Generating Electronic Health Records with Multiple Data Types and Constraints

Chao Yan*, MS[1], Ziqi Zhang*, BS[1], Steve Nyemba, MS[2], Bradley A. Malin, PhD[1,2]
[1]Vanderbilt University, Nashville, TN; [2]Vanderbilt University Medical Center, Nashville, TN

Abstract *Sharing electronic health records (EHRs) on a large scale may lead to privacy intrusions. Recent research has shown that risks may be mitigated by simulating EHRs through generative adversarial network (GAN) frameworks. Yet the methods developed to date are limited because they 1) focus on generating data of a single type (e.g., diagnosis codes), neglecting other data types (e.g., demographics, procedures or vital signs), and 2) do not represent constraints between features. In this paper, we introduce a method to simulate EHRs composed of multiple data types by 1) refining the GAN model, 2) accounting for feature constraints, and 3) incorporating key utility measures for such generation tasks. Our analysis with over 770,000 EHRs from Vanderbilt University Medical Center demonstrates that the new model achieves higher performance in terms of retaining basic statistics, cross-feature correlations, latent structural properties, feature constraints and associated patterns from real data, without sacrificing privacy.*

Introduction

Electronic health records (EHRs) have been widely adopted by healthcare organizations (HCO) to provide more accurate, timely and safer care for patients. Though initially designed to improve the efficiency of healthcare delivery, EHRs have shown great promise in secondary endeavors, including facilitating software systems development, enabling clinical training, and boosting biomedical research.[1,2] As a result, HCOs are incentivized to share EHR data, but are concerned that doing so could lead to privacy intrusions and loss in trust.[3,4] Various computational approaches have been proposed to maintain patient anonymity and confidentiality[5], but providing raw data, at any degree of specificity, leads to a tradeoff between privacy and data utility.[6] Moreover, as the amount of data alteration made to real data grows, one realizes greater privacy protections at the cost of lower utility.

Recent advances make it possible to alleviate such tensions by enabling the generation of synthetic EHR data with the look and feel of real data. If generated appropriately, synthetic data has a low risk of being linked to the real individuals that were in the training data for the model, which mitigates privacy concerns. As such, data simulation enables an opportunity to widely share data of a more granular nature, which is not often possible through traditional de-identification mechanisms. Deep learning approaches based on generative adversarial networks (GANs)[7,8] have demonstrated an uncanny ability to simulate realistic-looking data instances with high statistical generalizability, scalability and limited reliance upon knowledge drawn from domain experts.[9–11] This is because GAN-based models are trained in an adversarial environment, where a generator produces increasingly realistic instances, such that an evolving discriminator cannot distinguish them from real data.

However, there are several gaps between current GAN-based EHR simulation approaches and the practical needs of simulation and evaluation. First, current models are designed to simulate only one type of data, such as International Classification of Diseases (ICD) codes. Yet EHRs are an amalgamation of a multitude of types, such as demographics, procedures, medications, laboratory test results, and vital signs. Beyond differences in semantics, the data has different syntax (being composed of binary, categorical and continuous values), which induces higher dimensionality and greater sparsity. Second, there are various record-level constraints that exist between features. For example, the results of a blood pressure test should indicate that systolic pressure is greater than diastolic pressure. However, such constraints are not directly embedded in a GAN-based generative model, leading to situations where the corresponding semantics are violated in the generated dataset, which weakens the overall utility of synthetic EHRs.

Moreover, the current array of data utility measures for synthetic EHRs are deficient in several respects. First, they do not assess record-level consistency, which is the extent to which frequently associated conditions or events in real EHRs are sufficiently maintained by the generative model in records (e.g., correlated diseases and common diagnosis-procedure pairings). This is a concern because poor record-level consistency can invalidate record-level studies in the

* Equal contribution

synthetic data. Second, to simulate EHRs with multiple data types, we need to measure if the conditional distribution of one data type is well-represented with respect to another. For instance, the correlation between the distribution of blood pressure with respect to a certain diagnosis (e.g., hypertension) should be similar in the real and synthetic data.

In this paper, we address the aforementioned challenges by refining GAN-based generative models and enriching the current set of data utility measures. Specifically, we 1) incorporate a penalization into the GAN learning process such that, if a violation transpires during training, then the generator will be penalized and, thus, forced to output records with more desirable feature constraints; 2) refine the deep neural networks of both the generator and discriminator to make signal sparser and, thus, more efficient, 3) introduce measures for constraint violations, feature associations, and conditional distributions to assess EHR simulation models. We then demonstrate the effectiveness of the new model (as well as the privacy risks) and measures, using a dataset based on over 770,000 EHRs from Vanderbilt University Medical Center (VUMC).

Related Work

In this section, we review recent GAN-based developments in EHR simulation. Choi et al. first customized a GAN framework to generate a set of structured and discrete features in the form of ICD-9 codes.[10] Recognizing that the original GAN framework[7] could not generate discrete values, an autoencoder was incorporated to learn the distribution of discrete data and a trained decoder was applied to generate data in a discrete space. A limitation of this model was that it could suffer from a mode collapse (that is, the generator maps different inputs to the same output) and a mode drop (that is, the generator only captures certain regions of the underlying distribution of the real data). To mitigate these problems, Baowaly et al. introduced an approach that used Wasserstein divergence to more effectively measure difference in real and synthetic data distribution.[11] Zhang et al. enhanced the billing code simulation model by removing the auto-encoder and incorporating utility measures to evaluate structural properties of the data.[9] Still, all of these techniques focused on generating only a single feature type. Recently, Chin-Cheong et al. explored the use of off-the-shelf models[12,13] to generate EHRs with more than one data type.[14] However, their work did not improve the learning model, nor did it address feature constraints. They also did not evaluate the data with respect to consistency between features. By contrast, in this paper, we focus on model refinement and evaluation in more challenging scenarios than previous simulation settings.

Data Overview

The data in this study was derived from the VUMC Synthetic Derivative (SD), a de-identified warehouse of over 2.2 million EHRs. We collected the EHRs of patients with at least one recorded visit during a ten-year period (2005-2015). For each EHR, we extracted several types of data: 1) age (at the latest time in the resource), 2) gender, 3) ICD-9 codes, 4) Current Procedural Terminology-Fourth Version (CPT-4) codes, 5) body mass index (BMI), and 5) systolic and diastolic blood pressures. We restricted our analysis to EHRs with at least one documented BMI and blood pressure reading. We note that we use BMI and blood pressure due to the fact that they are frequently populated in EHRs and are common covariates in biomedical research. We rolled up 1) ICD codes to their subcategories (by removing the portion of the codes to the right of the ".") and 2) CPT codes to the minor categories (including 115 distinct categories)[15]. As such, through the remainder of this paper, ICD and CPT codes refer to their rolled-up versions. We refer to this dataset, which includes 928,089 records, as the SD.

To mitigate noise in the data, we further refined the set of records. First, we ranked all ICD codes based on their prevalence and removed those with a rate less than 1/1000 (which corresponded to 928 patients). The same process was applied to CPT codes. Second, we removed records that were composed of less than 5 distinct ICD and CPT codes. Third, we removed obviously-errored test results, which specifically corresponded to cases where 1) systolic pressure was less than diastolic pressure in the same observation, 2) systolic pressure was greater than 300, and 3) BMI was less than 5. We refer to the refined dataset as the cleaned SD (or CSD). This dataset contains 770, 231 records, 693 distinct ICD codes and 65 CPT codes. Summary statistics for both datasets are provided in Table 1.

Table 1: Summary statistics of the EHR datasets.

Dataset	Patients	Gender	ICD Codes	ICD Codes Per Patient	Patients Per ICD Codes	CPT Codes	CPT Codes Per Patient	Patients Per CPT Codes	BMI Instances Per Patient	Blood Pressure Instances Per Patient
SD	928, 089	M:43% F:57%	926	12.03	15, 208	88	7.02	104, 044	10.32	33.60
CSD	770, 231	M:44%: F:56%	693	13.88	20, 221	65	8.00	140, 815	12.05	40.05

For reference, Table 2 illustrates the format of a record, where the second row indicates the length of the sub-vector needed to represent each feature space for the CSD dataset. As can be seen, the CSD contains multiple types of features and two syntactic types, which are binary and continuous. To incorporate a statistical summary of the vital signs of each record, we computed the minimum, median and maximum values of BMI, systolic, and diastolic pressure, respectively. Each patient's record can then be represented as a vector over age, gender, ICD codes, CPT codes, minimum BMI (systolic and diastolic), median BMI (systolic and diastolic) and maximum BMI (systolic and diastolic). Note that categorical features can be embedded by applying a one-hot encoding strategy, which leads to a binary representation.

Table 2: Data representation of the CSD dataset. (*B*: binary; *C*: continuous).

Data Type	Age	Gender	ICD Codes	CPT Codes	BMI			Systolic			Diastolic		
					Min	Median	Max	Min	Median	Max	Min	Median	Max
Variable Type	B	B	B	B	C	C	C	C	C	C	C	C	C
Length	100	1	693	65	1	1	1	1	1	1	1	1	1

Constraints. In addition to the heterogeneity in data types, there are multiple constraints that can exist between features in practical data synthesis tasks. Consider the CSD dataset. In each record, the minimum value of a continuous feature (i.e., BMI, systolic and diastolic pressure) should be no greater than its corresponding median value. Similarly, the median value of a continuous feature should be no greater than its maximum value. The generative model needs to capture these constraints so that it can simulate synthetic EHRs without violating them.

Method

In this section, we introduce the GAN-based generative model for the simulation of EHR data with heterogenous data types and feature constraints. We then define new utility measures to assess the quality of synthetic data.

Generative Model

Preliminary of GAN in EHR simulation. There are two core components in GANs for EHR data simulation: 1) a generator neural network $G(\mathbf{z}, \theta_g)$ which accepts random noise $\mathbf{z} \in \mathbb{R}^r$ and 2) a discriminator neural network $D(\mathbf{x}, \theta_d)$ which accepts EHR data represented as vector $\mathbf{x}$ (θ_g and θ_d denote the parameters of G and D, respectively). These two components are updated iteratively and alternatively through competition with one another. The generator $G(\mathbf{z}, \theta_g)$ is trained to minimize the statistical divergence between the distribution of real data $\mathbb{P}_r$ and the distribution of generated data $\mathbb{P}_g$; i.e., $\min_G Div(\mathbb{P}_r, \mathbb{P}_g)$. Ideally, G is able to learn $\mathbb{P}_r$ and generate synthetic records that are indistinguishable to the discriminator D. By contrast, the discriminator D is trained to distinguish synthetic data generated by G from real data. In EHR simulation, the state-of-the-art models[9] adopt Wasserstein Divergence[12] to characterize the earth mover distance between two distributions. When combined with a gradient penalty technique[13], such a divergence measure can effectively mitigate the problem of poor convergence, which is the main cause of mode collapse and mode drop.

Heterogeneous GAN. We build our generative model, which we henceforth refer to as *Heterogeneous GAN* (HGAN), based on the basic structure of the state-of-the-art model, *EMR-CWGAN*[9]. The architecture is shown in Figure 1. First, we apply a conditional generation framework, where the conditional batch normalization[16,17] (in G) and conditional layer normalization (in D) are leveraged to control the specific generation and discrimination. In this work, we use age and gender as the conditioning features to simulate EHR data. Specifically, we build one distributed representation vector (i.e., an embedding) for each integer age and one embedding vector for each gender. We train HGAN using a batch of records from CSD in each training step and then use their conditioning features to extract the associated embedding representations to build the normalization layers.

Second, we reorder the neural networks for both the generator and discriminator. Instead of applying *conditional normalization* $\rightarrow$ *rectified linear unit (ReLU) layer* between fully connected layers (as adopted by EMR-CWGAN[9]), we use *ReLU* $\rightarrow$ *conditional normalization* $\rightarrow$ *ReLU* to filter signals. Though the former implementation has been widely utilized in the deep learning community[18], we anticipated that our design can make data sparser, which helps disentangle the signals and, thus, make the representation robust and efficient.[19] We confirm this expectation in the experimental results.

Third, to encourage simulated EHRs to adhere to the record-level feature constraints, we incorporate a penalization term as part of the loss function of the GAN model. Formally, the *minmax* objective with gradient penalty and

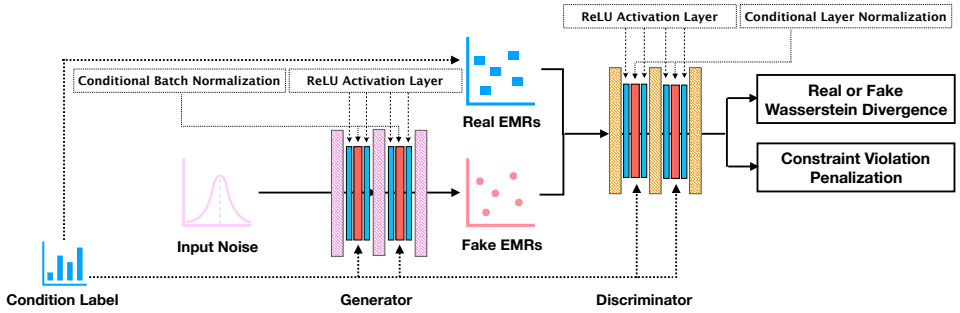


Figure 1: An architectural representation of an HGAN.

constraint violation penalty becomes:

$$\min_G \max_D \underbrace{\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_r}[D(\mathbf{x})] - \mathbb{E}_{\widetilde{\mathbf{x}}\sim\mathbb{P}_g}[D(\widetilde{\mathbf{x}})]}_{\text{Original Discriminator Loss}} - \underbrace{\lambda\mathbb{E}_{\widehat{\mathbf{x}}\sim\mathbb{P}_{\widehat{\mathbf{x}}},\delta\sim N_d(\mathbf{0},aI))}[(||\bigtriangledown_{\widehat{\mathbf{x}}} D(\widehat{\mathbf{x}}+\delta)||_2 - \mathbf{1})^2]}_{\text{Gradient Penalty}} + \underbrace{\beta\sum_{c\in\mathcal{C}}\mathbb{E}_{\widetilde{\mathbf{x}}\sim\mathbb{P}_g}[P(\widetilde{\mathbf{x}},\mathcal{F}_c(\cdot))]}_{\text{Constraint Violation Penalty}}, \tag{1}$$

where the first term denotes the classification loss of D and the second term penalizes the situation where the gradient of D is far from 1. In the third term, c denotes a feature constraint from set $\mathcal{C}$ and $P(\widetilde{\mathbf{x}}, \mathcal{F}_c(\cdot))$ represents the quantity of penalty applied to the synthetic record $\widetilde{\mathbf{x}}$ when using the constraint-specific penalization function $\mathcal{F}_c(\cdot)$. The coefficient λ and β control the weights of the two penalty terms during optimization. Recall that in CSD there is one type of constraint within multiple continuous feature pairs due to their ordinal relationship in semantics (e.g., for each type of vital signs, the *min* value is no greater than the *median* value, and the *median* value is no greater than the *max* value). We define the constraint violation penalty function for such a constraint as:

$$\mathcal{F}_c(x) = \max\{f_1(x) - f_2(x), 0\}, \tag{2}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ denote the feature extractor of EHR data, which correspond to *min* and *median* values for the former case and *median* and *max* values for the latter case, respectively. Another constraint is the mutual exclusiveness between two binary features. In this case, the constraint violation penalty function can be defined with the format $\mathcal{F}_c(x) = \omega[f_1(x) \cdot f_2(x)]$. Consider an example that a synthetic EHR of a man cannot own the CPT code "59400 vaginal delivery". In this case, $f_1(x)$ and $f_2(x)$ represent the gender (1 for male) and the CPT code 59400 (1 for presence), respectively. As long as both values are close to 1 in the training process, the generator will be heavily penalized. The existence of feature constraints depends on the feature space for EHR simulation tasks and $\mathcal{F}(\cdot)$ can be flexibly defined based on specific constraints.

Utility Measures

Various utility measures for EHR simulation have been proposed[9,10,20], including dimension-wise statistics (DWS), dimension-wise prediction (DWP), latent space representation (LSR), and first-order proximity (FOP). DWS investigates the degree to which the distribution of each binary feature (e.g., ICD or CPT code) in the synthetic EHR records is similar to real data. By contrast, DWP evaluates the degree to which a generative model captures the cross-dimensional relationships of real data. LSR and FOP measure the ability of a generative model to maintain the structural properties of real data in the latent and original space, respectively. Due to space limitations, we refer the reader to the supplemental materials of a recent paper[9] for the implementation details of these measures. In this section, we introduce three new utility measures to provide greater intuition into the quality of simulated EHR data in the general simulation scenarios.

Constraint Violation Test (CVT). Given that there may be known constraints between features in EHR data, we need to determine the extent to which the synthetic data satisfy these constraints. Though there may be a large number of constraints among the features in an EHR dataset, in this paper we focus on solving the ordinal relationship between

continuous features. To assess if the minimum value is no greater than the median value, we compute the difference between the values in median and minimum columns of each vital sign (including BMI, systolic and diastolic pressure) in each synthetic record, and then investigate whether this difference is positive. Similarly, we perform this evaluation for the max versus median case for each vital sign. As a baseline, for each setting we provide the corresponding distribution of difference obtained from the real data. An ideal generative model can simulate EHR data that obeys the record-level feature constraints and has a similar distribution with the real data.

Frequent Association Rules (FAR). This utility measure investigates the extent to which the record-level medical condition associations in real EHRs maintained in the synthetic ones. This measure functions over a set of categorical features, such as a mixture of ICD and CPT codes. The two key criteria in association rule mining are *support* and *confidence*. Support is an indication of how frequently the condition set appears in the dataset, whereas confidence is an indication of how often a condition rule is true. With respect to rule mining in EHR, the support of condition set X (e.g., a set of diseases and procedures) with respect to record dataset T is defined as the proportion of records in T that contain X. The confidence of a condition rule, $X \Rightarrow Y$, with respect to T, is the proportion of records that contain X that also contain Y.

We first obtain all frequent condition sets (FCS), which form a set $\mathcal{S}$ with frequency larger than a threshold min_s, such that any subset of any FCS is not in $\mathcal{S}$. For each FCS $f \in \mathcal{S}$, we then determine the set of association rules $R : f' \Rightarrow f - f'$, where each rule satisfies that the number of records which have f' also have f is greater than a threshold min_c. By applying such a process to both real and synthetic EHR data, we measure the proportion of the association rules that are from the synthetic data that are in the real data and vice versa, which we refer to as recall and precision, respectively. We use a popular association rule mining technique – *Apriori*[21] – to learn FCSs and the association rules from the real and synthetic EHRs. It is notable that FAR can be regarded as an expansion on the structural measure FOP. This is because FAR does not limit the number of features to consider and, thus, consider deeper and broader dependencies between features. By contrast, FOP focuses on the condition sets containing only two features. As a consequence, in this paper we report the FAR results, instead of FOP.

Cross-type Conditional Distribution (CCD). This utility measure evaluates the ability of a generative model to maintain the distribution of one data type conditioned on another. Since the correlation between ICD and CPT codes (binary one-hot representation) can be extracted and evaluated by FAR, we focus on the correlation between continuous and binary features. Specifically, we investigate the distribution of vital sign features conditioned on ICD and CPT codes. We compare the mean and standard deviation of each conditional distribution for real and synthetic EHR data. Similarly, we investigate the distributions of the conditioning labels of GAN model (i.e., age and gender in this study) on each ICD and CPT codes. This indicates the degree to which the demographic distributions for each health condition and procedure are similar in the real and synthetic data.

Privacy Measures

We investigate the extent to which the synthetic data is susceptible to *membership* and *attribute inference* attacks.[10] For the membership inference, the attacker is assumed to have the entire records of a set of real patients and attempts to infer which patients are in the training dataset of the generative model. This is achieved by calculating the Hamming distance between each compromised record and each synthetic record. We then apply a threshold such that one compromised record has a distance (to any synthetic record) less than the threshold is considered in the training dataset. For the attribute inference, an attacker is assumed to possess a subset of features of certain real EHRs and aims to infer the value of a missing feature. k-nearest neighbors algorithm is adopted to determine the missing values. As these two approaches were designed for the situation where all features are represented in a binary form, we uniformly discretize all continuous features into categories and then use one-hot distributed representations.

Results

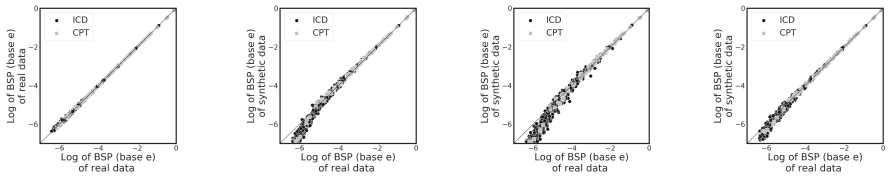
In this section, we compare the utility of synthetic EHR data generated by our model and several alternative models. The first alternative is the state-of-the-art model, *EMR-CWGAN*, which generates one data type, and, thus, will be relied upon to investigate whether generating a mixture of data types decreases the data utility on the binary features (i.e., ICD and CPT codes). To investigate how the new reordered filter *ReLU* $\rightarrow$ *conditional normalization* $\rightarrow$ *ReLU* between fully connected layers influences data generation, we design the second alternative model by applying the

filter in EMR-CWGAN to HGAN, which is *conditional normalization* $\rightarrow$ *ReLU*. We refer to this model as *HGAN-U*. It should be noted that, for evaluation purposes, we first train HGAN and HGAN-U and use them to generate heterogeneous records with the feature space depicted in Table 2. We then extract the features related to specific data utility measure to evaluate the performance. In addition, we build a baseline by randomly partitioning the real dataset into two equal sized datasets to construct a real *vs* real setting. This allows us to derive an upper bound on what an ideal generative model can achieve with respect to each utility measure. Note that we evaluate ICD and CPT as one set of features as both are one-hot encoded.

Experimental Setup

To compare EHR data simulation methods, we fixed the hyperparameter settings for all models in experiments. Specifically, we structured the generator and discriminator with a network formation of (128, 256, 256, 512, 512, 512, 512, 767) and (767, 512, 384, 256, 256, 128, 128, 1), respectively. To construct the conditional normalization layers, we set the length of the embedding vectors for age and gender to 96 and 32, respectively. All generative models were trained over $1,000$ epochs. We applied the Adam optimizer[22] with a learning rate of $4 \cdot 10^{-6}$ and $2 \cdot 10^{-5}$ for the generator and discriminator, respectively.

Dimension-wise statistics (DWS). The DWS results are shown in Figures 2(a)-2(d), where Figure 2(a) describes the real *vs* real setting. In all other subfigures, the x-axis corresponds to the Bernoulli success probability (or incidence rate) of one (ICD or CPT) code in the CSD dataset and the y-axis corresponds to this probability in the synthetic data. The 45 degree diagonal line corresponds to what a perfect correlation would indicate. There are several findings to highlight. First, the incidence rates of codes in the real *vs* real setting are closely distributed along the diagonal line, which indicates highest stability of basic statistics. Second, by comparing Figures 2(b) and 2(d), it can be seen that HGAN achieves very similar performance to EMR-CWGAN for codes with high prevalence ($x > -4$) and induces less bias with respect to codes with low prevalence ($x < -5$). Third, by comparing Figures 2(c) and 2(d), similar observations can be made in that HGAN-U exhibits a biased and less stable result for low prevalence codes ($x < -5$) than HGAN. This evidence suggests that HGAN provides a more faithful representation of the basic statistics in real EHR data. Fourth, it can be seen from all Figure 2 subfigures that the difference in the distributions between the two code types (i.e., ICD and CPTs) is negligible, which implies that the learning is unbiased between them.

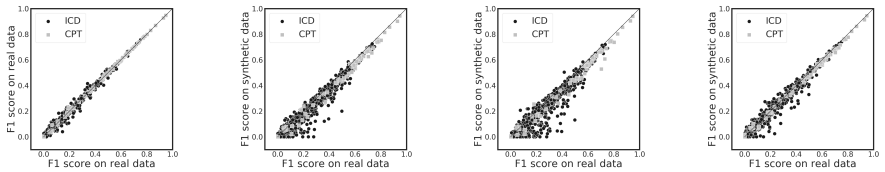


(a) Real (b) EMR-CWGAN (c) HGAN-U (d) HGAN

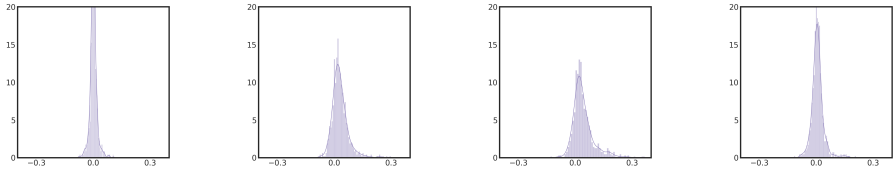
Figure 2: Dimension-wise statistics. Bernoulli success probabilities (BSP) in logarithmic scale for ICD and CPT codes.

Dimension-wise prediction (DWP). For each (ICD or CPT) code, two logistic regression classifiers were trained–one on the real data and one on synthetic data (with the same number of records as the real data). The binary status of a code served as the dependent variable while all remaining codes served as the independent variables. These classifiers were evaluated on a test dataset of the real EHRs (20% of CSD). In the first row of Figure 3, each point denotes an ICD or CPT code, whose x and y value are the *F1* score of the model trained on real and synthetic data, respectively (except 3(a)–the real *vs* real setting). There are several notable findings. First, Figure 3(a) shows that the F1 scores in the real *vs* real setting are closely distributed along the diagonal line without obvious skew. As such, the corresponding distribution of dot-to-diagonal distances, as shown in 3(e), is roughly symmetric, which indicates the stability of the cross-dimensional relationship in the original system. Second, the dot-to-diagonal distribution shown in Figure 3(h) is less skewed than the one in Figure 3(f), which indicates that HGAN is better at representing the cross-dimensional relationships than EMR-CWGAN. Third, as depicted in Figure 3(g), it can be seen that there is skew towards the real data, suggesting that HGAN-U reduces the ability to learn relationships between features. Fourth, it can be seen from Figure 3(d) that both types of codes behave similarly in the real *vs* HGAN setting, which indicates that there is high similarity in the cross-dimensional performance between two types of codes.

Latent space representation (LSR). At a high level, the investigation into LSR has three main steps: 1) use real



(a) Real (b) EMR-CWGAN (c) HGAN-U (d) HGAN



(e) Distribution of dot-to-diagonal distance of (a) (f) Distribution of dot-to-diagonal distance of (b) (g) Distribution of dot-to-diagonal distance of (c) (h) Distribution of dot-to-diagonal distance of (d)

Figure 3: Dimension-wise prediction. (a) F1 scores of logistic regression classifiers in the real *vs* real setting. (b, c, d) Results of real *vs* synthetic setting for the generative models. (e, f, g, h) Distributions of shortest distances from dots to the diagonal line for panels (a), (b), (c), and (d), respectively.

data to train a β-Variational Auto-Encoder (β-VAE)[23], a tool to disentangle the latent factors in data by forgetting less important latent dimensions, to discover their efficient latent dimensions for data reconstruction, as well as the corresponding distribution of variances, 2) input the generated data into the trained β-VAE model and assess their variance distributions on the latent dimensions, and 3) for each efficient latent dimension, assess whether the distribution of variances obtained from 2) is pushed closer to 1, which is an indicator of losing important structural properties in real data. Figure 4 shows the results for LSR, where we selected all efficient latent dimensions with the mean of the variance distribution in the real data less than 0.70. Both EMR-CWGAN and HGAN are highly similar to the real data, suggesting that both retain the latent structural properties. By contrast, HGAN-U exhibits a large shift in the variance distributions in all dimensions, indicating it is less capable of representing key structures in the latent space.

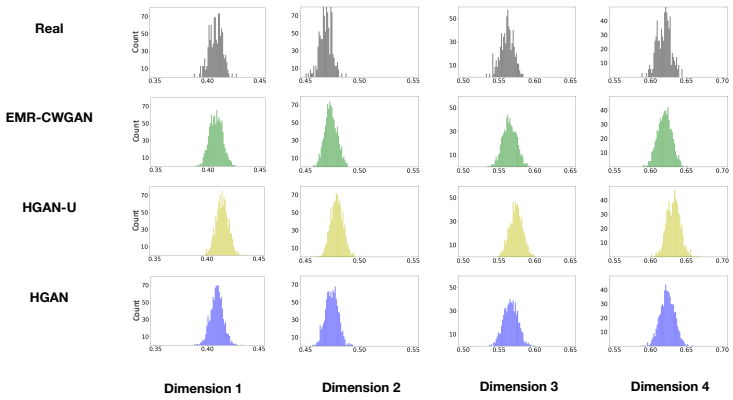


Figure 4: Latent space representation. Each subfigure illustrates the distribution of the variances in one latent dimension. The topmost row corresponds to the real EHR data. Each subsequent row corresponds to a synthetic data generation model.

Constraint Violation Test (CVT). To evaluate the effectiveness of the constraint violation penalty, we consider the following scenarios. First, for each vital sign (BMI, systolic, and diastolic pressure), we compute the record-level difference (*max-median*) and (*median-min*). Second, for each of the basic statistics (*max*, *median*, and *min*), we construct the distribution of the difference in record-level systolic and diastolic pressure. Third, we adopt two baselines to compare our model with. The first baseline is the real data, while the second is HGAN without the constraint violation penalty. Figure 5 shows the CVT results, where the first row is the comparison of real *vs* HGAN and the second row is real *vs* HGAN without the constraint violation penalty. As can be seen in the first six columns, all of

the difference distributions built from HGAN without a penalty have negative values, which suggests violations of the constraints. By contrast, HGAN always ensured the difference was positive. From the final three columns, it can be seen that HGAN (with the penalty) yield a more similar distribution with the real data than HGAN without the penalty. This indicates that HGAN is able to solve the feature constraint violation problem in our generation settings.

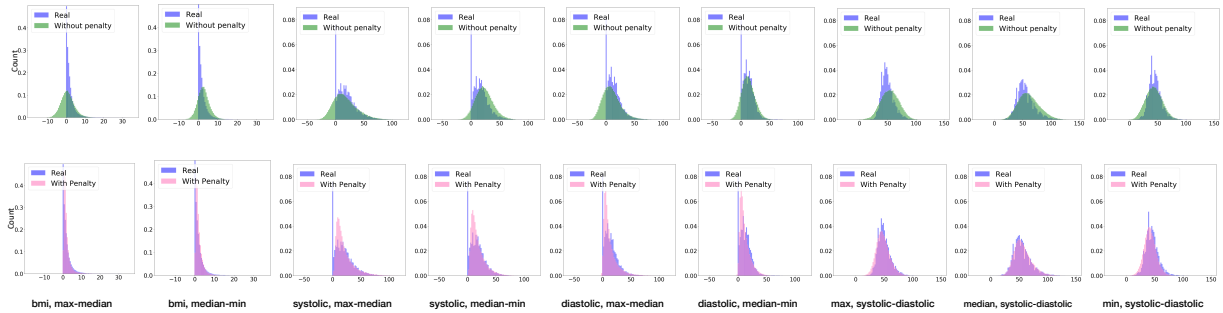
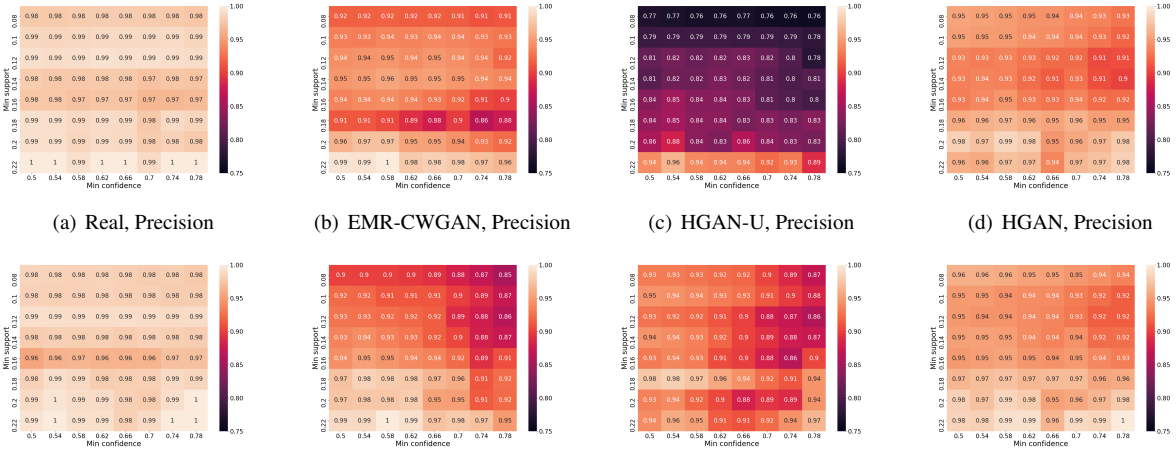


Figure 5: Constraint violation test. *Without penalty* corresponds to the model HGAN without constraint violation penalty. *With penalty* corresponds to HGAN. Distributions marked with *Real* are built from real EHR data.

Frequent Association Rules (FAR). Figure 6 shows the results for FAR. Given a range of thresholds for support ($min_s \in [0.08, 0.22]$) and confidence ($min_c \in [0.50, 0.78]$), we measure both the precision (Figures 6(a)–6(d)) and recall (Figures 6(e)–6(h)) of association rules (learned from ICD and CPT codes). There are multiple observations to highlight. First, as can be seen from Figures 6(a) and 6(e), both the precision and recall from two equal-sized subsets of real EHR data are close to 1. This indicates that the association rules in the real data are robust. Second, by comparing Figure 6(d) with 6(c) and Figure 6(h) with 6(g), it can be observed that HGAN almost always outperforms HGAN-U in terms of precision and recall. Third, by comparing Figures 6(h) with 6(f), it can be seen that the recall of HGAN is greater than EMR-CWGAN in the majority of conditions. However, there is no obvious domination in precision when comparing Figures 6(d) with 6(b). These results suggest that HGAN is more capable at maintaining the record-level feature association (or record-level consistency) than other baselines.



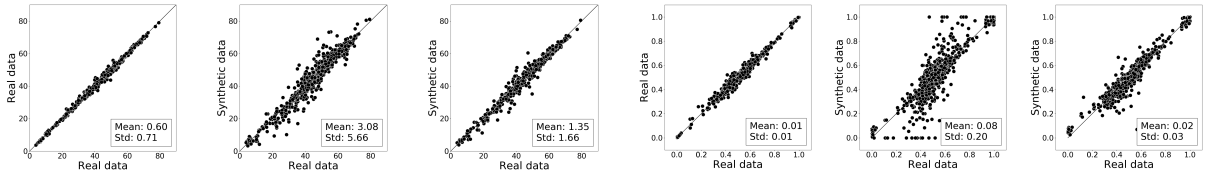
(a) Real, Precision (b) EMR-CWGAN, Precision (c) HGAN-U, Precision (d) HGAN, Precision

(e) Real, Recall (f) EMR-CWGAN, Recall (g) HGAN-U, Recall (h) HGAN, Recall

Figure 6: Frequent association rules. The cell numbers in subfigures 6(a)-6(d) and 6(e)-6(h) show the precision and recall of the association rules (for ICD and CPT codes), respectively. Figures 6(a) and 6(e) correspond to the real *vs* real setting.

Cross-type Conditional Distribution (CCD). The results of CCD with respect to the demographic distribution are shown in Figure 7, where each dot denotes a (ICD or CPT) code. The x-coordinate corresponds to real data, and the y-coordinate of subfigures 7(b)-7(c) and 7(e)-7(f) correspond to the synthetic data. In each subfigure, the mean and standard deviation of the dot-to-diagonal shortest distance distribution is marked at the bottom right corner. Figures 7(a) and 7(d) illustrate that the original systems (the baselines of the real *vs* real setting) are stable. When applying

HGAN-U, both the age and gender distributions are poorly represented, as shown in Figures 7(b) and 7(e). In particular, there are highly skewed gender ratios in multiple ICD and CPT codes. By comparing Figure 7(b) with 7(c) and Figure 7(e) with 7(f), it is evident that HGAN outperforms HGAN-U. This is because it achieves smaller distances from the diagonal. The CCD results exhibit highly similar patterns to the demographic results. Thus, our findings for CCD, with respect to three different data types, are consistent. Due to space limits, we report the results for the standard deviations of CCD in an extended version of this paper[24].



(a) Age, Real (b) Age, HGAN-U (c) Age, HGAN (d) Gender, Real (e) Gender, HGAN-U (f) Gender, HGAN

Figure 7: Cross-type conditional distribution on age and gender. Subfigures 7(a)–7(c) show the mean age for each code in the real *vs* real and real *vs* synthetic settings. Subfigures 7(d)–7(f) show the proportion of genders that are female for each code.

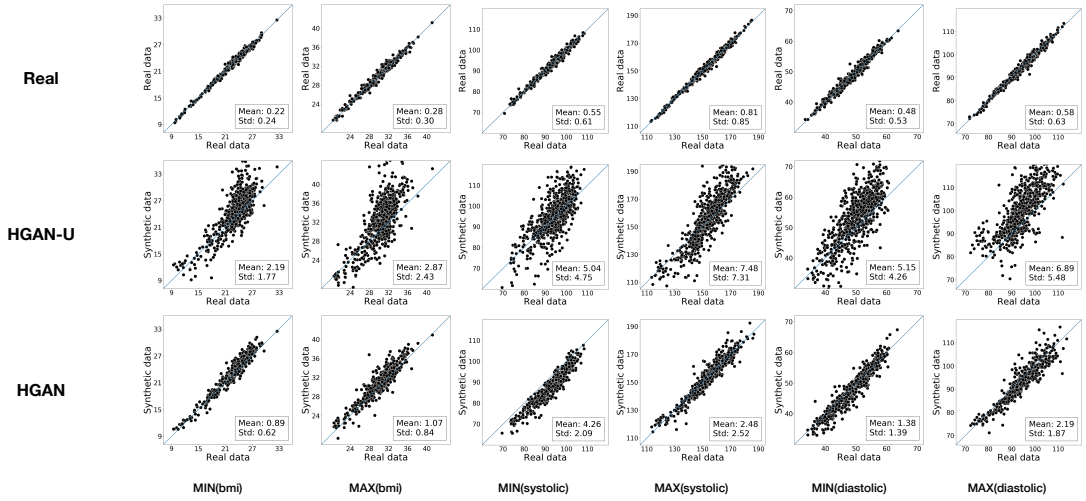


Figure 8: Cross-type conditional distribution on vital signs. The coordinates of each dot correspond to the means of the vital sign distributions on the real and synthetic dataset.

Privacy Risk Analysis

We evaluated the privacy risks of the synthetic data generated by HGAN and EMR-CWGAN. Since EMR-CWGAN was designed to generate billing codes, for comparison we performed two sets of investigations on HGAN–one on ICD and CPT only and the other on all data types. Though age and gender served as the conditionals for simulation in both models, we combined these features with the corresponding synthetic instances for analysis. We report the results of member-ship and attribute inference in the extended version[24]. There are two principle findings. First, considering only ICD and CPT codes, HGAN and EMR-CWGAN induce highly similar membership and attribute inference risks. This suggests that HGAN achieves better utility with no loss in privacy. Second, considering all data types, HGAN achieved 1) a similar precision, but a much lower recall of success than EMR-CWGAN in membership inference, and 2) a lower F1 score than EMR-CWGAN for attribute inference. This implies that a larger feature space enables HGAN to generate records that are very similar to real records and, thus, achieve a higher degree of privacy.

Discussions and Conclusions

This investigation yields several notable implications for EHR data simulation. First, the refinement of the neural network design improved the data utility by making the signal sparser. We verify this in the extended version of this

paper by investigating the number of activated neurons in the generators of HGAN and HGAN-U[24]. Second, it appears that a well-designed penalization can effectively prevent feature constraint violations. Importantly, based on the utility measures reported in this work, the incorporation of such penalization appears to not bias learning the distribution of real data. Third, this investigation illustrates the importance of the new measures (i.e., CVT, FAR, and CCD) for evaluating the utilities of synthetic data in the simulation tasks with feature constraints and heterogenous data types.

Despite the merits of this work, there are several limitations we wish to highlight. First, we did not assessed 1) the generalizability of our model?s performance on different category of features or in additional datasets or 2) the scalability of our model on larger and sparser feature spaces (e.g., medications and laboratory tests). Second, the model needs to refined to ensure that all meaningful association rules can be retained. Third, in this study we did not investigate how to automate the task of finding the meaningful constraints among features, which is ripe for further investigation. Fourth, we focused on static EHR data simulation, but it is necessary to incorporate temporal factors to simulate more complex phenotypes with trajectories that emerge over time.

Acknowledgements

This research was sponsored in part by NIH grants RM1HG009034 and U2COD023196.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395–405.
2. Casey JA, Schwartz BS, Stewart WF, et al. Using electronic health records for population health research: a review of methods and applications. Annu Rev of Public Health. 2016;37:61–81.
3. Filkins BL, Kim JY, Roberts B, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8(3):1560–80.
4. McGuire AL, Fisher R, Cusenza P, et al. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. Genet in Med. 2008;10(7):495–9.
5. Fung BC, Wang K, Chen R, et al. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys. 2010;42(4):1–53.
6. Dwork C, Pottenger R. Toward practicing privacy. J Am Med Inform Assoc. 2013;20(1):102–8.
7. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proc Advances in Neural Information Processing Systems; 2014. p. 2672–80.
8. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press: Cambridge, MA; 2016.
9. Zhang Z, Yan C, Mesa DA, et al. Ensuring electronic medical record simulation through better training, modeling, and evaluation. J Am Med Inform Assoc. 2020;27(1):99–108.
10. Choi E, Biswal S, Malin B, et al. Generating multi-label discrete patient records using generative adversarial networks. In: Proc Machine Learning Research. vol. 68;. p. 286–305.
11. Baowaly MK, Lin CC, Liu CL, et al. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc. 2019;26(3):228–41.
12. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proc International Conference on Machine Learning; 2017. p. 214–23.
13. Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs. In: Proc Advances in Neural Information Processing Systems; 2017. p. 5767–77.
14. Chin-Cheong K, Sutter T, Vogt JE. Generation of heterogeneous synthetic electronic health records using GANs. Workshop on Machine Learning for Health (ML4H) at Conference on Neural Information Processing Systems; 2019.
15. CPT Code Hierarchy; 2017. Available from: `http://medpricemonkey.com/cpt_hierarchy_list`.
16. Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style. arXiv preprint arXiv:1610.07629; 2016.
17. De Vries H, Strub F, Mary J, et al. Modulating early visual processing by language. In: Proc Advances in Neural Information Processing Systems; 2017. p. 6594–6604.
18. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc International Conference on Machine Learning; 2015. p. 448–56.
19. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proc International Conference on Artificial Intelligence and Statistics; 2011. p. 315–23.
20. El Emam K. Seven Ways to Evaluate the Utility of Synthetic Data. IEEE Security & Privacy. 2020;18(4):56–9.
21. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier: New York, NY; 2011.
22. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980; 2014.
23. Higgins I, Matthey L, Pal A, et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proc International Conference on Learning Representations; 2017.
24. Yan C, Zhang Z, Nyemba S, et al. Generating electronic health records with multiple data types and constraints. arXiv preprint arXiv:2003.07904; 2020.