of medical intervention prediction tasks, under various differential privacy (DP) budgets, measured by Macro-AUROC.

equivalently indicates a 75% reduction of sample size required in real data collection. Overall, results presented in Table 6 demonstrate that by exploiting only a limited ratio of the real data, EHR-M-GAN and EHR-M-GAN<sub>cond</sub> 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. Generative models can overcome the explicit one-to-one mapping towards the underlying original data in contrast to data anonymization. However, GAN could potentially raise privacy concerns of information leakage if they simply “memorize” the training data, or synthesize samples nearly identical to the real samples (often due to mode collapse). 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<sup>50,51</sup>. The threat model is implemented under the membership inference for GANs in the *black-box settings*<sup>50</sup>. 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 Fig. 6a). 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. If EHR-M-GAN simply “memorizes” the training data and can only generate synthetic samples (nearly) identical to the real samples, it would be straightforward for the adversary to identify samples that are used as training data. 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 Fig. 6a, 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. In other words, patient samples used in EHR-M-GAN's training are not recoverable by the threat model. 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<sup>52</sup>.

The concept differential privacy (DP)<sup>53</sup>, 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:  $P[\mathcal{M}(D) \in S] \leq e^\epsilon P[\mathcal{M}(D') \in S] + \delta$ . In our case,  $\mathcal{M}(\cdot)$  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 have 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. We then perform the same 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 6b 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 Fig. 6b, the performance of the downstream tasks operated based on the synthetic data generated by EHR-M-GAN improves as the DP budget relaxes ( $\epsilon$  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 *TRTR*) even when the  $\epsilon$  decrease to 4, which is an empirically reasonable value for implementing DP in practice<sup>54</sup>. 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

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. A comprehensive list of evaluation metrics is introduced for the systematic assessment, in terms of the fidelity, correlation, utility, and privacy of the synthesis model. First, both EHR-M-GAN and its conditional version, EHR-M-GAN<sub>condr</sub>, demonstrate consistent improvements against the state-of-the-art benchmark GANs in synthesizing timeseries data with high-fidelity. This indicates that the distributional characteristics of the EHR timeseries can be well-preserved in the synthetic data provided by EHR-M-GAN, therefore ensuring its usability during clinical data sharing. Second, 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 and

correlation between features. By accurately reconstructing the interdependencies and complex clinical relationships between features, downstream studies such as association analysis and outcome prediction can be supported. Notably, the proposed models also outperform the GAN variants that allow mixed-type inputs in the ablation study, indicating that the components in EHR-M-GAN are effective in synthesizing mixed-type timeseries with high fidelity, while successfully reconstructing the interdependencies between them. Then, 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 by data augmentation. Lastly, the generative capability of the proposed model avoids the "one-to-one" mapping to the original data, and enables the collaborative uses and sharing of EHRs by creating realistic novel samples. The assessment of privacy risks further demonstrates the synthetic data provided by EHR-M-GAN can preserve the sensitive information in patient records while maintaining an acceptable level of data utility.

The results in our study have several notable implications with respect to the synthesis of EHR data. First, as the proposed model can be used to provide synthetic longitudinal EHRs for various data types while preserving their underlying correlations, it is now feasible to use the synthesized data to improve the performance of ML models for downstream applications such as the prediction of next intervention, or understanding the disease dynamics and patient phenotyping, based on both the continuous and discrete components of EHR timeseries<sup>55,56</sup>. Second, the experimental results indicate that the quality of the synthetic EHR data can be improved by the integration of mixed-type information, in contrast to the benchmarks that utilize single-type data for learning. This also enables us to mimic how information is presented in clinical practices. Furthermore, we can generate condition/outcome-specific patient trajectories along with corresponding interventions, to facilitate clinical prediction and decision-making. Third, though facing the privacy-utility tradeoffs, the synthetic EHRs data provided by the proposed model leads to negligible privacy risks under the membership inference attacks. This paves the way for a series of applications in clinical research, including but not limited to, enabling the development of ML models by accessing the synthetic data, overcoming the paucity of medical data and improving the robustness of ML algorithms through data augmentation.

Due to the heterogeneous nature of EHR data, besides the ICU setting in our empirical evaluation, there are needs for synthesizing mixed-type EHR timeseries in various clinical scenarios. For example, patients' encounters in hospitals are documented as structured EHRs recorded in the temporal order. Each visit is typically associated with the corresponding medical events presented in the form of discrete-valued ICD codes<sup>27</sup>, and continuous-valued measurements. These mixed-type EHR timeseries capture a patient's health status and better align with clinical decision-making process than those using the single-type data alone. Therefore, developing GANs targeting mixed-type EHRs generation have the potential to pave the way for complex deep-learning systems that are capable of integrating information from various sources. However, it is worth noting that the validation of our proposed model is based on critical care settings with limited feature dimensions, can only serve as a proof of concept. When extending the proposed model to other clinical settings, such as synthesizing ICD codes with hundreds or thousands of feature dimensions<sup>27</sup>, the scalability and utility of our proposed model when dealing with the enlarged, sparse feature space needs further investigation.

There are limitations in the current work. First, data curation strategies on clinical timeseries, including truncating, smoothing and imputation, are applied before the EHR timeseries are used for

the training of generative models. As during the data preprocessing, we first extract the timeseries with a fixed duration (i.e., 24 h before the ICU clinical endpoints), and then hourly aggregate patients' physiological and intervention signals based on their mean statistics, followed by completing the missing value in the timeseries through the "Simple Imputation" approach<sup>57</sup>. Although these preprocessing steps are commonly used in clinical research under the critical care settings<sup>46</sup>, the proposed model cannot model the irregular time intervals between signals nor missing values within the timeseries. 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<sup>27</sup>. 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. Also, 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<sup>58</sup>. Furthermore, as evaluations are performed based on clinical timeseries with a fixed length, no comparisons are made between the model's scalability when dealing with timeseries with varying lengths. Recent studies have found the quality of the synthetic longitudinal data degenerates over time, also called as the "drift problem"<sup>28</sup>. Such problems when dealing with long sequences should be recognized and mitigated with techniques such as conditional fuzzing and regularization methods<sup>28</sup>, in both the generation and evaluation steps.

The evaluation of GANs is still a challenging task. Recent findings have suggested that systematical assessment for EHR synthesizers is critical before their applications in different use cases<sup>59</sup>. In this study, a comprehensive evaluation list is provided with regards to the fidelity, correlation, utility and privacy of the synthesis models. It is also worth noting that evaluation metrics should be properly chosen and implemented based on the purpose of the task, otherwise may lead to biased results. For example, recent findings<sup>28</sup> have reported that the traditional implementation of the discriminative score which trains the critic using the randomly initialized parameters, though widely used<sup>29</sup>, may lead to unreliable results. Improvement has been made to this evaluation metric for a more robust assessment, where the parameters of the trained generative model can be used for the critic's initialization.

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 personalized 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<sup>60,61</sup>. Ultimately, by constructing the "synthetic twin" of patients using GANs, the synthesis tool can become more generalizable 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-GAN<sub>cond</sub> can generate realistic longitudinal EHR timeseries with mixed data types. By providing synthetic EHR data better mimicking the nature of clinical decision-making, 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 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

### Dataset description

The following three publicly accessible ICU datasets with de-identified EHR data are used for evaluating the performance of EHR-M-GAN in generating the longitudinal data:

- **MIMIC-III** (Medical Information Mart for Intensive Care)<sup>36</sup>—a freely accessible database that comprises EHR data associated with approximately 60,000 ICU admitted patients and 312 million observations to Beth Israel Deaconess Medical Center.
- **eICU** (eICU Collaborative Research Database)<sup>37</sup>—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)<sup>38</sup>—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 featurize the patient trajectories as the combination of continuous-valued physiological timeseries (such as heart rate, oxygen saturation, and measurements from blood gas tests) and discrete-valued medical intervention timeseries (such as the usage of therapeutic devices or intravenous medications). Temporal trajectories **24 h prior to patients' ICU endpoints** (discharge or death) are extracted for the three critical care databases. Data are preprocessed following an open-source framework—MIMIC-Extract<sup>46</sup>, where the patients' physiological and intervention signals are hourly aggregated for denser representations. Details on data curation, including the cohort selection criteria, full list of features, and imputation method, are explained in Supplementary Note 2. Overall, the summarizing statistics of the finalized cohorts for three databases are shown in Table 7.

### Problem formulation

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

### Challenges in mixed-type timeseries generation

There are two main challenges when synthesizing mixed-type EHR timeseries. First, GANs have serious limitations on the type of data they can model<sup>30</sup>. Specifically, as GANs require generators and

**Table 7.** Summary of the cohorts after preprocessing on three critical care databases.

	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	1870 (6.59%)
					Hospital mortality	911 (3.21%)
					30-day readmission	1122 (3.95%)
					No 30-day readmission	24,441 (86.22%)
eICU	99,015	99,015	55	19	ICU mortality	4500 (4.54%)
					Hospital mortality	3291 (3.32%)
					Hospital discharge	91,224 (92.13%)
HiRID	14,129	14,129	50	39	ICU mortality	1266 (8.96%)
					ICU discharge	12,963 (91.74%)

Number of patients and ICU admissions, as well as the dimensions of continuous-valued and discrete-valued variables, are provided for each dataset. Temporal trajectories 24 h prior to patients' ICU endpoints are extracted for the three critical care databases. 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<sub>cond</sub> and the corresponding counts for each class are also listed.

discriminators to be both fully differentiable, generating discrete-valued timeseries using traditional GANs architectures would raise problems during backpropagation as no direct gradient can be provided<sup>31,32</sup>. Therefore, it is intrinsically difficult to model the underlying joint distribution of mixed data type timeseries using a single unified framework. Second, as the mixed-type timeseries are correlated (such as correlations between ICU patients' physiological signals and treatment status in the critical care setting), it is therefore important to model the interdependencies among heterogeneous types of timeseries.

**Intuition behind EHR-M-GAN**

First, to jointly model the distribution of continuous-valued and discrete-valued timeseries using GANs, we build the generative model based on the latent space encoded by VAE networks. Instead of directly synthesizing discrete-valued timeseries that deactivate the backpropagation in GANs, the generator first synthesizes latent representations that allow the direct gradient in the network, therefore satisfying the prerequisite for GANs architecture to be fully differentiable. The synthetic latent representations for both types of data can be further decoded into raw timeseries using the decoders in VAEs.

Even though the aforementioned network architectures enable the joint modeling of mixed-type data distribution, it still lacks the capability of capturing the inter-dependencies in heterogeneous data. In order to address the second issue, we devised *dual-VAE* module for pretraining step and *sequentially coupled generator* module for generation step. The *dual-VAE* incorporates multiple loss constraints, which were previously adopted in domains such as self-supervised learning (SSL), timeseries representation learning, and domain adaptation, to extract useful hierarchical representations from heterogeneous but correlated data types. The *sequentially coupled generator* module replaces the traditional LSTM cell with the bilateral LSTM (BLSTM) cell we propose, where the "communication" of the two types of information are introduced into the networks. Therefore, the temporal dynamics between the mixed-type data can be preserved during the iteration.

**Network architecture**

As illustrated above, EHR-M-GAN can be factorized into two key components (see Fig. 1b): (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. 1b, during the *pretrain* stage, both continuous-valued and discrete-valued temporal trajectories are first jointly mapped into a shared latent space using the *dual-VAE* component (Step 1). Then, the *sequentially coupled generator* in EHR-M-GAN produces the synthetic latent representations (Step 2), which further can be recovered into features in the observational space by the pretrained decoders in the *dual-VAE* (Step 3). Finally, the adversarial loss is provided based on discriminative results and backpropagated to update the network (Step 4). 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  $\mathbf{x}_{1:T_i}^c$  and  $\mathbf{x}_{1:T_i}^d$  can be encoded into the *same* latent space  $\mathcal{H}^S \subset \mathbb{R}^{|S|}$ , where  $|S|$  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, which exploits two VAE networks to encode both continuous and discrete multivariate timeseries into dense representations within  $\mathcal{H}^S$  based on multiple constraints.

Supplementary Fig. 2 diagrams the details of the proposed *dual-VAE* framework for learning the shared latent representations. We start with training two encoders, i.e.,  $Enc^c: \phi_{T \times \mathcal{X}^c} \rightarrow \phi_{T \times \mathcal{H}^S}$  and  $Enc^d: \phi_{T \times \mathcal{X}^d} \rightarrow \phi_{T \times \mathcal{H}^S}$ , with the embedding functions:

$$\mathbf{z}_{1:T}^c = Enc^c(\mathbf{x}_{1:T}^c) \quad \mathbf{z}_{1:T}^d = Enc^d(\mathbf{x}_{1:T}^d) \tag{3}$$

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

$$\tilde{\mathbf{x}}_{1:T}^c = Dec^c(\mathbf{z}_{1:T}^c) \quad \tilde{\mathbf{x}}_{1:T}^d = Dec^d(\mathbf{z}_{1:T}^d) \tag{4}$$

Also, to incentivize dual-VAE to better bridge the gap between domains of mixed-type timeseries, we enforce a weight-sharing

constraint<sup>62,63</sup> within specific layers of both the encoders pairs and the decoders pairs (See Supplementary Note 1 for details).

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<sub>cond</sub>. Among these losses, *ELBO loss* ensures that the mixed-type timeseries can be successfully reconstructed after being encoded into latent representations. The *matching loss* ensures that heterogeneous types of features from a single patient share contexts during representation learning (instance-wise). The goal of *contrastive loss* is to ensure that patients with similar trajectories stay close to each other in the latent space (population-wise). And *semantic loss* used in EHR-M-GAN<sub>cond</sub> encourages patients with the same conditional labels (e.g., outcomes) to share similar latent representations. 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 embeddings, where features can then be reconstructed by sampling from that space. The reparameterization tricks enable differentiable stochastic sampling and network optimization. For encoder and decoder in the dual-VAE for domain  $d \in \{C, D\}$ , the objective function is defined as:

$$\mathcal{L}_d^{\text{ELBO}} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\psi(\mathbf{x}|\mathbf{z})] + \beta_{\text{KL}} D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\psi(\mathbf{z})) \quad (5)$$

where  $\mathbf{z} \sim \text{Enc}(\mathbf{x}) \triangleq q_\phi(\mathbf{z}|\mathbf{x})$ ,  $\tilde{\mathbf{x}} \sim \text{Dec}(\mathbf{z}) \triangleq p_\psi(\mathbf{x}|\mathbf{z})$ , and  $D_{\text{KL}}$  is the Kullback–Leibler divergence. The first term in Eq. (5) 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  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ ).  $\beta_{\text{KL}}$  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. Therefore, embedding vectors ( $\mathbf{z}_{i,1:T_i}^C, \mathbf{z}_{i,1:T_i}^D$ ) projected from the *same* patient  $i$ , are supposed to be positioned closely in the shared latent space (see Supplementary Fig. 2). Therefore, in this study, we borrow the concept of matching loss from domain alignment in DA, which enables efficient representation learning crossing domains/modalities<sup>64</sup>. In this study, 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:

$$\mathcal{L}^{\text{Match}} = \mathbb{E}_{\mathbf{z} \sim p_z} \left[ \sum_{t \in T} \|\mathbf{z}_t^C - \mathbf{z}_t^D\|^2 \right] \quad (6)$$

The pairwise matching loss achieve its optimal when the distance proxy  $\mathcal{L}^{\text{Match}}$  becomes zero.

**Contrastive loss.** On the flip side, pairwise reconstruction error (i.e., intra-correlations within one instance) measured by *matching loss* neglects the commonalities present across patients (inter-correlations of data)<sup>65</sup>. 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 popularized in self-supervised learning (SSL)<sup>66</sup>, which aims to capture intrinsic patterns from input data without human annotations. In this study, we instantiate the contrastive loss via

*NT-Xent*, which is proposed by Chen et al. in their work SimCLR<sup>67</sup>. 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 adapt the corresponding auxiliary tasks for calculating contrastive loss to the scenario of learning representations from mixed-type timeseries. The objective of the task is to determine whether a set of representations transformed from the observational space belong to the *same* patient. And this leads to the corresponding positive pairs (true) and negative pairs (false).

For patient data of  $N$  records, we can obtain  $N$  pairs of latent representations from the encoders in *dual*-VAE. For patient indexed with  $i$ ,  $\mathbf{h}_i^C$  and  $\mathbf{h}_i^D$  denotes the embeddings derived from the continuous-valued and discrete-valued observational space, respectively. Due to the symmetric architecture of *dual*-VAE, here we use  $d$  and  $d'$  to represent one of each different domain, i.e.,  $d, d' \in \{C, D\}$  and  $d \neq d'$ . Therefore, the positive pairs for patient  $i$  can be referred as  $(i^d, i^{d'})$ , while the other  $2(N-1)$  samples are regarded as negative pairs. Then the contrastive loss for a positive pair  $(i^d, i^{d'})$  is defined as:

$$\mathcal{L}_{i^d, i^{d'}}^{\text{Contra}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_{i^d}, \mathbf{h}_{i^{d'}})/\tau)}{\sum_{j^{d''} \neq i^{d''}} \mathbb{1}_{[j^{d''} \neq i^{d''}]} \exp(\text{sim}(\mathbf{h}_{i^d}, \mathbf{h}_{j^{d''}})/\tau)} \quad (7)$$

where  $\text{sim}(u, v) = u^T v / \|u\| \|v\|$  denotes the cosine similarity between two vectors.  $\tau > 0$  denotes a temperature hyperparameter.  $\mathbb{1}_{[n \neq m]} \in \{0, 1\}$  is an indicator evaluating to 1 iff  $n \neq m$ . And  $i^{d''} \in \{1, 2, \dots, 2N\}$  represents the index of latent embeddings from *both* data types. The final contrastive loss is computed across the total number of  $|i^d - i^{d'}| = N$  positive pairs for both  $(i^d, i^{d'})$  and  $(i^{d'}, i^d)$ , and is defined as:

$$\mathcal{L}^{\text{Contra}} = \frac{1}{2N} \sum_{i^d=1}^N \sum_{i^{d'}=1}^N [\mathcal{L}_{i^d, i^{d'}}^{\text{Contra}} + \mathcal{L}_{i^{d'}, i^d}^{\text{Contra}}] \quad (8)$$

**Semantic loss.** In EHR-M-GAN<sub>cond</sub>, 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  $d \in \{C, D\}$ , given the latent embedding vector  $\mathbf{z}^d$  and the conditional information vector  $\mathbf{y}$ :

$$\mathcal{L}_d^{\text{Class}} = \mathbb{E}_{\mathbf{z}^d \in \mathcal{H}^S} \text{CE} \left( f_{\text{linear}}^d(\mathbf{z}^d), \mathbf{y} \right) \quad (9)$$

where  $f_{\text{linear}}^d$  denotes the linear classifier for the corresponding domain. And  $\text{CE} = -\sum_j y_j \log(\hat{y}_j)$ , ( $j = 1, 2, \dots, |L|$ ) denotes the cross entropy loss, where  $\hat{y}_j$  is the output of the linear classifier, and  $y_j$  is the ground truth label for class  $j$  in condition vector  $\mathbf{y}$ .

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

$$\mathcal{L}_d = \beta_0 \mathcal{L}_d^{\text{ELBO}} + \beta_1 \mathcal{L}^{\text{Match}} + \beta_2 \mathcal{L}^{\text{Contra}} \quad (10)$$

Under the conditional learning scenario of EHR-M-GAN<sub>cond</sub>, the total loss becomes:

$$\mathcal{L}_d = \beta_0 \mathcal{L}_d^{\text{ELBO}} + \beta_1 \mathcal{L}^{\text{Match}} + \beta_2 \mathcal{L}^{\text{Contra}} + \beta_3 \mathcal{L}_d^{\text{Class}} \quad (11)$$

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 Supplementary Note 3. As shown in Supplementary Table 7, 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

We propose the *sequentially coupled generator* for generating latent representations for mixed-type timeseries, which is built based on the network architecture of *coupled recurrent network* (CRN). Specifically, a CRN exploits bilateral long short-term memory (BLSTM) cells as its recurrent layer to preserve the temporal dependencies between the continuous and discrete-valued sequences. The network architecture of bilateral-LSTM we proposed can extract and transmit the correlations between the mixed-type timeseries, as opposed to vanilla-LSTM which has only one recursive connection. In the following section, we first discuss the structure of BLSTM in detail as its essential recurrent layer of CRN, and then build the *sequentially coupled generator* based on CRN.

**Bilateral long short-term memory.** As the traditional LSTM only considers temporal dynamics from single-type timeseries, therefore is incapable to extract and transmit temporal correlation from heterogeneous features. Therefore, we propose the bilateral-LSTM cell with two network connections to characterize the correlations between two types of data. Given  $d, d' \in \{\mathcal{C}, \mathcal{D}\}$ ,  $\mathbf{u}_t^d$  and  $\mathbf{h}_t^d$  denotes the input vector (i.e., the random noise during GANs' training) and hidden state vector for domain  $d$  at time step  $t$ , respectively. An additional set of weights for introducing hidden states representations  $\mathbf{h}_t^{d'}$  from domain  $d'$  is included when updating the input gate  $\mathbf{i}_t^{d'}$ , forget gate  $\mathbf{f}_t^{d'}$ , output gate  $\mathbf{o}_t^{d'}$ , and cell memory  $\mathbf{c}_t^{d'}$ . The state transition functions for BLSTM are:

$$\begin{aligned} \mathbf{i}_t^d &= \sigma(\mathbf{W}_{idu} \mathbf{u}_t^d + \mathbf{W}_{idh^{d'}} \mathbf{h}_{t-1}^{d'} + \mathbf{W}_{idh^d} \mathbf{h}_{t-1}^d + \mathbf{b}_{id}) \\ \mathbf{f}_t^d &= \sigma(\mathbf{W}_{fdu} \mathbf{u}_t^d + \mathbf{W}_{fdh^{d'}} \mathbf{h}_{t-1}^{d'} + \mathbf{W}_{fdh^d} \mathbf{h}_{t-1}^d + \mathbf{b}_{fd}) \\ \mathbf{o}_t^d &= \sigma(\mathbf{W}_{odu} \mathbf{u}_t^d + \mathbf{W}_{odh^{d'}} \mathbf{h}_{t-1}^{d'} + \mathbf{W}_{odh^d} \mathbf{h}_{t-1}^d + \mathbf{b}_{od}) \\ \mathbf{c}_t^d &= \tanh(\mathbf{W}_{cdu} \mathbf{u}_t^d + \mathbf{W}_{cdh^{d'}} \mathbf{h}_{t-1}^{d'} + \mathbf{W}_{cdh^d} \mathbf{h}_{t-1}^d + \mathbf{b}_{cd}) \\ \mathbf{c}_t^d &= \mathbf{f}_t^d \odot \mathbf{c}_{t-1}^d + \mathbf{i}_t^d \odot \mathbf{c}_t^d \\ \mathbf{h}_t^d &= \mathbf{o}_t^d \odot \tanh(\mathbf{c}_t^d) \end{aligned} \quad (12)$$

As indicated by Eq. (12), the proposed BLSTM network overcomes the limitation of vanilla-LSTM network on modeling the correlation between the mixed-type timeseries by establishing the supplemental recursive connection. The new connection facilitates the model to intrinsically decide how much information it should pass through the gates from its counterpart. A diagram of the BLSTM cell in contrast to vanilla-LSTM cell can be found in the Supplementary Fig. 3.

**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  $\mathbf{u}_t^d$  and  $\mathbf{u}_t^{d'}$  for two domains, which are sampled from uniform distributions (i.e.,  $\mathbf{u}_t^d, \mathbf{u}_t^{d'} \in \mathcal{U}(0, 1)$ ), are separately fed into the *input layers*. Then the *recurrent layers*  $f_{rec}$ , which are built based on two streams of BLSTM, one for each data type, are used to recursively iterate hidden states from both branches. Finally, the *fully connected layers*  $f_{conn}^d$  and  $f_{conn}^{d'}$  produce the generated latent vectors  $\hat{\mathbf{z}}_t^d$  and  $\hat{\mathbf{z}}_t^{d'}$  for the

decoding stage in *dual*-VAE. At time step  $t$ , CRN can be formulated as:

$$\begin{aligned} (\mathbf{h}_t^d, \mathbf{h}_t^{d'}) &= f_{rec}(\mathbf{u}_t^d, \mathbf{u}_t^{d'}, (\mathbf{h}_{t-1}^d, \mathbf{h}_{t-1}^{d'})) \\ \hat{\mathbf{z}}_t^d &= f_{conn}^d(\mathbf{h}_t^d) \\ \hat{\mathbf{z}}_t^{d'} &= f_{conn}^{d'}(\mathbf{h}_t^{d'}) \end{aligned} \quad (13)$$

In summary, heterogeneous timeseries that exhibits mutual influence on each other are integrated into CRN to model their interdependencies. By exploiting the BLSTM cell as its recurrent layer, 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:

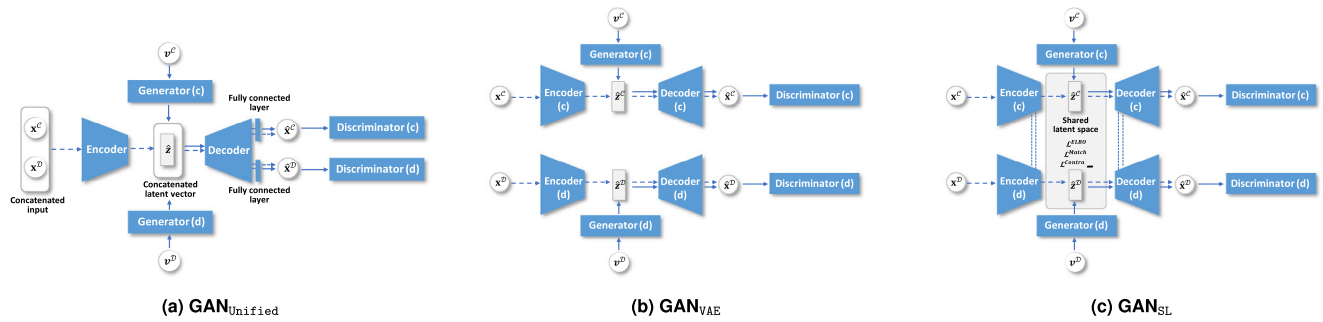
- The *pretraining of dual*-VAE: First, a dual-VAE network which consists of a pair of encoders ( $Enc^C, Enc^D$ ) and decoders ( $Dec^C, Dec^D$ ) is pretrained with both continuous and discrete data. Based on multiple objective constraints in Eq. (10), a shared latent space is learnt using *dual*-VAE, where the gap between the embedding representations ( $\mathbf{z}_{1:T}^C, \mathbf{z}_{1:T}^D$ ) from both domains is minimized.
- The *generation of latent representations based on CRN*: Then, during the joint training stage, the *sequentially coupled generator* which is built based on CRN, takes the random noise vector ( $\hat{\mathbf{z}}_{1:T}^C, \hat{\mathbf{z}}_{1:T}^D$ ) as inputs and iterating across the timesteps  $t \in \{1, 2, \dots, T\}$  by the internal transition functions. Therefore, the synthetic latent embedding representations ( $\hat{\mathbf{z}}_{1:T}^C, \hat{\mathbf{z}}_{1:T}^D$ ) for both continuous and discrete data can be obtained.
- The *decoding for the mixed-type timeseries*: Next, the generated latent embeddings ( $\hat{\mathbf{z}}_{1:T}^C, \hat{\mathbf{z}}_{1:T}^D$ ) are further fed into the pretrained decoders ( $Dec^C, Dec^D$ ) and decoded into the corresponding synthetic patient trajectories ( $\hat{\mathbf{x}}_{1:T}^C, \hat{\mathbf{x}}_{1:T}^D$ ) 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:

$$\begin{aligned} \min_G \max_D V_{\text{EHR-M-GAN}} &= \mathbb{E}_{\mathbf{x} \sim p_x} [\log D^C(\mathbf{x}^C) + \log D^D(\mathbf{x}^D)] \\ &+ \mathbb{E}_{\mathbf{u} \sim p_u} [\log(1 - D^C(\hat{\mathbf{x}}^C)) + \log(1 - D^D(\hat{\mathbf{x}}^D))] \end{aligned} \quad (14)$$

### Conditional version of EHR-M-GAN

For the conditional extension of EHR-M-GAN<sub>cond</sub>, the auxiliary label information is first used during the pretraining step of *dual*-VAE. Both the encoders and decoders condition on the auxiliary (one-hot) labels from  $\mathcal{L}$ , to make the model better adapted to particular contexts. In *dual*-VAE, the additional semantic loss is also incorporated during the optimization for the shared latent space as in Eq. (11). 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.<sup>68</sup>



**Fig. 7 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, **a**  $\text{GAN}_{\text{Unified}}$  learns the joint representations of heterogeneous types of data in a unified network; **b**  $\text{GAN}_{\text{VAE}}$  maintains the basic architecture of EHR-M-GAN, but ignore the dependency learning (i.e., separate networks for two streams of inputs are trained in parallel); **c**  $\text{GAN}_{\text{SL}}$  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.

The t-SNE visualization of the latent embeddings induced from *dual-VAE* can be found in Supplementary Note 4, which indicates that the conditional information carried into  $\text{EHR-M-GAN}_{\text{cond}}$  can be beneficial when synthesizing patient trajectories under specific medical conditions. Overall, the adversarial loss for  $\text{EHR-M-GAN}_{\text{cond}}$  can be denoted as follows:

$$\min_G \max_D V_{\text{EHR-M-GAN}_{\text{cond}}} = \mathbb{E}_{\mathbf{y}, \mathbf{x} \sim p_{\mathbf{y}, \mathbf{x}}} [\log D^c(\mathbf{x}^c | \mathbf{y}) + \log D^d(\mathbf{x}^d | \mathbf{y})] + \mathbb{E}_{\mathbf{y} \sim p_{\mathbf{y}}, \mathbf{u} \sim p_{\mathbf{u}}} [\log(1 - D^c(\hat{\mathbf{x}}^c | \mathbf{y})) + \log(1 - D^d(\hat{\mathbf{x}}^d | \mathbf{y}))] \quad (15)$$

The pseudocodes for *dual-VAE* and EHR-M-GAN are provided in the Supplementary Note 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 typically face challenges when modeling mixed-type EHR timeseries and can only synthesize single-type EHRs, we draw the comparison between EHR-M-GAN and the benchmark models 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<sup>69</sup>, R(C)GAN<sup>25</sup>, and TimeGAN<sup>29</sup>. For discrete-valued timeseries generation, classic medGAN<sup>32</sup>, seqGAN<sup>31</sup>, and two recently proposed work—SynTEG<sup>27</sup> and DualAAE<sup>26</sup> are used for comparison. Apart from these GAN-based models, we also incorporate PrivBayes<sup>70</sup> to synthesize discrete-valued timeseries, which falls in the class of non-GAN generative approaches using a Bayesian framework<sup>17</sup>. 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 ref. <sup>70</sup>). For medGAN and PrivBayes, we feed the flattened temporal sequence as the input since the models cannot produce timeseries data.

We further perform the **ablation study** to investigate whether our introduced novel components in the proposed model have advantages over its variants that also model mixed-type EHRs. First, we compare EHR-M-GAN with a variant that jointly models the mixed-type data using a single unified VAE network (denoted as  $\text{GAN}_{\text{Unified}}$ ). Second, we test the variant that encodes the mixed-type inputs separately with two independent VAE networks (denoted as  $\text{GAN}_{\text{VAE}}$ ). Then, we assess the effectiveness of the proposed *sequentially coupled generator* component in our model by implementing  $\text{GAN}_{\text{SL}}$ . Lastly, as the *dual-VAE* module alone can also be used for generating EHR timeseries, it serves as a non-GAN-based benchmark in the ablation study (see Supplementary Note 3). The architectures of different variants of EHR-M-GAN in

the ablation study are detailed as follows (also see Fig. 7 for illustration):

- $\text{GAN}_{\text{Unified}}$ : It contains a unified VAE module and two separate GANs. The continuous-valued and discrete-valued timeseries is concatenated together, via normalization and one-hot encoding, as input to the encoder in the unified VAE network. The decoder receives the concatenation of the generated latent vectors as the input, and then decodes it into synthetic timeseries with the corresponding data types using the separate fully connected layers. Each component in the architecture of  $\text{GAN}_{\text{Unified}}$  (unified encoder and decoder, separate generators and discriminators) is implemented with LSTMs, which are the same as EHR-M-GAN.
- $\text{GAN}_{\text{VAE}}$ : It is composed of a pair of VAE networks and GANs (one for each type of inputs). The continuous-valued timeseries and discrete-valued timeseries from the same patients are separately fed into the corresponding paths in  $\text{GAN}_{\text{VAE}}$ , and then run in parallel. The synthetic outputs of each data type are then combined as the final results. It maintains the basic structure of EHR-M-GAN but lacks the latent space sharing with *dual-VAE* and the *sequentially coupled generator* in the original EHR-M-GAN.
- $\text{GAN}_{\text{SL}}$ : In addition to  $\text{GAN}_{\text{VAE}}$ , it learns the shared latent space representations through *dual-VAE* by adding the corresponding loss functions in EHR-M-GAN, including *ELBO loss*, *Matching loss* and *Contrastive loss*. This model lacks the *sequentially coupled generator*.
- EHR-M-GAN: In addition to  $\text{GAN}_{\text{SL}}$ , it incorporates the *sequentially coupled generator* for the learning the correlated temporal dynamics in timeseries of different data types. This is the proposed full model.
- $\text{EHR-M-GAN}_{\text{cond}}$ : This version is implemented on the basis of conditional GAN<sup>68</sup>, where the conditional inputs are fed into EHR-M-GAN to generate patients under specific labels.

For training  $\text{EHR-M-GAN}_{\text{cond}}$ , auxiliary information from the patient status is used as conditional input. These conditional inputs are selected since synthesizing EHR information of patient subgroups with potential outcomes would be valuable for clinicians in their decision-making process. Other conditional labels (such as patient demographics in the categorized format) can also be used in the proposed conditional synthesizer for other research purposes. For MIMIC-III dataset, the classes are (1) *ICU mortality*: patient died within the ICU; (2) *Hospital mortality*: patient discharged alive from the ICU, and died in the hospital; (3) *30-day readmission*: patient discharged alive from the hospital, and readmitted to the hospital within 30 days; (4) *No 30-day readmission*: patient discharged alive from the hospital, and had no readmission record to the hospital within 30 days. For eICU and



HiRID datasets, the corresponding labels are also extracted based on the availability of the patient outcomes (see Table 7).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## DATA AVAILABILITY

All datasets are freely accessible. The MIMIC-III dataset (Version: 1.4) can be accessed at <https://physionet.org/content/mimiciii/1.4/>. The eICU-CRD database (Version: 2.0) can be accessed at <https://physionet.org/content/eicu-crd/2.0/>. The HiRID dataset (Version: 1.1.1) can be accessed at <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 at [https://github.com/jli0117/preprocessing\\_physionet](https://github.com/jli0117/preprocessing_physionet). Code for proposed EHR-M-GAN model can be found at <https://github.com/jli0117/ehrMGAN>.

Received: 16 September 2022; Accepted: 5 May 2023;

Published online: 27 May 2023

## REFERENCES

- Artzi, N. S. et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nat. Med.* **26**, 71–76 (2020).
- Raket, L. L. et al. Dynamic electronic health record detection (detect) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digital Health* **2**, e229–e239 (2020).
- Menger, V., Spruit, M., Van Est, R., Nap, E. & Scheepers, F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Netw. Open* **2**, e196709 (2019).
- Wilkinson, J. et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digital Health* **2**, e677–e680 (2020).
- Watson, D. S. et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* **364**, l886 (2019).
- Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health* **2**, e489–e492 (2020).
- Esteve, A. et al. Deep learning-enabled medical computer vision. *NPJ Digital Med.* **4**, 1–9 (2021).
- Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
- Wirth, F. N., Meurers, T., Johns, M. & Prasser, F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Med. Inform. Decis. Mak.* **21**, 1–13 (2021).
- Dinov, I. D. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience* **5**, s13742–016 (2016).
- Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **19**, 1236–1246 (2018).
- Kim, J. et al. Privacy-protecting, reliable response data discovery using covid-19 patient observations. *J. Am. Med. Inform. Assoc.* **28**, 1765–1776 (2021).
- Simon, G. E. et al. Assessing and minimizing re-identification risk in research data derived from health care records. *eGEMS* **7**, 6 (2019).
- Jordon, J., Yoon, J. & Van Der Schaar, M. PATE-GAN: generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations* (ICLR, 2019).
- Frid-Adar, M. et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).
- Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
- Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Med.* **3**, 1–13 (2020).
- El Emam, K., Mosquera, L., Jonker, E. & Sood, H. Evaluating the utility of synthetic covid-19 case data. *JAMIA Open* **4**, ooab012 (2021).

- N3c. Synthetic data workstream. [https://covid.cd2h.org/N3C\\_synthetic\\_data](https://covid.cd2h.org/N3C_synthetic_data) (2021).
- CPRD. Synthetic data. <https://www.cprd.com/content/synthetic-data> (2021).
- Goodfellow, I. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems* **27** (NIPS, 2014).
- Kearney, V. et al. Dosegan: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Sci. Rep.* **10**, 1–8 (2020).
- Yang, Q. et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans. Med. Imaging* **37**, 1348–1357 (2018).
- Marouf, M. et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**, 1–12 (2020).
- Esteban, C., Hyland, S. L. & Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. Preprint at *arXiv* <https://arxiv.org/abs/1706.02633> (2017).
- Lee, D. et al. Generating sequential electronic health records using dual adversarial autoencoder. *J. Am. Med. Inform. Assoc.* **27**, 1411–1419 (2020).
- Zhang, Z., Yan, C., Lasko, T. A., Sun, J. & Malin, B. A. Synteg: a framework for temporal structured electronic health data simulation. *J. Am. Med. Inform. Assoc.* **28**, 596–604 (2021).
- Zhang, Z., Yan, C. & Malin, B. A. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. *J. Am. Med. Inform. Assoc.* **29**, 1890–1898 (2022).
- Yoon, J., Jarrett, D. & Van der Schaar, M. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems* **32** (NIPS, 2019).
- de Rosa, G. H. & Papa, J. P. A survey on text generation using generative adversarial networks. *Pattern Recogn.* **119**, 108098 (2021).
- Yu, L., Zhang, W., Wang, J. & Yu, Y. SeqGAN: sequence generative adversarial nets with policy gradient. In *Proc. AAAI Conference on Artificial Intelligence* 2852–2858 (ACM, 2017).
- Choi, E. et al. Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference* 286–305 (PMLR, 2017).
- Yu, C., Liu, J. & Zhao, H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med. Inform. Decis. Mak.* **19**, 111–120 (2019).
- Ghassemi, M., Wu, M., Hughes, M. C., Szolovits, P. & Doshi-Velez, F. Predicting intervention onset in the ICU with switching state space models. *AMIA Summ. Transl. Sci. Proc.* **2017**, 82 (2017).
- Wang, L., Zhang, W. & He, X. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In *International Conference on Database Systems for Advanced Applications* 36–52 (Springer, 2019).
- Johnson, A. E. et al. Mimic-iii, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
- Pollard, T. J. et al. The eicu collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 1–13 (2018).
- Yèche, H. et al. Hirid-icu-benchmark – a comprehensive machine learning benchmark on high-resolution icu data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (NIPS, 2021).
- Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* 3–18 (IEEE, 2017).
- Dwork, C. Differential privacy. In *International Colloquium on Automata, Languages, and Programming* 1–12 (Springer, 2006).
- Borji, A. Pros and cons of gan evaluation measures. *Comput. Vis. Image Understanding* **179**, 41–65 (2019).
- Sutherland, D. J. et al. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations* (ICLR, 2017).
- Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* **26**, 228–241 (2019).
- Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A. & Miklau, G. Benchmarking differentially private synthetic data generation algorithms. In *The Third AAAI Workshop on Privacy-Preserving Artificial Intelligence* (PPAI-22).
- Benedetti, J. d., Oues, N., Wang, Z., Myles, P. & Tucker, A. Practical lessons from generating synthetic healthcare data with bayesian networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 38–47 (Springer, 2020).
- Wang, S. et al. Mimiextract: a data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proc. ACM Conference on Health, Inference, and Learning* 222–235 (ACM, 2020).

47. Wu, M. et al. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *J. Am. Med. Inform. Assoc.* **24**, 488–495 (2017).
48. Suresh, H. et al. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference* 322–337 (PMLR, 2017).
49. Kiyasseh, D. et al. Plethaugment: Gan-based ppg augmentation for medical diagnosis in low-resource settings. *IEEE J. Biomed. Health Inform.* **24**, 3226–3235 (2020).
50. Hayes, J., Melis, L., Danezis, G. & De Cristofaro, E. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies*, 133–152 (2019).
51. Chen, D., Yu, N., Zhang, Y. & Fritz, M. Gan-leaks: a taxonomy of membership inference attacks against generative models. In *Proc. 2020 ACM SIGSAC Conference on Computer and Communications Security* 343–362 (ACM, 2020).
52. Lin, Z., Jain, A., Wang, C., Fanti, G. & Sekar, V. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proc. ACM Internet Measurement Conference* 464–483 (ACM, 2020).
53. Dwork, C. Differential privacy: a survey of results. In *International Conference on Theory and Applications of Models of Computation* 1–19 (Springer, 2008).
54. Cormode, G. et al. Privacy at scale: local differential privacy in practice. In *Proc. 2018 International Conference on Management of Data* 1655–1658 (ACM, 2018).
55. Alaa, A. M. & van der Schaar, M. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems* 32 (NIPS, 2019).
56. Lee, C. & Van Der Schaar, M. Temporal phenotyping using deep predictive clustering of disease progression. In *International Conference on Machine Learning* 5767–5777 (PMLR, 2020).
57. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, 1–12 (2018).
58. Yoon, J., Jordon, J. & Schaar, M. Gain: missing data imputation using generative adversarial nets. In *International Conference on Machine Learning* 5689–5698 (PMLR, 2018).
59. Yan, C. et al. A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* **13**, 1–18 (2022).
60. Yoon, J., Jordon, J. & Van Der Schaar, M. Ganite: estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations* (ICLR, 2018).
61. Qian, Z., Zhang, Y., Bica, I., Wood, A. & van der Schaar, M. SyncTwin: treatment effect estimation with longitudinal outcomes. In *Advances in Neural Information Processing Systems* 34 (NIPS, 2021).
62. Liu, M.-Y., Breuel, T. & Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems* 700–708 (2017).
63. Liu, M.-Y. & Tuzel, O. Coupled generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **29**, 469–477 (2016).
64. Wan, Z. et al. Old photo restoration via deep latent space translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 2071–2087 (2022).
65. Kiyasseh, D., Zhu, T. & Clifton, D. A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning* 5606–5615 (PMLR, 2021).
66. Liu, X. et al. Self-supervised learning: generative or contrastive. In *IEEE Transactions on Knowledge and Data Engineering* 857–876 (IEEE, 2021).
67. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
68. Mirza, M. & Osindero, S. Conditional generative adversarial nets. Preprint at *arXiv* <https://arxiv.org/abs/1411.1784> (2014).
69. Mogren, O. C-rnn-gan: continuous recurrent neural networks with adversarial training. Preprint at *arXiv* <https://arxiv.org/abs/1611.09904> (2016).
70. Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. & Xiao, X. Privbayes: private data release via bayesian networks. *ACM Trans. Database Syst.* **42**, 1–41 (2017).

## ACKNOWLEDGEMENTS

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. Jin Li was funded by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. T.Z. was supported by the Royal Academy of Engineering under the Research Fellowship scheme and the National Institute for Health Research Oxford Biomedical Research Centre.

## AUTHOR CONTRIBUTIONS

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

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-023-00834-7>.

**Correspondence** and requests for materials should be addressed to Jingsong Li or Tingting Zhu.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023