ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models

Ahmed Ammar Naseer[1,2] ANASEER@VIRONIX.AI
Benjamin Walker[1] BENJAMIN.WALKER2@BALLIOL.OX.AC.UK
Christopher Landon[3] CHRIS.LANDON@VENTURA.ORG
Andrew Ambrosy[4,5] ANDREW.P.AMBROSY@KP.ORG
Marat Fudim[6,7] MARAT.FUDIM@DUKE.EDU
Nicholas Wysham[8,9] NWYSHAM@TVC.ORG
Botros Toro[2] BTORO@VIRONIX.AI
Sumanth Swaminathan[2,10,11] SSWAMI@VIRONIX.AI
Terry Lyons [1,12] TERRY.LYONS@MATHS.OX.AC.UK

[1] *Mathematical Institute, University of Oxford, Oxford, UK*
[2] *Vironix Health Inc, Austin, TX, USA*
[3] *Landon Pediatric Foundation, Ventura, CA, USA*
[4] *Dept. of Cardiology, Kaiser Permanente San Francisco Medical Center, San Francisco, CA, USA*
[5] *Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA*
[6] *Division of Cardiology, Duke University Medical Center, Durham, NC, USA*
[7] *Duke Clinical Research Institute, Durham, NC, USA*
[8] *The Vancouver Clinic, Vancouver, WA, USA*
[9] *Department of Medicine, University of Washington, Seattle, WA, USA*
[10] *Department of Mathematical Sciences, University of Delaware, Newark, DE, USA*
[11] *Industrially Focused Mathematical Modelling CDT, University of Oxford, Oxford, UK*
[12] *The Alan Turing Institute, British Library, London, UK*

Abstract

Global access to statistically and clinically representative patient health data holds potential for advancing disease research, enhancing patient care, and accelerating drug development. However, acquisition of health data such as electronic health records (EHRs) comes with challenges characterised by high costs, time constraints, and concerns related to patient privacy. An approach to tackling these challenges is by using synthetic data. In this paper we introduce ScoEHR, a novel deep learning method for generating synthetic EHRs, which combines an autoencoder with a continuous-time diffusion model. ScoEHR is shown to outperform three baseline synthetic EHR generation frameworks (medGAN, medWGAN, and medBGAN) on two publicly available datasets, MIMIC-III and the Yale New Haven Health System Emergency Department dataset, based on four widely accepted metrics of data utility. Additionally, a blind clinician evaluation was carried out to assess the qualitative realism of the synthetic data generated by ScoEHR. In this evaluation, a patient's data was labeled as 'unrealistic' if at least one clinician found it to be unrealistic. This evaluation showed that existing real EHR data and ScoEHR generated synthetic data were scored as equally realistic. Our code is available at https://github.com/aanaseer/ScoEHR.

1. Introduction

Patient health data gathered during the provision of medical services is often stored as electronic health records (EHRs). These records are widely used within the healthcare industry, with over 84% of hospitals in the US and 94% of hospitals in the UK making use of EHR systems (Li et al., 2020). EHRs contain temporal snapshots of structured data (e.g. patient demographics, physiologic measurements, medications) and unstructured data (e.g. medical images, discharge summaries, physician comments) (Tayefi et al., 2021), that can be mined, analysed, and modelled to elucidate disease progression and the underpinnings of health deterioration. Accordingly, EHRs represent a valuable source of information for predictive and prescriptive machine-learning models that enhance patient care. Historically, medical records have been used for predicting intensive care unit re-admissions (Rojas et al., 2018), risk of disease (Ruan et al., 2020), disease severity (Kogan et al., 2020), and forecasting clinical outcomes (Norgeot et al., 2019).

Despite the research opportunities presented by EHRs, sharing such records with external parties remains limited (Yan et al., 2022). The health insurance portability and accountability act (HIPAA) in the United States and general data protection regulation (GDPR) in the European Union outline stringent guidelines for handling and sharing of such data, even in de-identified forms (Kaissis et al., 2020). In addition to the data access problem, the inability to aggregate EHR data across practices and geographies limits the robustness of predictive models as many medical institutions maintain local patient demographics, standards-of-care, healthcare disparities, and local reimbursement policies (Gianfrancesco et al., 2018). A potential way to overcome issues of EHR data access and bias is through synthetic data generation (Hernandez et al., 2022). Synthetic healthcare data has been generated for modalities such as medical imaging (Frid-Adar et al., 2018), biomedical signals (Hernandez-Matamoros et al., 2020), and EHRs (Choi et al., 2017). Synthetically generated patient vignettes have also been used in machine-learning applications such as patient monitoring and early detection of health deterioration (Swaminathan et al., 2017, 2020; Morrill et al., 2021).

Generating synthetic EHRs is not a trivial task. Records are often high-dimensional, heterogeneous, and contain a mix of discrete and continuous values (Xu et al., 2019). Additionally, data in continuous columns often follow a multi-modal distribution while categorical columns exhibit non-uniform distributions with significant imbalances (Xu et al., 2019).

In this paper we limit our focus to data-driven methods for structured EHRs. The popularity gained by generative adversarial networks (GANs) (Goodfellow et al., 2014) in various domains, such as computer vision, has prompted progress in using GANs for synthesising EHRs (Choi et al., 2017; Baowaly et al., 2019; Torfi and Fox, 2020; Zhang et al., 2020). When generating EHRs it is important to ensure fairness and diversity in the records. A limitation of GANs is their inability to synthesise diverse data representing good mode coverage because of mode collapse (Goodfellow, 2016). Furthermore, GANs struggle to converge to a stable solution, resulting in a difficult training process (Goodfellow, 2016). More recently, a class of generative models referred to as diffusion models have attained state-of-the-art results in image generation (Song et al., 2021). Diffusion models involve a forward process by which a data distribution is perturbed gradually, followed by a reverse

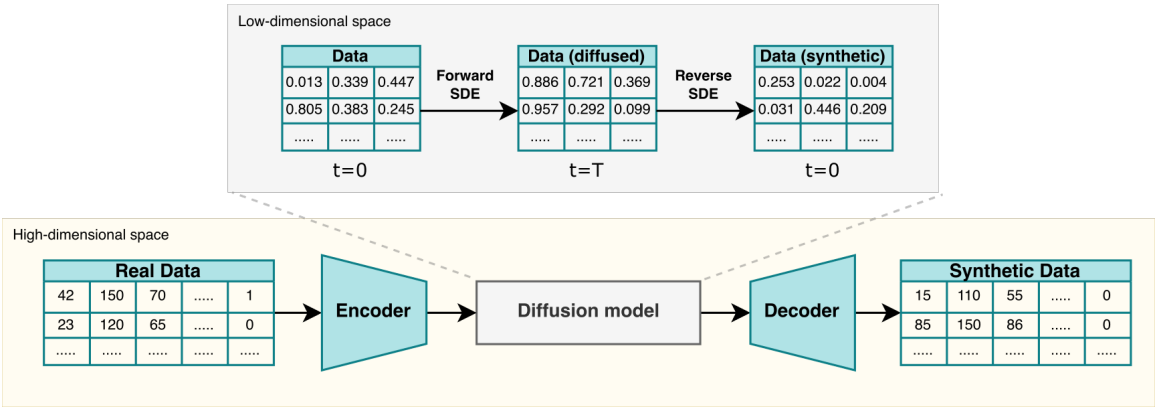

Figure 1: A schematic diagram of ScoEHR's architecture. Real data is transformed to a low-dimensional space using an encoder from a pre-trained autoencoder. In the low dimensional space, a forward stochastic differential equation (SDE) diffuses the data. The reverse SDE is learnt and used to produce new synthetic data. This synthetic data is transformed using the decoder from the autoencoder to generate the final synthetic EHR data.

de-noising process to generate data (Yang et al., 2022). These models have demonstrated the ability to synthesise high fidelity data with good mode coverage (Song et al., 2021). Furthermore, as they do not require adversarial training unlike GANs, their training is more stable. For these reasons, diffusion models present an ideal paradigm for synthetic EHR generation.

In this paper, we introduce ScoEHR, a framework for synthetic EHR generation using continuous-time diffusion models. A schematic of our architecture is shown in figure 1. We compare ScoEHR with three baseline models medGAN (Choi et al., 2017), medWGAN, and medBGAN (Baowaly et al., 2019) in four key performance metrics of data generation utility including: 1) preservation of feature marginal relationships, 2) preservation of feature correlations, 3) preservation of full feature distribution using log-clusters, and 4) synthetic data performance in downstream predictions of patient outcomes. In addition, the realism of ScoEHR generated data is evaluated by a group of USA board certified physicians to show strong concordance with actual EHRs. Finally, a brief examination of the privacy disclosure of our model is also presented. This paper demonstrates that ScoEHR is the current state-of-the-art in generating synthetic EHRs.

Generalisable Insights about Machine Learning in the Context of Healthcare

In addition to introducing a novel state-of-the-art method for synthetic EHR generation, this paper presents a number of generalisable insights about machine learning in the context of healthcare. These insights encompass:

- The framework proposed for generating synthetic EHRs is capable of generalising across different seed datasets, accommodating data from diverse geographical regions, diseases, and healthcare practices.
- The success of diffusion models can be harnessed for applications in healthcare.
- The effectiveness of models used to generate synthetic EHRs relies on the dataset utilised and the specific use case.
- The combination of metrics chosen for the comparative analysis of synthetic EHRs in this study is shown to be suitable for evaluating whether a model generates clinically realistic patients.

The first insight emphasises the significance of having access to representative patient data, which is critical for healthcare research and product development. ScoEHR offers a solution to enhance access to healthcare data by generating synthetic EHR from a seed dataset, thereby facilitating the availability of data for various research endeavours. This approach also enables the reduction of bias in machine learning modelling efforts by generating statistically and clinically representative data from any seed data source. It is essential to address this issue as machine learning models trained on EHRs are known to exhibit bias towards certain patient demographics, healthcare practices, geographic-specific governance, and socioeconomic groups (Gianfrancesco et al., 2018; Momenzadeh et al., 2022; Juhn et al., 2022; Thompson et al., 2021). By using ScoEHR to generate representative data from multiple sources, the sharing of a more inclusive dataset is promoted, contributing to a fairer and more comprehensive representation of patient populations in healthcare research.

Diffusion models are a powerful class of generative models. They have shown to outperform models such as GANs in domains such as computer vision, natural language processing, and computational chemistry (Yang et al., 2022). The second insight accounts for the fact that this paper is amongst the first to demonstrate the effectiveness of diffusion models in generating EHRs.

The third insight is drawn from this work's comparative analysis of ScoEHR with medGAN, medWGAN, and medBGAN. When considering the aggregate metrics in the comparative analysis, ScoEHR is the best performing model, though there are use cases where a medGAN variant would be preferable. Furthermore, even if considering a single use case, the general ranking of the models is dataset dependent. As such, this study suggests that clinicians should not ovelry rely on a single modelling paradigm when trying to generate patient data for research.

Finally, evaluation of synthetic EHRs is challenging as it is difficult to measure whether a change to a patient makes them clinically unrealistic without expert human input. It is therefore important to, where possible, evaluate generated healthcare data using clinician evaluation, which is often overlooked. The strong performance of ScoEHR in the comparative analysis along with the clinician evaluation provides evidence that the chosen combination of metrics are suitable for evaluating synthetic EHRs. Therefore, this study does yield a robust combination of performance metrics that could be a benchmark for evaluation of future modelling efforts.

2. Related Work

Many methods have been used for generating synthetic EHRs, including domain knowledge (Buczak et al., 2010), Bayesian networks (Rankin et al., 2020), and summary statistics (McLachlan et al., 2016). In this section we provide an overview of deep generative modelling for EHR generation, as it is most closely related to our approach.

One of the earliest studies in this regard is medGAN (Choi et al., 2017), in which Choi et al. (2017) make use of a pre-trained autoencoder and a GAN to generate synthetic binary and count data. In their framework, they first train an autoencoder to learn a low-dimensional continuous embedding of high-dimensional discrete features in an EHR dataset. They do so to overcome the limitations in using discrete data with GANs. They then train a GAN in the low-dimensional space using Jensen-Shannon divergence as their objective function. The generated data is projected back onto the high-dimensional space using the decoder component of the pre-trained autoencoder. Choi et al. (2017) makes use of mini-batch averaging to address the issue of mode collapse. Additionally, they use batch normalisation (Ioffe and Szegedy, 2015) and shortcut connections (He et al., 2016) to improve learning efficiency.

By taking advantage of improvements proposed to the traditional GAN architecture, Baowaly et al. (2019) introduced medWGAN and medBGAN. Both medWGAN and medBGAN closely follow the medGAN architecture; the key difference being replacement of GAN with a Wassertein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017) and a boundary-seeking GAN (Hjelm et al., 2018) respectively. By doing so, they generated more realistic synthetic EHRs compared to medGAN. In the work by Choi et al. (2017) and Baowaly et al. (2019), all neural networks are feed-forward multi-layer perceptrons. Building on this, Torfi and Fox (2020) developed corGAN using a convolutional GAN and autoencoder. They used such networks to capture the correlation between neighbouring features.

Our work is similar in architecture to these models given that we make use of an autoencoder and a generative model (continuous-time diffusion model) in the framework. In addition to supporting both binary and count data, our model also handles continuous values.

Further modifications to the medGAN architecture found in literature include removing the autoencoder component by Zhang et al. (2020). In their EMR-WGAN, a traditional GAN along with Wassertein divergence is used with normalisations to address the issue of exploding gradients. Similar efforts have been made in generating other forms of EHRs such as time-series data (Zhang et al., 2021) and those making use of differential privacy (Jordon et al., 2018b), which is beyond the scope of this paper. For further reading on the use of GANs in EHRs generation, we refer the reader to (Ghosheh et al., 2022) and (Hernandez et al., 2022).

3. Background

3.1. Autoencoders

Autoencoders (Rumelhart et al., 1985) are a type of neural network architecture used to learn a latent representation of its input. It consists of an encoder, $E(\cdot\,;\phi) : \mathbb{R}^n \mapsto \mathbb{R}^m$

parametrised by learnable parameters ϕ and a decoder $D(\cdot\,;\theta) : \mathbb{R}^m \mapsto \mathbb{R}^n$ parametrised by learnable parameters θ, with $m < n$. The encoder maps the input $\mathbf{x}$ to the low-dimensional representation, while the decoder maps from the low-dimensional representation to the data space, reconstructing the input data.

3.2. Continuous-time Diffusion Models

Continuous-time diffusion models are generalisations to past work in diffusion models (Ho et al., 2020; Song and Ermon, 2019) through the use of stochastic differential equations (SDEs) (Song et al., 2021). Such models have a forward diffusion process, a score-matching stage, and a reverse diffusion process for sample generation.

Notation Let $p : \mathbb{R}^n \to \mathbb{R}_{>0}$ represent a data distribution. Let t denote time. Let $\mathcal{D}$ represent a dataset with independent and identically distributed (i.i.d) samples $\mathbf{x} \in \mathbb{R}^n$. Let p_0 represent the data distribution of $\mathcal{D}$ at $t = 0$. Let $\mathbf{w}_t \in \mathbb{R}^n$ represent standard Brownian motion at t.

Data perturbation In continuous-time diffusion models (Song et al., 2021), perturbation of data is performed through a forward diffusion process from $t = 0$ to $t = T$, with an Itô SDE,

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{1}$$

along with initial data sample drawn from p_0 (i.e. $\mathbf{x}_0 \sim p_0$). Here $\mathbf{f}(\cdot\,, t) : \mathbb{R}^n \mapsto \mathbb{R}^n$ and $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ represent the drift and diffusion coefficients of $\mathbf{x}_t$ respectively. The data distribution during transition from $\mathbf{x}_s$ to $\mathbf{x}_t$ is given by the transition kernel $p_{st}(\mathbf{x}_t|\mathbf{x}_s)$ where $0 \leq s < t \leq T$. For an affine drift coefficient, the transition kernel is always Gaussian (Särkkä and Solin, 2019). Additionally, choosing a specific drift and diffusion coefficient ensures that at $t = T$ the perturbed data distribution approximates a fixed prior distribution p_T, such as a Gaussian (Song et al., 2021).

Sample generation Anderson (1982) derived the reverse-time SDE for equation (1),

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g(t)^2\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ is the score of the data distribution at t and $\bar{\mathbf{w}}$ is the reverse-time Browninan motion. In the reverse-time direction, the solution trajectories of equations (1) and (2) have the same marginal probability densities (Song et al., 2021). As a result, solving equation (2) from $t = T$ to $t = 0$ with initial condition $\mathbf{x}_T \sim p_T$ allows for sample generation. To generate samples, equation (2) can be solved using general-purpose numerical SDE solvers, predictor-corrector samplers, or by solving the probability flow ordinary differential equation (ODE) (Song et al., 2021).

Score matching To solve equation (2), the score at each time step is required. To obtain the scores we can perform score matching (Hyvärinen and Dayan, 2005). This can be achieved by training a time-dependent score-network $\mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\Theta})$ parameterised by learnable parameters $\boldsymbol{\Theta}$ to estimate the scores (i.e. $\mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\Theta}) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$). The explicit score-matching training objective is given by (Hyvärinen and Dayan, 2005),

$$\mathcal{L}_{ESM} \triangleq \mathbb{E}_{p(\mathbf{x}_t)}\left[\frac{1}{2}\|\mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\Theta}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\|_2^2\right]. \tag{3}$$

It is not possible to directly train using equation (3), as it explicitly contains the intractable score function, $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. However, up to a constant of integration, the explicit score-matching objective is equivalent to the de-noising score-matching objective,

$$\frac{1}{2}\mathbb{E}_t\mathbb{E}_{p(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\lambda(t)\|\mathbf{s}(\mathbf{x}_t, t; \mathbf{\Theta}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right], \tag{4}$$

where $\lambda(t) : [0, T] \mapsto \mathbb{R}_{>0}$ is a positive weighting function. As will be discussed in 4.2, for a particular choice of forward diffusion process, $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)$ can be written down explicitly, allowing score-networks to be trained.

4. Method: ScoEHR

4.1. Notation

Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ represent an EHR dataset where $\mathbf{x}_i \in \mathbb{R}^n$ denotes a record for a patient. Let $\mathcal{D}^{Tr} \in \mathbb{R}^{P\times n}$ and $\mathcal{D}^{Ts} \in \mathbb{R}^{Q\times n}$ represent the disjoint training and test datasets, where $\mathcal{D} = \mathcal{D}^{Tr} \cup \mathcal{D}^{Ts}$ and $Q+P = N$. Let each record comprise C continuous values and D discrete binary values, such that each record is given by $\mathbf{x}_i = \{\mathbf{c}_i, \mathbf{d}_i\} = \{c_{1,i}, ..., c_{n_c,i}, d_{1,i}, ..., d_{1,n_d}\}$, where $c_{j,i}$ and $d_{j,i}$ denote continuous and discrete values respectively and $n_c + n_d = n$. A dataset with synthetic records is denoted by $\widehat{\mathcal{D}} = \{\widehat{\mathbf{x}}_i\}_{i=1}^N = \{\widehat{\mathbf{c}}_i, \widehat{\mathbf{d}}_i\}$.

4.2. ScoEHR framework

The ScoEHR framework comprises an encoder, continuous-time diffusion model, and decoder as shown in figure 1. The encoder and decoder are components from a pre-trained autoencoder.

Autoencoder Given that each record in $\mathcal{D}$ comprises both continuous and discrete features, before perturbing the data using the diffusion model, we learn a low-dimensional representation of the high-dimensional features in $\mathcal{D}^{Tr}$. Similar to Choi et al. (2017), we do so to ensure that our generative model is able to learn the distribution of discrete values in $\mathcal{D}^{Tr}$. The encoder E and decoder D are single layer perceptrons. The loss function during training is

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{MSE}, \tag{5}$$

where

$$\mathcal{L}_{BCE} = -\mathbf{c}_i \log(D(E(\mathbf{c}_i))) + (1 - \mathbf{c}_i) \log(1 - D(E(\mathbf{c}_i))), \tag{6}$$

and

$$\mathcal{L}_{MSE} = \|\mathbf{d}_i - D(E(\mathbf{d}_i))\|_2^2. \tag{7}$$

Diffusion model After obtaining a low-dimensional continuous representation of our dataset we perturb the data over a time interval $t \in [0, T]$ using the variance preserving (VP) SDE (Song et al., 2021). The drift and diffusion coefficients of the VP SDE are

$$\mathbf{f}(\mathbf{x}_t, t) \triangleq -\frac{1}{2}\beta(t)\mathbf{x}_t \quad \text{and} \quad g(t) \triangleq \sqrt{\beta(t)} \tag{8}$$

respectively. Here $\beta(t) : \mathbb{R} \mapsto \mathbb{R}$ represents a time-dependent linear noise schedule

$$\beta(t) \triangleq \beta_{min} + t(\beta_{max} - \beta_{min}), \tag{9}$$

where $\beta_{min}, \beta_{max} \in \mathbb{R}$. Since our drift coefficient is affine, we can obtain a Gaussian transition kernel given by

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \mathbf{x}_0 e^{-\frac{1}{2}\xi(t)\mathrm{d}s}, \mathbf{I}(1 - e^{-\xi(t)})\right), \tag{10}$$

where

$$\xi(t) = \int_0^t \beta(s)\mathrm{d}s. \tag{11}$$

As we have an explicit transition kernel, the denoising score-matching objective given in equation (4) can be used to train our score-network.

Sample generation To generate samples we numerically solve the reverse-time SDE using the Euler-Maruyama numerical scheme.

4.3. Evaluation metrics

Evaluation of synthetic data is a challenging task as there is no universally established metric to compare against (Hernandez et al., 2022). Therefore, we evaluate the utility of the data synthesised using a combination of quantitative and qualitative metrics. In addition, we also investigate the privacy of our synthetic data through an adversarial attack.

4.4. Data utility evaluation

In our quantitative evaluation of utility, we seek to check four areas: (i) whether the marginal distributions in real data are captured by the synthetic data, (ii) the correlations in both real and synthetic datasets are similar, (iii) the full similarity of both datasets in terms of marginals and correlations, and (iv) the utility of the synthetic data for use in downstream machine learning tasks.

Dimension-wise distribution To evaluate the extent to which marginal relationships in real data are captured by the synthetic data we use a dimensional distribution metric. For binary data, this is evaluated using the sum of the absolute difference in dimension-wise mean for binary data, denoted by DWM (Yan et al., 2022). This is defined mathematically as

$$\text{DWM} = \sum_{i=1}^{N} \left| \frac{1}{n_d} \sum_{j=1}^{n_d} d_{j,i} - \widehat{d}_{j,i} \right|. \tag{12}$$

In order to extend this to continuous variables for the ED EHR dataset, we use the sum of the dimension-wise $1-$Wasserstein distance (Villani, 2008), denoted DEM, between the continuous variables

$$\text{DEM} = \sum_{i=1}^{N} \frac{1}{n_c} \sum_{j=1}^{n_c} |c_{j,i} - \widehat{c}_{j,i}|. \tag{13}$$

The score reported as the dimensional distribution is

$$\frac{\text{DWM} + \text{DEM}}{N}. \tag{14}$$

A lower dimensional distribution indicates that the real and synthetic data have similar marginal relationships.

Pairwise correlation difference In an EHR dataset, feature-wise correlations are clinically significant. Therefore, to verify whether the synthetic data has the same correlations as real data we make use of pairwise correlation difference (PCD). To compute PCD we obtain the Pearson correlation matrices $\text{Corr}(\cdot)$ for the real and synthetic data, after which we compute the Frobennius norm of their difference,

$$\text{PCD} = \|\text{Corr}(\mathcal{D}) - \text{Corr}(\widehat{\mathcal{D}})\|_F. \tag{15}$$

The closer PCD is to zero, the better the inter-dimensional relationships are captured by the synthetic data.

Log-cluster The log-cluster metric, denoted $\mathcal{U}$, attempts to capture the similarity of the entirety of the real and synthetic data distributions (Woo et al., 2009; Goncalves et al., 2020). This is done with unsupervised clustering to evaluate the similarity of the latent structure of real and synthetic datasets (Ghosheh et al., 2022; Goncalves et al., 2020). First, the real and synthetic datasets are concatenated and k-means clustering is performed with G clusters. Then, the log-cluster metric is given by

$$\mathcal{U} = \log\left(\frac{1}{G}\sum_{j=1}^{G}\left[\frac{n_j^R}{n_j} - c\right]^2\right), \tag{16}$$

where n_j denotes the number of samples in j-th cluster, n_j^R the number of samples from the real dataset in the $j-$th cluster, and $c = n^R/(n^R+n^S)$ in which n^S is the number of samples in the synthetic dataset. A lower log-cluster score indicates more similarity between the synthetic and real data.

Synthetic ranking agreement We are often interested in using synthetic data for a downstream machine learning task. Therefore, it is important to understand whether we can use synthetic data in place of real data for such tasks. In order to evaluate this, Jordon et al. (2018b) proposed the synthetic ranking agreement (SRA). SRA is viewed as an empirical probability of a comparison made using synthetic data being the same as the results obtained when using real data (Jordon et al., 2018b). To compute SRA, we use two settings, A and B. In setting A, we train and test L machine learning models using real data and obtain their area under the receiver operating characteristic curve (AUROC). We denote the set of these AUROC values by $\{A_i\}_{i=1}^L$. Similarly, in setting B, we obtain the AUROC values after training and testing the models using synthetic data, denoted $\{B_i\}_{i=1}^L$. The SRA is then computed using

$$\text{SRA} = \frac{1}{L(L-1)}\sum_{j=1}^{L}\sum_{k\neq j}\mathbb{I}\left((A_j - A_k)(B_j - B_k) > 0\right), \tag{17}$$

where $\mathbb{I}$ denotes the indicator function. A higher SRA indicates that the synthetic and real data behave similarly in downstream machine learning tasks.

4.5. Privacy disclosure

Membership inference attack We use a membership inference attack (Shokri et al., 2017) to infer whether a particular patient record was used during the training of the

Table 1: Statistics for pre-processed MIMIC-III and ED EHR datasets.

Dataset	Continuous Features	Binary features	No. of Records
MIMIC-III	0	1,071	46,520
ED EHR	63	562	232,592

synthetic data generation framework. Consider the scenario where a certain subset of features from a synthetic dataset is shared (e.g. synthetic dataset without disease labels). If an adversary is aware that the model used to generate a synthetic dataset comprised of people with a certain characteristic, then determining that a particular person's record was used during model training would allow the adversary to infer that that person has the given characteristic. This leads to a privacy issue. To carry out the attack, a random sample of k records from $\mathcal{D}^{Tr}$ and $\mathcal{D}^{Ts}$ are obtained (Choi et al., 2017). The cosine similarity for each sample from the synthetic dataset $\widehat{\mathcal{D}}$ is computed. A match is identified if the cosine similarity for any of the records is above a threshold we set (Choi et al., 2017). In this situation, there are four outcomes for the adversary: correctly predicting that a record is in $\mathcal{D}^{Tr}$ (true positive) or correctly predicting it is not in $\mathcal{D}^{Tr}$ (true negative), and incorrectly predicting that a record is in $\mathcal{D}^{Tr}$ (false positive) or incorrectly predicting that a record is not in $\mathcal{D}^{Tr}$ (false negative) (Choi et al., 2017). Using these results we compute the precision and sensitivity of the attack.

4.6. Clinician evaluation

Blinded clinician evaluation In image generation tasks, qualitative evaluation is performed by visual inspection of images to evaluate the realism or similarity of the generated images (Zhu et al., 2017). In a similar manner, we perform a qualitative evaluation of our synthetic EHRs with the help of clinicians. We selected a random set of K patients from both the training set and synthetic data, with chief complaints relevant to each clinician's speciality. The clinicians are provided with a randomly mixed set of real and synthetic data, and asked to label each patient as either realistic or not realistic.

5. Experiments

5.1. Datasets and data handling

We make use of two publicly available datasets: MIMIC-III clinical database (Johnson et al., 2016; Goldberger et al., 2000) and emergency department (ED) visits data used by Hong et al. (2018), which we refer to as ED EHR.

MIMIC-III MIMIC-III consists of 61,298 identified patient records for patients admitted in critical care units at the Beth Israel Deaconess Medical Centre in the United States from 2001 to 2012. It includes data on patient vitals, laboratory readings, diagnostic codes, imaging data, notes from healthcare providers and other clinically relevant recordings (Johnson et al., 2016).

ED EHR The ED data consists of deidentified records of 560,486 adult patient visits to three EDs within the Yale New Haven Health system between March 2014 and July 2017 (Hong et al., 2018). We select a subset of patient features from these, which include demographics, hospital usage statistics, chief complaint, past medical history, and medications.

Data pre- and post-processing is discussed in Appendix C.

5.2. Experimental setup

Autoencoder All of our models use the medGAN autoencoder, which consists of single layer feed-forward neural networks for the encoder and decoder (Choi et al., 2017). The activation functions are tanh and sigmoid for the encoder and decoder respectively. Following Baowaly et al. (2019), the GAN-based models use an encoded dimension of 128. ScoEHR uses an encoded dimension of 144, chosen to be similar to that used by the GAN-based models, but such that it meets the constraints of our chosen UNet architecture. We train the autoencoder using Adam optimisation algorithm with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Diffusion Model In the following experiments, the parameters for the time-dependent linear noise schedule $\beta(t)$ in (8) were taken as $\beta_{min} = 0.1$ and $\beta_{max} = 20$. These are the values used by Song et al. (2021). The VP SDE is run from $t = 0$ to $t = 1$.

Score-network The time-dependent score-network is a UNet with self-attention. A small hyperparameter optimisation was performed to find the UNet architecture. The final architecture for MIMIC-III was two down-sampling layers with two residual blocks per layer and the best for ED EHR was three down-sampling layers and three residual blocks per layer. The network is trained using Adam, with an adaptive learning rate

$$\text{lr} = 0.001 \left(1.0 - \frac{\max(0, \text{epoch} - 20)}{190}\right), \tag{18}$$

and $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During sample generation, the Euler-Maruyama solver is run with a time step of 10^{-3}.

5.3. Baseline models

We consider three baseline models in this paper, medGAN, medWGAN, and medBGAN (Baowaly et al., 2019). Each of these methods is based on a different GAN, specifically the original GAN (Goodfellow et al., 2014), a Wasserstein-GAN with gradient penalty (WGAN) (Arjovsky et al., 2017), and a boundary-seeking GAN (BGAN) (Hjelm et al., 2018).

All three GAN-based models have the same architecture and hyperparameters, which are taken from Baowaly et al. (2019). The generator and discriminator are multi-layer perceptrons (MLPs), where the generator has one hidden layer of width 128, and the discriminator has two hidden layers of width 256 and 128. The generator and discriminator are trained using the Adam optimisation algorithm with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Table 2: Utility metrics measured with data generated using MIMIC-III dataset for all models. The best results are indicated in **bold**.

Model	Log-cluster	SRA	PCD	Dimensional distribution
`medGAN`	-2.9 ± 0.1	0.83 ± 0.02	120 ± 8	0.0019 ± 0.0001
`medBGAN`	-3.2 ± 0.1	0.86 ± 0.03	146 ± 6	$\mathbf{0.0016 \pm 0.0001}$
`medWGAN`	-5.4 ± 0.2	0.81 ± 0.03	22.1 ± 0.4	0.0025 ± 0.0001
`ScoEHR`	$\mathbf{-6.0 \pm 0.1}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{21.8 \pm 0.3}$	0.0029 ± 0.0001

Table 3: Utility metrics measured with data generated using ED EHR dataset for all models. The best results are indicated in **bold**.

Model	Log-cluster	SRA	PCD	Dimensional distribution
`medGAN`	-3.7 ± 0.3	0.81 ± 0.02	24.1 ± 0.2	0.012 ± 0.001
`medBGAN`	-2.5 ± 0.2	0.85 ± 0.03	20.7 ± 0.5	0.014 ± 0.001
`medWGAN`	-6.1 ± 0.4	0.83 ± 0.04	$\mathbf{15.2 \pm 0.3}$	0.0088 ± 0.0001
`ScoEHR`	$\mathbf{-7.8 \pm 0.5}$	$\mathbf{0.86 \pm 0.03}$	33.6 ± 0.2	$\mathbf{0.0037 \pm 0.0001}$

5.4. Metric Evaluation

Utility metrics In order to evaluate the trained models, each model is used to generate 5 sets of $10,000$ patients. Each of these sets is then evaluated on dimension-wise distribution, pairwise correlation difference, log-cluster, and synthetic ranking agreement using a randomly chosen subset of $10,000$ from the real data. The mean and standard deviation of the metric evaluations over the 5 sets are reported for each model.

Clinician evaluation A total of 200 records containing 100 synthetic patients generated using ScoEHR and 100 real patients from ED EHR are shared with three clinicians board certified in internal medicine within the United States. Each clinician is also provided a set of instructions on how to evaluate the records. Additionally, the clinicians are told that the records may consist of either all real or all synthetic data.

Privacy disclosure Membership inference is carried out to investigate whether ScoEHR results in privacy disclosure. Two sets of experiments are performed. In the first, the number of patients known to the adversary is varied for a fixed synthetic dataset of size $50,000$. In the second experiment, we investigate the impact of increasing the size of the synthetic dataset for a fixed set of $1,000$ patients known to the adversary. We perform each experiment using four threshold values five times and report the results.

6. Results and Discussion

6.1. Utility metrics

The results for the utility metrics are reported in tables 2 and 3 for MIMIC-III and ED EHR data respectively.

Across both datasets and all models, ScoEHR performs the best with respect to log-cluster, obtaining an average of -6.0 for MIMIC-III and -7.8 for ED EHR. We note that the worst performance for log-cluster when using MIMIC-III is obtained with medGAN. In the case of ED EHR data, the worst performer is medBGAN.

It is observed that PCD is generally consistent across both datasets and all models, except when medGAN and medBGAN are used in synthesising MIMIC-III data, where they perform significantly worse. With regard to PCD, medWGAN performs the best on ED EHR while ScoEHR performs the worst. On MIMIC-III, the worst performer is medGAN and best is ScoEHR.

Regarding dimensional distribution, except for `ScoEHR No Auto`, all the models perform well. For MIMIC-III, the best performer is medBGAN, while ScoEHR performs the worst, Whilst for ED EHR, the worst performer is medWGAN and best is ScoEHR. We visualise the dimension-wise probabilities for the categorical columns for both datasets in Appendix B.

We observe that across all models and datasets, SRA is over 0.80, indicating that all models generate data suitable for downstream machine learning tasks. For both datasets, the best score is obtained on ScoEHR while medWGAN and medGAN perform the worst on MIMIC-III and ED EHR data respectively.

6.2. Clinician evaluation

We applied three different methods of aggregation to the clinician evaluations. Firstly, a patient was labelled as unrealistic if at least one clinician labelled them as unrealistic. Here, 81% of the real patients were labelled as real and 81% of the synthetic patients were labelled as real. In the second method, a patient was labelled as realistic if at least one clinician labelled them as realistic. Here, 100% of both the real and synthetic patients were labelled as real. The final method labelled a patient using the label given by the majority of the clinicians. Here, 93% of the real patients were labelled as real and 95% of the synthetic patients were labelled as real.

6.3. Privacy

In the membership inference attack, the proportion of correct decisions made by the adversary is quantified by precision. A value greater than 0.5 indicates a risk of privacy disclosure. While sensitivity is an indicator of the proportion of records correctly identified by the adversary on records the adversary has prior knowledge of being used to train the synthetic data generation model. Given that Choi et al. (2017) carried out a similar analysis on medGAN, we carry out this attack only on ScoEHR. The results of the attack are illustrated in figure 2.

In the case whereby the number of patients known to the adversary is increased, regardless of the threshold value used, we observe that the precision remains at 0.5 with increase

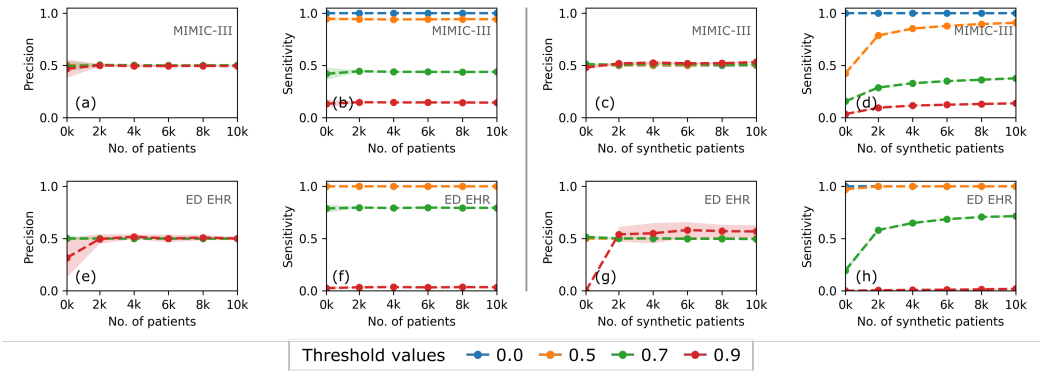

Figure 2: Precision and recall curves for membership inference attack performed on synthetic data generated using ScoEHR. The shaded region indicates standard deviation computed after three adversarial attacks. (a)-(d) results obtained on the MIMIC-III dataset. (e)-(h) results obtained on the ED EHR dataset.

in number of patients known to the adversary with some variability with fewer number of records (figure 2(a) and 2(e)). Regardless of the number of patients known to the adversary, the sensitivity for a given threshold remains constant. It is observed that with increasing threshold values, the sensitivity decreases.

Similarly, despite increasing the number of synthetic patients available to the adversary, precision remains close to 0.5 for both datasets, except when ED EHR dataset is used with a threshold of 0.9. In this case, sensitivity is not as high for fewer records, however increasing synthetic patients to 6000, for a given threshold, it remains constant.

6.4. Discussion

Of the four metrics considered, ScoEHR attains the best performance on three of the metrics on both the MIMIC-III dataset and the ED EHR dataset. Specifically, on MIMIC-III, ScoEHR achieves the best performance on log-cluster, SRA, and PCD, while for ED EHR, ScoEHR achieves the best performance on log-cluster, SRA and dimensional distribution. However, it is worth noting that on the metric where ScoEHR does not achieve the highest performance, specifically dimensional distribution for MIMIC-III and PCD for ED EHR, ScoEHR performs the worst of the models evaluated.

ScoEHR consistently achieving the best score on log-cluster metric demonstrates the ScoEHR framework's ability to capture joint distribution of all features in real data. The results on ED EHR demonstrate that this can be true, even when PCD is high. This is primarily because PCD evaluates the data generation frameworks ability to capture feature level correlations. Additionally, our results on capturing joint distribution are corroborated by the clinician evaluation in which parity is achieved between clinician evaluation with respect to the realism between real and synthetic data in the most conservative case.

The fact that ScoEHR performs the best on SRA indicates that the synthetic data generated using ScoEHR has the highest utility for downstream machine learning tasks.

However, it is important to note that SRA cannot be used in isolation, as a model could potentially perform well even if the joint distributions are not captured (Jordon et al., 2018a).

With regards to privacy disclosure, the adversary is unable to infer membership of a patient to ScoEHR's training set, irrespective of the number of patients known to the adversary or the size of the synthetic dataset. This is demonstrated by precision being at 0.5 (unless very few patients are known) regardless of variable sensitivity measurements for a given threshold. This demonstrates that the adversary is unable to make anything more than a random guess. Additionally, as remarked by Goncalves et al. (2020), it is difficult for an adversary to ascertain the optimal threshold value to use without access to two sets of patient records, one which was used to train the model and one which was not.

ScoEHR's capability to generate synthetic EHRs using diverse seed data from different geographies, practice modalities, and demographics presents an opportunity to improve the robustness and performance of multiple models and forecasts in healthcare. These improvements extend to areas such as disease progression, hospital utilisation, staffing needs, and the management of radiology, surgery, laboratory, equipment, as well as the evolving healthcare costs.

7. Limitations

Despite ScoEHR's ability to generate high fidelity structured EHRs, we did not consider both unstructured and temporal EHR data. However, often in healthcare settings EHR data is found in this manner. We believe extending ScoEHR framework to support such data is an important next step. In addition, even though we are able to demonstrate low risk of membership inference empirically, our model can be improved to incorporate a more rigorous privacy guarantee through differential privacy (Dwork, 2006).

8. Conclusion

In this study, we develop and validate a novel deep learning method, ScoEHR, for generating synthetic EHRs that combines an autoencoder with a continuous time diffusion model. Comparison of ScoEHR to state-of-the-art EHR generation models, medGAN, medWGAN, and medBGAN, demonstrates the superior performance of ScoEHR on two datasets, MIMIC-III and Yale New Haven Health System Emergency Department, based on four metrics of data generation utility: 1) preservation of feature marginal relationships, 2) preservation of feature correlations, 3) preservation of full feature distribution, and 4) synthetic data performance in downstream modelling. Expert clinical opinion in a blinded experiment further showed that ScoEHR generated data and real EHR data were scored as equally realistic. Finally, a privacy study showed that ScoEHR showed low risk of privacy disclosure that could give a potential attacker knowledge of the identity of any real seed data used during generation.

The methods used in this study can be generalised for synthetic record generation in a variety of fields and applications. Moreover, this work has important implications in accelerated testing of medical software, hardware, disease modelling and research, global

inequities in data access, problems with EHR, data bias, and the economics of healthcare delivery.

Code and data availability

The code is available at https://github.com/aanaseer/ScoEHR. All of the datasets used in our experiments are publicly available.

Acknowledgements

Benjamin Walker was funded by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). Terry Lyons was funded in part by the EPSRC [grant number EP/S026347/1], in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1, the Data Centric Engineering Programme (under the Lloyd's Register Foundation grant G0095), the Defence and Security Programme (funded by the UK Government) and the Office for National Statistics & The Alan Turing Institute (strategic partnership) and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. 2017. doi: 10.48550/ARXIV.1701.07875. URL https://arxiv.org/abs/1701.07875.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

Anna L Buczak, Steven Babin, and Linda Moniz. Data-driven approach for creating synthetic electronic medical records. *BMC medical informatics and decision making*, 10(1): 1–28, 2010.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Ghadeer Ghosheh, Jin Li, and Tingting Zhu. A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources. *arXiv preprint arXiv:2203.07018*, 2022.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 2022.

Andres Hernandez-Matamoros, Hamido Fujita, and Hector Perez-Meana. A novel approach to create synthetic biomedical signals using birnn. *Information Sciences*, 541:218–241, 2020.

R Devon Hjelm, Athul Paul Jacob, Adam Trischler, Gerry Che, Kyunghyun Cho, and Yoshua Bengio. Boundary seeking gans. In *International Conference on Learning Representations*, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Woo Suk Hong, Adrian Daniel Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7): e0201016, 2018.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Measuring the quality of synthetic data for use in competitions. *arXiv preprint arXiv:1806.11345*, 2018a.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018b.

Young J Juhn, Euijung Ryu, Chung-Il Wi, Katherine S King, Momin Malik, Santiago Romero-Brufau, Chunhua Weng, Sunghwan Sohn, Richard R Sharp, and John D Halamka. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the houses index. *Journal of the American Medical Informatics Association*, 29(7):1142–1151, 2022.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

Emily Kogan, Kathryn Twyman, Jesse Heap, Dejan Milentijevic, Jennifer H Lin, and Mark Alberts. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC medical informatics and decision making*, 20(1):1–8, 2020.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

Scott McLachlan, Kudakwashe Dube, and Thomas Gallagher. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 439–448. IEEE, 2016.

Amanda Momenzadeh, Ali Shamsa, and Jesse G Meyer. Bias or biology? importance of model interpretation in machine learning studies from electronic health records. *JAMIA open*, 5(3):ooac063, 2022.

James Morrill, Klajdi Qirko, Jacob Kelly, Andrew Ambrosy, Botros Toro, Ted Smith, Nicholas Wysham, Marat Fudim, and Sumanth Swaminathan. A machine learning methodology for identification and triage of heart failure exacerbations. *Journal of Cardiovascular Translational Research*, 15(1):103–115, 2021.

Beau Norgeot, Benjamin S Glicksberg, Laura Trupin, Dmytro Lituiev, Milena Gianfrancesco, Boris Oskotsky, Gabriela Schmajuk, Jinoos Yazdany, and Atul J Butte. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA network open*, 2(3):e190606–e190606, 2019.

Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, 8 (7):e18910, 2020.

Juan C Rojas, Kyle A Carey, Dana P Edelson, Laura R Venable, Michael D Howell, and Matthew M Churpek. Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, 15(7): 846–853, 2018.

Yue Ruan, Alexis Bellot, Zuzana Moysova, Garry D Tan, Alistair Lumb, Jim Davies, Mihaela Van Der Schaar, and Rustam Rea. Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. *Diabetes care*, 43(7):1504–1511, 2020.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Sumanth Swaminathan, Klajdi Qirko, Ted Smith, Ethan Corcoran, Nicholas Wysham, Gaurav Bazaz, and Anthony N. Gerber. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS One*, 12(11):e0188532, 2017.

Sumanth Swaminathan, James Morrill, Klajdi Qirko, Ted Smith, Nicholas Wysham, and Botros Toro. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *European Respiratory Journal*, 56(Suppl 64):1356, 2020.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021.

Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Journal of the American Medical Informatics Association*, 28(11):2393–2403, 2021.

Amirsina Torfi and Edward A Fox. Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. In *The Thirty-Third International Flairs Conference*, 2020.

Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin, Heidelberg, 2008.

Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D Mooney, and Bradley A Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1):1–18, 2022.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108, 2020.

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

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Appendix A. Autoencoder Ablation Study

In order to evaluate the effectiveness of the autoencoder in ScoEHR, an ablation study is performed, where ScoEHR is trained on the Hong dataset without the autoencoder. As the dimension of the Hong dataset does not meet the requirements of our chosen UNet architecture, the data is zero-padded to have dimension 784.

As can be seen in table 4, removing the autoencoder is severely detrimental to log-cluster and dimensional distribution. Furthermore, we were unable to calculate SRA using the data generated without an autoencoder, as all of the data had positive labels. However, the pairwise correlation difference improves when the autoencoder is removed. This demonstrates the importance of analysing synthetic data generation models using various metrics, as if you were to just consider PCD, removing the autoencoder would seem to be a benefit, when it is actually very detrimental.

Table 4: Utility metrics measured with data generated using ED EHR dataset for `ScoEHR` and `ScoEHR No Autoencoder`.

Model	Log-cluster	SRA	PCD	Dimensional distribution
`ScoEHR No Auto`	-1.4 ± 0.1	–	22.8 ± 0.1	5.0 ± 0.1
`ScoEHR`	-7.8 ± 0.5	0.86 ± 0.03	33.6 ± 0.2	0.0037 ± 0.0001

Appendix B. Dimension-wise Probabilities

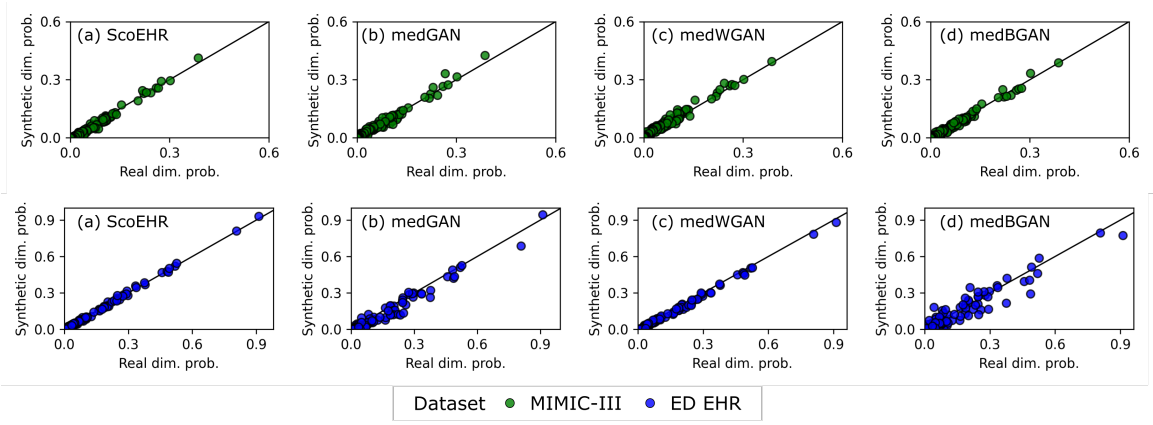

Figure 3: Dimension-wise probabilities for the binary columns in both datasets. Top row (a)-(d) are plots generated using MIMIC-III data. Each point corresponds to one of the features in the dataset. Bottom row (a)-(d) are plots generated using ED EHR data. Each point corresponds to one of the binary features in the dataset.

Appendix C. Data Processing

Data pre-processing For the MIMIC-III dataset, we pre-process the data in the same way as Choi et al. (2017), after extracting ICD-9 codes generalised up to the first three digits, longitudinal records for patients are aggregated as a binary dataset. For the ED EHR, we one-hot encode all the categorical features, use a cyclical encoding with sin and cosine transformation for date and day, and perform a min-max scaling on the continuous and count features.

Data post-processing After generating data, data corresponding to those in binary columns are rounded up to one if the value generated is greater than or equal to 0.5, and zero otherwise. We also reverse the sin and cosine time-encoding and min-max scaling for the continuous features.