Neurocomputing 416 (2020) 244–255

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Generation and evaluation of privacy preserving synthetic health data

Andrew Yale[a,*], Saloni Dash[c], Ritik Dutta[d], Isabelle Guyon[b], Adrien Pavao[b], Kristin P. Bennett[a]

[a] Rensselaer Polytechnic Institute, Troy, New York, USA
[b] UPSud/INRIA Université Paris-Saclay, France
[c] BITS Pilani, Department of CSIS, Goa Campus, India
[d] IIT Gandhinagar, India

ARTICLE INFO

Article history:
Received 16 July 2019
Revised 16 December 2019
Accepted 16 December 2019
Available online 10 April 2020

Keywords:
Synthetic data
Health data
Generative adversarial networks
Privacy

ABSTRACT

We develop metrics for measuring the quality of synthetic health data for both education and research. We use novel and existing metrics to capture a synthetic dataset's resemblance, privacy, utility and footprint. Using these metrics, we develop an end-to-end workflow based on our generative adversarial network (GAN) method, HealthGAN, that creates privacy preserving synthetic health data. Our workflow meets privacy specifications of our data partner: (1) the HealthGAN is trained inside a secure environment; (2) the HealthGAN model is used outside of the secure environment by external users to generate synthetic data. This second step facilitates data handling for external users by avoiding de-identification, which may require special user training, be costly, or cause loss of data fidelity. This workflow is compared against five other baseline methods. While maintaining resemblance and utility comparable to other methods, HealthGAN provides the best privacy and footprint. We present two case studies in which our methodology was put to work in the classroom and research settings. We evaluate utility in the classroom through a data analysis challenge given to students and in research by replicating three different medical papers with synthetic data. Data, code, and the challenge that we organized for educational purposes are available.

1. Introduction

Teaching data analysis and doing research with actual patient level medical data such as electronic healthcare records (EHR) are greatly restrained by laws protecting patients' privacy, such as the Health Insurance Portability and Accountability Act (HIPAA) [1,2] in the United States and the General Data Protection Regulation (GDPR) [3] in the European Union. While beneficial, these laws severely limit access to patient level medical data thus stagnating innovation and limiting educational and research opportunities. The process of obfuscation of medical data is costly and time consuming with high penalties for accidental release. Research and education using EHR are highly skewed to a few shareable datasets such as MIMIC-III (Medical Information Mart for Intensive Care) [4], which consists of de-identified ICU (intensive care unit) longitudinal data from 2001 to 2012 that adheres to the HIPAA restrictions and therefore can be shared. The only requirement is that the user completes a "Data or Specimens Only Research" certification. Datasets like MIMIC protect patients' privacy with classical anonymization techniques consisting of removing or regrouping quasi-identifiers in higher level categories (such as broad geographical areas) and removing or obfuscating sensitive information. Hence data utility can be severely altered. While the MIMIC data is extremely useful and has generated many research papers, it is limited to ICU data. It does not give access to the entire medical history of patients, hence limiting the type of analyses that can be carried out. This paper addresses this problem by proposing to use generators of synthetic data. The balance that needs to be hit in this project is to create synthetic data with enough quality to be useful for teaching purposes and ideally even for research, while preserving the privacy of the real data. In order to be useful in an education setting, synthetic data must preserve the relationships that exist in real patient-level data, so that assignments and projects using or discovering these relationships can be taught to students with the privacy preserving synthetic data. Other synthetic data generators like Synthea [5] pursue a similar goal but are based on publicly available summary statistic data, and therefore do not provide the flexibility of creating generative models faithfully resembling real data.

* Corresponding author.
E-mail address: yalea@rpi.edu (A. Yale).

https://doi.org/10.1016/j.neucom.2019.12.136
0925-2312/

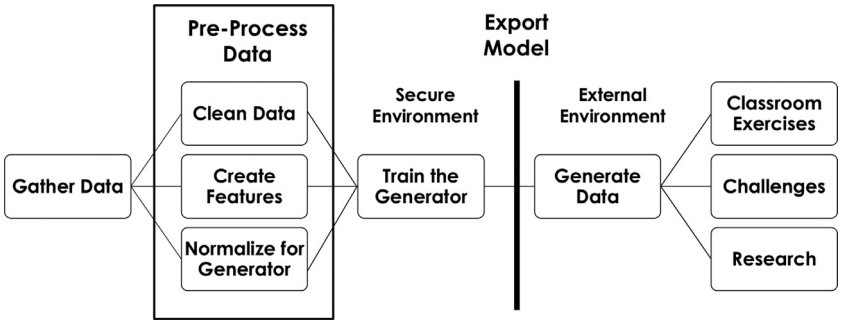


Fig. 1. Workflow used to generate synthetic data securely. The data is gathered, processed, and used to train the generator model inside the secure environment. Then the model, which does not contain any real data, or require real data to generate synthetic data, is exported outside the secure environment. Finally, data is generated using the model and used for multiple types of applications.

Our proposed workflow (Fig. 1) consists of training a generative model of synthetic data, using real data in a secure sand-boxed environment, exporting the model to the outside, and then synthesizing data. This procedure complies with our healthcare partners' regulatory requirements. We use novel and existing metrics to capture (1) *resemblance*: data generated are sufficiently close to the real data and (2) *privacy*: data generated are significantly different from training samples. We also assess (3) *utility*: data generated preserves some utility (for research and education purposes) and (4) *footprint*: trained model may not contain or require real data to generate synthetic data and should not be on the order of the real data in size. We develop a novel Wasserstein GAN-based method called HealthGAN and conduct a benchmark study on MIMIC data comparing it to five other approaches using a battery of metrics of utility, resemblance, and privacy.

We emphasize that privacy and resemblance are conflicting goals. By overfitting the data, the generative models can memorize the data thus potentially generating the actual real data points [6–8]. Models such as Parzen Windows can potentially accurately capture the data, but may reveal actual data points in the generated data or the modeling code making their footprint unacceptable.

In this paper, we extend the European Symposium on Artificial Neural Networks (ESANN) 2019 conference paper "Privacy Preserving Synthetic Health Data" [9] to include a more complete description and evaluation of HealthGAN and our proposed approaches for evaluating the quality of synthetic data. Section 2 discusses classical privacy preservation methods that are traditionally used to protect datasets such as MIMIC. These methods work to maintain as much of the real data as possible while still being private. In Section 3, we present new metrics for measuring the resemblance and privacy of synthetic data using nearest neighbors methods. Section 4 explores a previous GAN method for generating synthetic medical data called medGAN [10] and discusses how the method fares on other datasets. Section 5 contains descriptions of the six different methods that will be evaluated including our proposed solution, HealthGAN. Evaluation of the methods using our new metrics is presented in Section 6. Further results in the form of case studies using the HealthGAN for both education and research applications are in Section 7. Finally, in Section 8, we provide conclusions and discuss future possibilities for this work.

2. Related work

Classical privacy preservation techniques are focused on creating a new version of the real data set that ensures that no record in the data can exclusively identify an individual ("Unique Identity Disclosure"). These methods must also ensure that an attacker having prior knowledge about an individual is not able to obtain sensitive information from the disclosed attributes ("Sensitive Attribute Disclosure"). This can be done using processes such as generalization, anatomization, and perturbation [11]. The effectiveness of these methods is proved theoretically and ensures a specific level of privacy preservation.

In general, privacy models can be categorized as follows[11]:

- *Generalization:* Replacement of a value for a more general one (parent). For instance, a zip code of 12345 can be replaced by 123** or a Profession of Actor can be replaced by Artist.
- *Suppression:* Removal of some attribute values to prevent information disclosure (identifiers).
- *Anatomization:* De-associates quasi-identifiers (QIDs) and sensitive attributes in two separate tables.
- *Perturbation:* Adding noise to real data to create new data that preserves summary statistical information.

Our approach deviates from these classical methods by focusing on methods that create new data points that approximately mimic the real data rather than altering real data points. We also quantify the loss in privacy rather than relying on theoretical guarantees. Since our generative methods attempt to create data that look exactly like the real data, they minimize the utility loss, as compared to techniques such as generalization which would remove details from the data. Utility loss measures how well the synthetic data performs the task for which the real data was used. For example if a classification task is being performed the loss measures how close performance metrics for the classification task is between the synthetic data and the real data. Empirical evaluation is important to our research because privacy and utility loss can be concretely measured on the synthetic data with respect to organizations' sensitive data before releasing it to the public. As a baseline, the ubiquitous differential privacy data obfuscation method [12,13], an instance of the perturbation model of privacy, will be used to compare how well these data altering methods perform against generative methods.

In addition to classical techniques, there are newer methods that use GANs to create synthetic medical data. MedGAN, developed a GAN on MIMIC data as discussed in "Generating multi-label discrete patient records using generative adversarial networks" [10]. The medGAN implementation was tested with diagnosis data derived from MIMIC discrete patient records using generative adversarial networks. Each row in the data represents a patient and contains binary columns or count columns for each diagnosis, although the binary data was the focus of most of the work. In the binary data, for every instance where a patient had ever been diagnosed with a condition in the MIMIC data, the diagnosis was set to one, and the rest are zero. The many real-valued features in MIMIC were not synthesized using medGAN.

The medGAN version of the GAN architecture was also slightly different from the original GAN implementation in several ways, in an attempt to fix problems and make the GAN work for discrete

data. In their version, they added an autoencoder to the middle of the network to help the generator create more realistic samples. This autoencoder was trained from the real data so that the decoder part could be used on the data created from the generator. They also used minibatch averaging which helped enforce the column mean of the real data on the synthetic data. Finally, in the initial GAN structure discrete values could not be generated. This implementation took the continuous values from zero to one created by the GAN and rounded them based on the cutoff of 0.5 to create binary data.

To measure the privacy preserved by the data generated by medGAN, they sampled 1% of the training set R and a subset s of known attributes from each record. Then using the synthetic data, they inferred the missing attributes using k-nearest neighbors. They tried this for multiple values of k, s, and the total number of synthetic samples. Throughout all of their experiments they consistently showed that the attacker would have poor sensitivity and precision in this method when inferring the missing attributes. In this case the sensitivity is the percentage of positive diagnoses inferred out of the total number of positive diagnoses in the compromised record. Precision is measure the percentage of inferred diagnoses that were in fact true in the compromised record. This notion of nearest neighbors being used in privacy metrics led us to the metrics we have created for testing privacy and resemblance in our synthetic data.

3. Metrics of resemblance and privacy

We introduce metrics of resemblance and privacy. Consider two data distributions P_T and P_S, where T and S designate a Target and a Source domain respectively, for instance True (real) and Synthetic data. We draw empirical samples $\mathcal{S}_T = \{(\mathbf{x}_T^1, y_T^1), \dots (\mathbf{x}_T^n, y_T^n)\}$ from P_T and $\mathcal{S}_S = \{(\mathbf{x}_S^1, y_S^1), \dots (\mathbf{x}_S^n, y_S^n)\}$ from P_S. We assume that in all cases $\mathbf{x}$ variables belong to a common metric space *e.g.,* IR^d and y is a categorical or continuous variable (*i.e.,* defining classification or regression tasks). We also assume that all variables have been normalized, e.g. by subtracting the minimum and dividing by the range of the data.

The proposed metrics are based on nearest neighbors. We call $d_{TS}(i) = \min_j \|\mathbf{x}_T^i - \mathbf{x}_S^j\|$ the distance (Euclidean or otherwise) between $\mathbf{x}_T^i \in \mathcal{S}_T$, a point in the sample from the Target distribution, and its nearest neighbor in $\mathcal{S}_S$, the sample from the Source distribution. Therefore $d_{ST}(i)$ is the opposite measuring the distance between $\mathbf{x}_S^i \in \mathcal{S}_S$, a point in the sample from the Source distribution, and its nearest neighbor in $\mathcal{S}_T$, the sample from the Target distribution. Therefore $d_{ST}(i)$ is the opposite me measuring We call $d_{TT}(i) = \min_{j, j\neq i} ||\mathbf{x}_T^i - \mathbf{x}_T^j||$ the "leave-one-out" distance to the nearest neighbor in a sample of size $(n-1)$ drawn from the same distribution. We define $\mathcal{AA}_{TS}$, the *nearest neighbor Adversarial Accuracy* between T and S as:

$$\mathcal{AA}_{TS} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(d_{TS}(i) > d_{TT}(i)) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(d_{ST}(i) > d_{SS}(i))\right) \tag{1}$$

where the indicator function $\mathbf{1}(.)$ takes value one if its argument is true and zero otherwise. If we think of T as the true data and S as the synthetic data, by this definition, a real point i in T, which is sufficiently far away from any point in S, is a "true positive" point with respect to privacy. Similarly, a simulated point j in S must be sufficiently far from any point in T in order to be a "true negative" point. We can think of $\mathcal{AA}_{TS}$ as the performance of an adversarial classifier that distinguishes between real versus synthetic data. The $\mathcal{AA}$ definition is a "balanced accuracy", which averages the true positive rate and the true negative rate. If datasets T and S are indistinguishable, then $\mathcal{AA}_{TS}$ should be 0.5.

We use various datasets, all of size n, to define resemblance and privacy: R_{tr} is the real data training set used to train the generator; R_{te} is the real data test set, drawn independently from the same distribution as R_{tr}; A_1 and A_2 are any two artificial datasets generated by the generator network, G. We denote by $E(.)$ the mean value over all A_i and define three kinds of losses:

$$\begin{aligned}
\textbf{TrResemblLoss}\ (\textit{Train Adversarial Acc.}) &= E[\mathcal{AA}_{RtrA_1}] \\
\textbf{TeResemblLoss}\ (\textit{Test Adversarial Acc.}) &= E[\mathcal{AA}_{RteA_2}] \\
\textbf{PrivacyLoss}\ (\textit{Test } \mathcal{AA} - \textit{Train } \mathcal{AA}) &= E[\mathcal{AA}_{RteA_2} - \mathcal{AA}_{RtrA_1}]
\end{aligned} \tag{2}$$

Intuitively, if the generator G does a good job, then the adversarial classifier cannot distinguish between generated data and real data, train and test adversarial accuracy should both be 0.5, and the privacy loss will be zero. If G does a poor job and underfits, it will serve generated data that does not resemble real data. Thus the adversarial classifier will have no problem classifying real *vs.* artificial so the train and test adversarial accuracy will *both* be high (>0.5) and similar, and the privacy loss will also be near zero. In this last case, privacy is preserved but the utility of the data may be low. If the generator overfits the training data, the Train $\mathcal{AA}$ will be near zero (good training resemblance), but the Test $\mathcal{AA}$ will be around 0.5 (poor test resemblance). Thus the privacy loss will be high (near 0.5). Fig. 2 provides a two-dimensional synthetic example of these three cases in which blue is the real data and red is the artificial data. We generate train and test data ($n = 50$ from a semicircle plus Gaussian noise then standardized). We generate two artificial datasets A_1 and A_2 of the same size with the Parzen Windows density estimator (this method approximates a density by a mixture of local continuous density functions centered at data points and having a certain bandwidth size), using a Gaussian kernel of varying bandwidth to create three models, from left to right: (1) underfitted, (2) properly fitted, and (3) overfitted. The Train and Test adversarial accuracy ($\mathcal{AA}$) is shown for each case. For the same example, Fig. 3 provides curves representing Train $\mathcal{AA}$, Test $\mathcal{AA}$, and Privacy Loss for decreasing Parzen Windows kernel widths.

4. HealthGAN formulation

Based on the improvements the medGAN makes on the initial GAN, and the fact that it actually uses medical data, we used this as a first attempt at a GAN method. Through evaluating the data produced from medGAN, we found that there are some major issues with the implementation. First is the fact that it is only built for binary data. Second, our assessment reveals it has some flaws in the resemblance of the generated data to the real data.

The problem with the resemblance of the data from medGAN is that the architecture has been optimized to preserve the probability of each diagnosis occurring. Stated another way, it is optimizing for matching each column's mean to the real data. This is shown in Fig. 4 where the real data column mean is on the X-axis and the synthetic data column mean is on the Y-axis. Even though most of the diagnoses are very rare the synthetic data matches the univariate probabilities closely.

While this metric seems to show that the datasets have a very close resemblance, looking at the other dimension shows the true story. Another way that the synthetic data should resemble the real data is in the row sum. The row sum can be interpreted as the total number of unique diagnoses that a patient has. Patients typically have a small number of diagnoses. Comparing the overall distribution of row sums reveals that medGAN overestimates the number of diagnoses for some patients. In Fig. 5, the row sums of medGAN and the real data are compared. The X-axis shows different values of row sums, or total unique diagnoses. The Y-axis

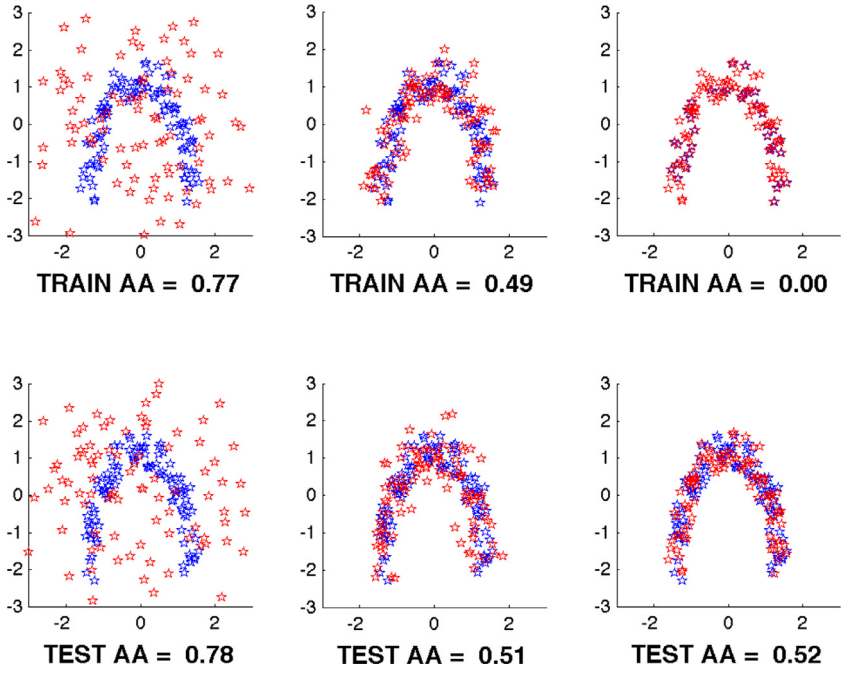


Fig. 2. *Parzen windows, toy example.* Blue markers represent real 2-d data samples (R_{tr} and R_{te}) and red markers represent synthetic data generated with Parzen windows (A_1 and A_2). *Top row:* R_{tr} and A_1. *Bottom row:* R_{te} and A_2. *Form left to right:* Large kernel ⇒ underfitting; optimized kernel ⇒ fitting right; and small kernel ⇒ overfitting.

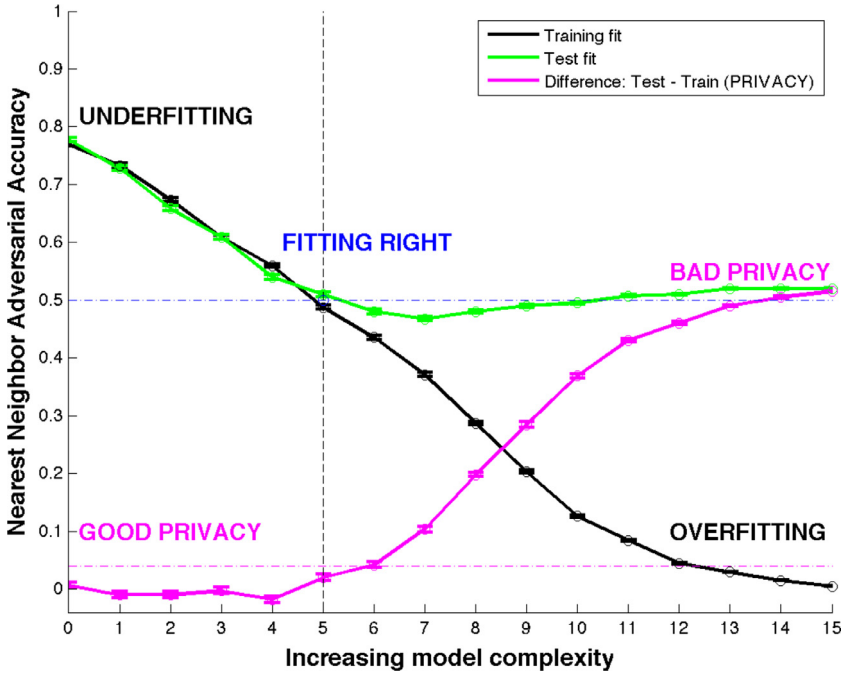


Fig. 3. *Parzen windows, learning curves.* The Train $\mathcal{AA}$ keeps decreasing, but not the test $\mathcal{AA}$. The privacy is good when the difference Test $\mathcal{AA}$ – Train $\mathcal{AA}$ is small. The best compromise is attained around the point where the black curve crosses the dashed blue line at 0.5. The pink dashed line shows the difference (Test $\mathcal{AA}$ – Train $\mathcal{AA}$).

shows the counts of the occurrences of these bins on a log scale in order to show the tail of the synthetic data. This plot can be interpreted to say that the generator is creating synthetic patients with four times as many diagnoses as the patient with the highest number of diagnoses in the real data. While the number of patients generated in that tail might not be high, this exposes another issue which is that these synthetic patients with over 100 diagnoses become a catch-all for the rare diagnoses. They become patients that are only placed there to make the column means balance. This issue was indirectly referenced in the paper when they showed their synthetic patients and real patients to a medical doctor and ask the medical doctor to distinguish between the two. In some of the instances where the doctor was able to distinguish between the two, the reason was the patient was marked down for a female specific and male specific diagnosis at the same time. MedGAN introduces spurious comorbidities that are not in the real dataset, thus limiting the resemblance and utility of the synthetic data.

Even though medGAN was created to solve a similar problem to ours and included some good results, it was ultimately ruled out as

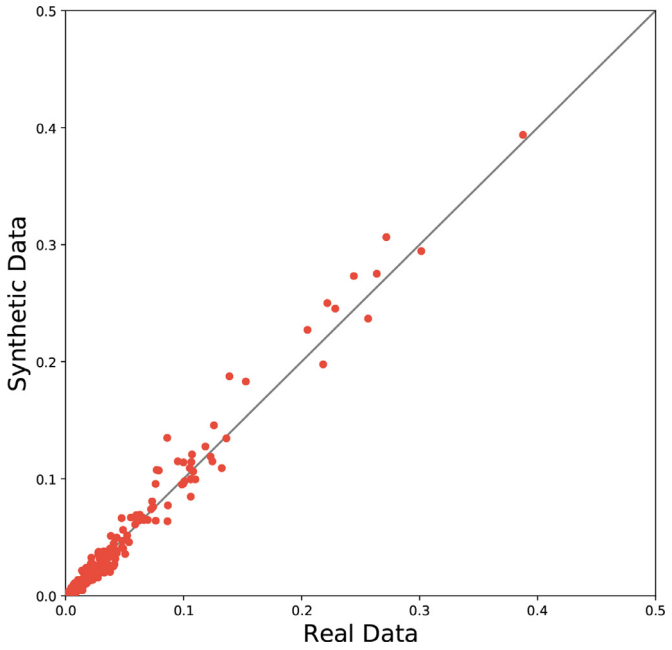


Fig. 4. medGAN dimension-wise probability (column mean) comparison. Each point represents a column mean in the real and synthetic data. A perfect match would be indicated by all the points lying on the line $y = x$.

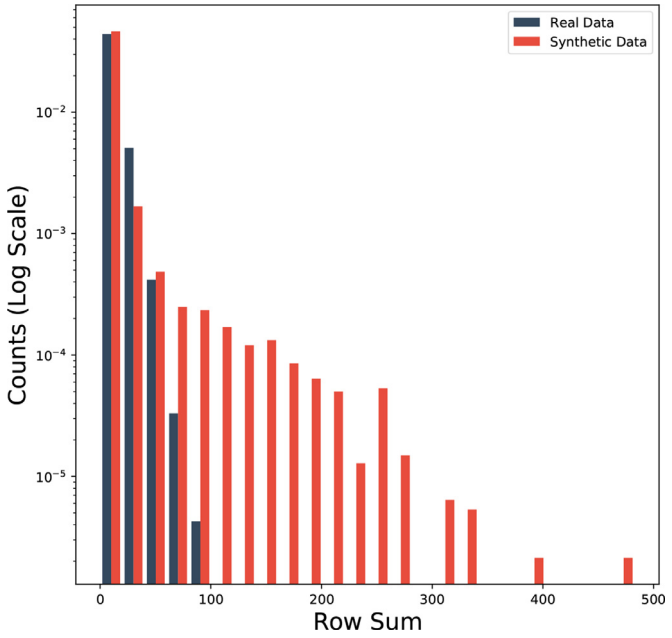


Fig. 5. medGAN row sum distribution comparison. A perfect match would be indicated by matching histograms.

a basis for the generative method because of problems with resemblance. First, the fact that it could only generate binary data was not flexible enough. Second, the issues with resemblance in the row sums indicated that unrealistic data was being generated. Finally, the model had been so specifically crafted for the binary diagnosis dataset, that it was not robust enough to work with many different datasets of different sizes and columns, which is the final goal. This led us to creating a new method we are calling HealthGAN.

The HealthGAN architecture and design took ideas from medGAN and combined them with the Wasserstein GAN gradient penalty (WGAN-GP) [14,15] to accommodate multiple data types. The WGAN-GP was used as a basis for the HealthGAN because it uses the earth mover's distance or Wasserstein distance versus the Kullback–Leibler (KL) divergence used in original GAN [16] or the medGAN [10]. That paired with a data transformation to accommodate mixed continuous and categorical data, such as MIMIC-III. EHR are always mix of continuous features (*e.g.,* age, bmi, lab results) and categorical features (*e.g.,* diagnosis codes and providers) and therefore this new GAN was built to accommodate multiple data types from the start. Furthermore, a patient has many EHR records through time. While this version of the HealthGAN does not accommodate time series data it is an improvement being attempted on other datasets [17].

The architecture is similar to the original GAN with a generator network with three layers, and a discriminator network with four layers. Due to the large variance present in medical datasets a large batch size is used to ensure that outliers and rare values are captured in each batch and therefore learned by the generator. The batch size is determined based on the size of the input data to maximize the size. The dimensions of the generator network is also determined by the number of features in the dataset to scale in complexity according to the data it is modeling. HealthGAN represents an attractive black box method with a very compact footprint (parameters of the model) since the bottleneck in HealthGAN is constructed to prevent memorization.

5. Data generation methods

We performed a comparison of 6 data generative methods[1] on the MIMIC-III mortality problem: (1) *Gaussian Multivariate* [18], (2) *HealthGAN*, (3) Parzen Windows [19], (4) Additive Noise Model (ANM) [20], (5) Differential Privacy preserving data obfuscation (DP) [12], and (6) Copy the real data (CP).[2] Our usage of HealthGAN and ANM in this context are novel, to the best of our knowledge.

5.1. Data generation methods

We describe the 6 data generation methods in more detail.

- *Gaussian multivariate:* This method simply consists of modeling data by a multivariate Gaussian distribution whose parameters are then found using Maximum Likelihood Estimation (MLE), *i.e.,* using the mean and covariance matrix of the training data. This method fulfills our footprint specifications because the model is much smaller in size than the real data and does not directly represent any sample (provided that the means are not actual data points).
- *HealthGAN:* Described in the preceding section, a GAN based method for generating mixed continuous and categorical data.
- *Additive noise model:* Inspired by methods used for imputation of missing data, a suitable predictor (here we use Random Forests) is trained to predict one feature of a given sample, given all the other features. Predicting each feature for each sample in this way gives a dataset A_0 consisting entirely of predicted values, which can then be sampled from to generate synthetic datasets. Noise is drawn from a Gaussian distribution with zero mean and variance equal to the mean-square-error of the fit and is added to each predicted value to increase the diversity of the data produced. The model itself has a small footprint, but data generation requires storing A_0 and therefore exporting data, which rules out this model for our application purposes. We keep it as a baseline method.

[1] https://github.com/yknot/ESANN2019

[2] medGAN was not included as a compared method as it can only generate discrete data.

- *Parzen windows:* Parzen Windows density estimation approximates a density by a mixture of local continuous density functions K, called kernels, centered at data points and with bandwidth equal to h: $\hat{f}_h(x) = \frac{1}{Z}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$ with x_1, ..., x_i the data points and Z a proper scaling factor. Generating data boils down to picking a data sample at random, then drawing a sample at random around the sample by applying the kernel density function. This method has an unacceptable footprint since each data point is represented in the Parzen Windows function.
- *Copy real data:* We exactly duplicate the data; more precisely we use the train set instead of synthetic data. Resemblance is high but the model maximally overfits, thus privacy is at a minimum. The footprint duplicates the data and thus is of course unacceptable.
- *Privacy-preserving data obfuscation:* Differential privacy is a widely accepted privacy requirement for data publishing [12]. We generated a ϵ, δ differentially private version of the MIMIC-III dataset by creating generalization hierarchies for the seven quasi-identifier attributes[3] using ARX, an open source anonymization tool for medical data [21] based on the SafePub Algorithm [22]. The footprint of this method is unacceptable because it requires export of most of the real data and privacy is limited to quasi-identifiable fields.

5.2. Data transformation

Data transformation was essential for the success of many of the methods including HealthGAN. Recall MIMIC-III contains a mix of categorical and discrete variables. We adapted data transformation strategies used in the Synthetic Data Vault (SDV) [23]. We map all features to a range between zero and one, synthesize the data, and finally transform the synthetic data back to its original form, using the mapping from the real data. Numeric variables are scaled by subtracting the min and dividing by (max-min). For each categorical variable, we first sort from most frequent to least frequent. Then we split the interval from zero to one into sections based on the cumulative probability for each category. Finally, lining up each category with its section on the interval from zero to one, we take a sample from that section using a truncated Gaussian distribution. The reverse transformation maps the synthetic data to the original categories. This transformation is part of all of the methods discussed in the previous section that cannot accommodate categories by default.

6. Experimental results

We evaluated the synthetic generation data on the MIMIC-III dataset which contains records for about 40,000 intensive care unit (ICU) patients and indicates whether they died in the ICU. It includes demographics, vital signs, diagnoses, and procedures performed. The dataset we used is tabular data related to 48-h mortality in the ICU. The data has 342 features, the output column of mortality and approximately 27,000 observations. We generated synthetic data and then evaluated the different approaches using visualization techniques and the proposed metrics.

6.1. Principal component analysis plots

We found principal component analysis (PCA) plots created using projection of the real train data to be very useful for getting a quick understanding of resemblance of the real test data (black dots) to the generated synthetic data (red dots). Here we can see that data generated by the Gaussian Multivariate and Parzen Windows methods span a larger space than the real data, which aligns with the fact that those methods create differences in the data in both directions uniformly. The PCA of the Differential Privacy data obfuscation method data spans a smaller space, which represents fact that the quasi-identifiers are changed enough to not reveal outlier data (Fig. 6).

Both the real and synthetic data distributions of HealthGAN and the ANM have high resemblance, which aligns with their greater ability to define relationships that exist in the real data and apply that to their generated synthetic data.

Other dimensionality reduction methods were also explored, but the best tool for visual assessing the synthetic data was determined to be the PCA. For example T-distributed Stochastic Neighbor Embedding (t-SNE) produced plots that were more uniform looking across the methods and therefore did not show obvious differences between the methods.

6.2. Adversarial accuracy results

We compared the adversarial accuracy (Eq. (1)) in terms of TrainResemblanceLoss, TestResemblanceLoss, and PrivacyLoss = TestResemblanceLoss - TrainResemblanceLoss (Eq. (2)). As shown in Table 1, Gaussian Multivariate preserves privacy, but suffers from high testing adversarial accuracy (0.55). A well fitted Parzen Window (optimized kernel width) and HealthGAN both perform well with respect to resemblance and privacy. But the footprint of Parzen Windows rules it out for this purpose since the real data. While the DP method obscures the quasi-identifiers, it leaves open the rest of the data and therefore scores very poorly on the training data. For the ANM, we used few and deep trees to illustrate a case of overfitting: indeed the ANM overfits the data badly, completely exposing the real data. It is possible to tune the ANM hyperparameters to prevent overfitting, however, its footprint would still make it unacceptable for our applications. In both methods, the privacy of the data is at its worst. In Table 1 the colors for adversarial accuracy indicate how far the value is from optimal. In the case of adversarial accuracy the optimal value is 0.50 and $0.50 \pm .01$ is in the blue or excellent range. Yellow indicates a good value and specifically is a value of $0.50 \pm .03$. Anything outside that range is orange or poor. For privacy loss the optimal value is zero and anything less than or equal to.01 is excellent, less than or equal to 0.03 is good, and above that is poor.

We also assessed utility of the data generated by the methods by using the synthetic data to train a classifier to predict patient mortality, then testing the classifier on the real test dataset. A logistic regression classifier was selected and by comparing the area under the curve on the test data, we can see that the DP and CP methods have the best performance, but also have unacceptable privacy scores. The next best methods are Parzen Windows, ANM, and HealthGAN. The Additive Noise Model predictably performs poorly on privacy because it is overfitting, but cannot be used as a final method in any form due to the model footprint requiring real data. The Parzen Windows and HealthGAN perform well for both privacy and area under the curve, but the Parzen Windows method does not fulfill the model privacy requirements. Finally, Gaussian Multivariate, performs the worst on utility, but still has good data and model privacy. The utility metric is important to consider, because it roughly captures usefulness of the synthetic data in the classroom setting. The colors in the table correspond to excellent being a value from 0.80 to 1.00, good from 0.65 to 0.80, and poor for any value below that.

As discussed throughout the results the footprint is a major factor in selecting the final model as well. The footprint measures whether the information needed to generate synthetic data, specifically the model and any inputs, contains real data or is on the order of the size of the real data. This is critical to being able to export the model from the secure environment and keep the real

[3] 'Insurance', 'Language', 'Religion', 'Marital-Status', 'Ethnicity', 'Gender' and 'Age'.

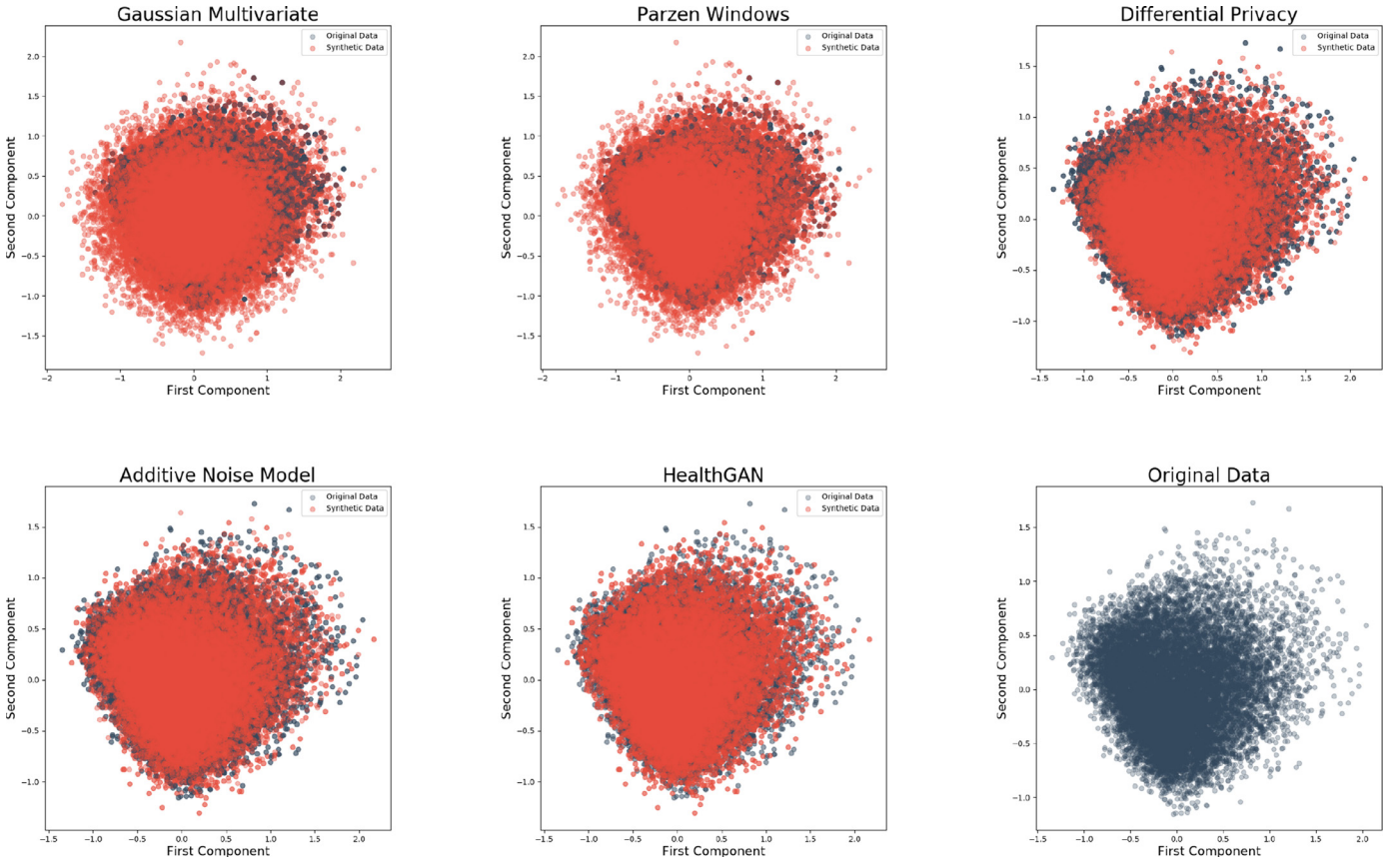


Fig. 6. Comparison of generative methods using PCA projection created using the real data. Blue is the real data and red is the synthetic data.

Table 1
Comparison of models with respect to various metrics. Blue: Excellent; Yellow: Good; Orange: Poor. Our advocated method marked with (*) performs best. Train $\mathcal{AA}$ and Test $\mathcal{AA}$ measure resemblance loss. $PrivacyLoss = Test\mathcal{AA} - Train\mathcal{AA}$. Utility measures test accuracy of predicting mortality. Footprint indicates whether we can export a small footprint model out of the secure area.

		Methods					
		HealthGAN (*)	**Gaussian Multivariate**	Fitted Parzen Win.	Overfitted ANM	Differential Privacy (DP)	Copy (CP)
Adversarial Accuracy	Train $\mathcal{AA}$	0.50	0.53	0.50	0.00	0.05	0.00
	Test $\mathcal{AA}$	0.51	0.55	0.50	0.50	0.52	0.50
	Privacy Loss	0.00	0.02	0.00	0.50	0.47	0.50
Utility	Area Under the Curve	0.66	0.62	0.77	0.74	0.87	0.88
Footprint	Up-to-specs	yes	yes	no	no	no	no

data secure. The Gaussian Multivariate and HealthGAN are the only methods that satisfy this condition.

7. Education and research case studies

After ensuring that the synthetic data generated by HealthGAN is satisfying the privacy and resemblance metrics in the previous section, the next goal was to ensure the synthetic data was of high enough utility to be used in the education and research settings. To do this we conducted both education and research case studies.

The education studies involve setting up a challenge for students with synthetic data in which they must create a classifier and are scored based on the performance of the classifier. This synthetic data challenge was designed for use in undergraduate health informatics curricula at Rensselaer Polytechinic Institute developed with support from the United Health Foundation. This curricula is designed to rapidly recruit and prepare undergraduate students to be data scientists in healthcare using early data analytics courses and experiential research projects centered on real-world health challenges.

In the research case study, we attempt to replicate three different published medical papers that use MIMIC data with synthetic versions of the datasets generated using HealthGAN. The data replicated from the research papers can also be used for educational purposes as students can either replicate the papers themselves with the synthetic data or be given the synthetic data and attempt to find their own method of accomplishing the goals set forth in the papers. All of these results show how well the synthetic data performs at modeling specific relationships in datasets.

7.1. Education case study

To assess educational utility, we used synthetic data in classroom challenges. “To be, or not to be?” is a mortality prediction

challenge hosted on CodaLab.[4] used in two courses The mortality prediction challenge has been used in two undergraduate courses at Rensselaer: "Introduction to Data Mathematics and Health Analytics Challenge Lab" at Rensselaer Polytechnic Institute. This challenge and other subsequent ones support the curriculum goal of exposing students in early data science courses to compelling healthcare problems in order to attract them to careers in health informatics.

We give students 80,000 synthetic records generated from the MIMIC mortality dataset used for testing in the previous section. This data includes demographic data, vital signs, diagnoses, and mortality. Using this dataset the students create models in R [24] on the training data and then predict mortality for 20,000 synthetic test records. Their predictions are uploaded to the server and evaluated. They never get to see the mortality values for the test dataset. The models will then be evaluated with the balanced accuracy metric and ranked. A major advantage of using the synthetic data challenge was that the students did not have to undergo the training required to use the real MIMIC data.

The best students achieved values of 0.77 and 0.76 balanced accuracy. When the model is trained on real data and evaluated on the same test set the result is 0.80, which shows how high the level of utility is on the generated data.

This case study demonstrates the utility of synthetic data for health informatics education. Students get the experience of working with patient level data while preserving privacy and avoiding additional precautions needed to work with actual patient data. Potentially, synthetic generation could make data from many published papers and electronic health care records available to enhance many types of educational programs including medical and public health programs.

7.2. Research case study

To assess utility of synthetic data in a research setting, we examined the effectiveness of synthetic data on three of the numerous medical studies published using the MIMIC data. We attempted to replicate the analysis in the three papers using synthetic versions of the MIMIC data. The results were also replicated with real MIMIC data to enable a comparison of the results using real and synthetic data. The papers were selected to utilize a wide variety of analysis methods.

7.2.1. Impact of race on ICU mortality study

The first paper used MIMIC data to try to evaluate the impact of race on 30-day mortality [25]. The authors took different demographic variables and comorbidities and found the odds ratios for each of them to find out whether race specifically impacted the 30-day mortality of the patient.

The original paper utilized the MIMIC-II dataset, [26] while we used MIMIC-III. The main difference between these datasets is that MIMIC-III includes years 2001 to 2012 in the data whereas MIMIC-II is only 2001 to 2007. In addition to those extra years of data, MIMIC-III also enriches the dataset with more variables to be looked at and fixing up some of the older values. Therefore the MIMIC-III dataset is just a better version of the MIMIC-II dataset and a better basis for doing this analysis and creating synthetic data. We also did cleaning with the race (called ethnicity in MIMIC), insurance, and age variables to give them better categories. For race there is a long tail of races which are put into "Other" and another category "Unknown" for missing values. Insurance had several categories that are separated into "Private", "Medicaid", and "Medicare". Age is put into buckets to more easily determine the effect of an age range. Finally, 30-day mortality is not tracked that accurately in the MIMIC data, so we will use mortality in the ICU, which is another example in the paper.

The authors of the paper used Kruskal–Wallis, Wilcoxin two-sample, Pearson's chi-squared, and Fisher's exact tests [26] to account for collinearity between race and covariates. Two multivariate logistic regression models are trained in order to account for confounders. The first regression does not include the "Unknown" race group and the second does. The second model is the focus of the results below.

The model creates odds ratios for each of the variables. In Table 2 we can see a comparison of the odds ratios computed from the real data versus the synthetic data. In this case we can see that 15 out of the 17 values have overlapping confidence intervals, thus indicating the results for the synthetic and real data were not significantly different from each other on the 15 variables.

These results are promising because they seem to replicate almost all of the relationships that we see in the real data. This satisfies another targeted utility example and shows that students can work with and learn the relationships between different features specifically found in health data. For research purposes it is harder to determine if quality of the synthetic data was high enough. This is because the results from the paper are not able to be exactly replicated in the real data and therefore also do not appear in the synthetic data. In the paper they found that the "Black" and "Asian" categories for race had odds ratios significantly less than one, meaning they did not overlap with one. Using the newer MIMIC-III dataset, this result does not appear in the data. This could be caused by many things including the specific definition of the features, model, or the fact that there are more records in MIMIC-III. The result is that our model shows inconclusive results about the effects of race on mortality in the real and synthetic data.

7.2.2. Mortality of elderly patients in the ICU study

The next paper attempted to analyze the characteristics and mortality of elderly patients in the ICU [27]. This study seeks to evaluate the association between the demographic and clinical characteristics of patients over the age of 65 and their 28-day and one-year mortality. Logistic regression is used to analyze 28-day mortality and a Cox regression model is used to analyze one-year mortality.

The data used in this paper are age, gender, sequential organ failure assessment (SOFA) [28], do not resuscitate (DNR), and the Elixhauser Comorbidity Index (ELIX) [29]. They only include patients over 65 and categorize the patients into age 65–74, 75–84, and 84+. In the MIMIC-III data we found 15,771 patients in the first category, 5664 in the second, and 3517 in the third.

Using the same significant variables as the paper we created a logistic regression model with both the real and synthetic data to see how they compared. The odds ratios confidence intervals for those models are in Table 3. The results between the synthetic and real data are similar but in some cases do not overlap. In the case of SOFA and DNR, the odds ratios do not overlap but have similar effects in that both show an odds ratio greater than one. The other variable that does not overlap is the 75–84 age range which in the synthetic data does not seem to have an effect as the confidence interval straddles one, but in the real data the value is decidedly greater than one. The discrepancies between the real and synthetic data highlight the challenges of synthesizing imbalanced data. The results suggests that oversampling rare classes during training of HealthGAN may lead to improved results.

These results mostly match the results in the paper. In the paper all of the selected factors besides being male have a greater than one odds ratio and are significant as we found for the real data.

[4] https://competitions.codalab.org/competitions/19365

Table 2
Comparing odds ratios for each variable in real data and synthetic data. Blue indicates overlapping intervals, red indicates non-overlapping intervals.

Variable		Real Data CI	Synthetic Data CI
Admission Loc.	ER	(1.00, 1.64)	(0.90, 1.56)
	Transfer	(1.07, 1.85)	(1.27, 2.51)
Insurance	Medicaid	(0.77, 1.42)	(0.65, 1.54)
	Medicare	(0.71, 1.21)	(1.00, 1.78)
Gender	Female	(0.73, 1.06)	(0.73, 1.12)
Ethnicity	Asian	(0.53, 1.58)	(0.60, 1.96)
	Black	(0.51, 1.08)	(0.80, 1.81)
	Other	(0.89, 1.90)	(0.58, 1.45)
	Unknown	(1.84, 3.13)	(0.60, 1.48)
Age	46-65	(0.79, 1.46)	(0.83, 1.72)
	66-80	(0.91, 1.93)	(0.70, 1.56)
	81+	(1.01, 2.33)	(0.78, 1.73)
First Careunit	SICU	(0.61, 0.95)	(0.76, 1.26)
Resuscitation Pref.	DNR	(4.13, 7.03)	(3.87, 6.51)
Disease	CHF	(0.44, 0.70)	(1.05, 1.80)
	Any Malignancy	(0.77, 1.44)	(0.92, 1.78)
	Both	(0.27, 0.88)	(0.33, 2.09)

Table 3
Logistic regression results. Blue indicates overlapping intervals, red indicates non-overlapping intervals.

Variable	Real Data CI	Synthetic Data CI
Age		
75-84	(1.25, 1.59)	(.84, 1.06)
85+	(1.29, 1.70)	(1.06, 1.38)
Gender, Male	(0.82, 1.00)	(0.82, 1.00)
SOFA, per point	(1.24, 1.27)	(1.15, 1.20)
DNR	(4.94, 6.18)	(1.23, 3.97)
Elixhauser, per point	(1.01, 1.02)	(1.02, 1.03)

Table 4
Cox regression results. Blue indicates overlapping intervals, red indicates non-overlapping intervals. INF indicates there was not enough synthetic data generated in that range and therefore the interval could not be computed.

Variable	Real Data CI	Synthetic Data CI
Age		
75-84	(0.64, 1.28)	(0.23, 4.69)
85+	(0.41, 1.05)	INF
Gender, Male	(0.79, 1.50)	(0.25, 5.09)
SOFA, per point	(1.21, 1.32)	(0.67, 1.24)
DNR	(1.36, 3.12)	INF
Elixhauser, per point	(1.03, 1.06)	(0.97, 1.09)

For the one-year mortality prediction a Cox regression model was used. The same variables were used as in the logistic regression model. In Table 4 we can see the results of this model.

Using the real data we were able to replicate most of the results of the original paper; the synthetic data was not able to at all. This again falls into an imbalanced class issue. The chance of a 28-day mortality was about 14.7% in the real data and was replicated in the synthetic data at 12.4%. On the other hand, one-year mortality occurred 1.1% of the time and therefore was much harder to replicate in the synthetic data. The fact that the one-year percentage is so much lower than the 28-day value does not make sense, and stems from the fact that the ability for the ICU to track patients for a year after they leave the ICU is poor.

7.2.3. Mortality in acute kidney injury

In this final paper, the authors predict mortality of patients with Acute Kidney Injury (AKI) [30]. To ensure the patients have an AKI diagnosis we look at patients with a longer than 72 h stay in the ICU. To predict mortality we use the Simplified Acute Physiology Score (SAPS) [31]. The SAPS score looks at systolic blood pressure, heart rate, temperature, urine output, blood urea nitrogen, white blood count, serum potassium, serum sodium, serum

Table 5
Coefficient values and significance per variable per day.

Scores	Day 1		Day 2		Day 3	
	Estimate	*p* value	Estimate	*p* value	Estimate	*p* value
WBC	0.2864	0.0002	0.3702	0.0000	0.6266	0.0000
ACI	−0.0579	0.6915	1.0185	0.0001	1.3380	0.0000
Urine	0.9031	0.0000	0.8968	0.0000	0.8513	0.0000
Sys BP	0.0518	0.0006	0.0563	0.0007	0.0629	0.0002
HR	0.0700	0.0108	0.0647	0.0355	0.0418	0.1744
Temp	−3.2122	0.9610	−3.0327	0.9632	3.5671	0.9567
BUN	0.0262	0.1430	0.0279	0.1298	0.0437	0.0198
Potassium	0.0915	0.0047	0.0770	0.0370	0.0908	0.0169
Sodium	0.1835	0.0052	0.2705	0.0006	0.1946	0.0154
Bilirubin	0.1485	0.0000	0.1650	0.0000	0.1727	0.0000
Age	0.0482	0.0000	0.0502	0.0000	0.0506	0.0000

Table 6
Area under the curve results.

	Real data	Synthetic data
Day 1 SAPS variables	0.6817	0.6071
Day 2 SAPS variables	0.7127	0.7133
Day 3 SAPS variables	0.7308	0.692
Day 1 + Day 2 + Day 3 (all variables)	0.7351	0.7301
Day 1 + Day 2 + Day 3 (forward selection)	0.7329	0.7279

bicarbonate, serum bilirubin, and age. Just like SOFA or comorbidity indexes each one of these variables results in points based on its value. By selecting patients who are in the ICU for at least three days we can get a score for each one of these categories for each day. Therefore we have 33 variables for eleven categories across three days.

With the variables selected we then ran a logistic regression to predict mortality for these patients using the SAPS variables. In Table 5 we can see the resulting coefficient and significance level of each variable using the real data.

Based on these results we can specifically select the significant variables and look at those. These variables are urine, age, and heart rate from day one, ACI and sodium from day two, and WBC, ACI, BUN, bilirubin, and sodium from day three.

Using these variables, we then construct five different logistic regression models. The first three use just day one, two, and three variables respectively. With the real data, as the days go on we have a better chance of predicting mortality, based on the area under the curve (AUC) measure. The fourth model uses all of the variables, and achieves a slightly better value AUC than just the day three variables. Finally, the fifth model just uses the significant variables from the previous figure to predict mortality. This model performs roughly equivalently to the all-variable model. These results are seen in Table 6.

These same five models were constructed for the synthetic data. The results were almost identical except for the day one and day three AUC values being slightly worse on the synthetic data than the real data. Given such a close result across five different models, we can say that the synthetic data achieves the desired level of targeted utility for this task.

In all three papers, the synthetic data sets exhibited high utility for education usage. In many cases, the conclusions found were not qualitatively different from those obtained from the real data. HealthGAN synthetic data can be published in lieu of the real data for cases where privacy does not permit publication of the real data. New algorithms and methods could be developed and compared on the public synthetic data. Ideally, the results of these methods should be verified on the real data in a secure environment that maintains privacy.

8. Conclusions and future work

Although GANs have increased in popularity, their effectiveness in the health domain was not clear. Through replicating the results from with the medGAN [10] architecture, the potentials and limitations of these methods became apparent. HealthGAN and the proposed evaluation metrics provide both an improved algorithm and better metrics for evaluating the quality of synthetic health data in the future. The workflow that we presented for generating synthetic data from real data and exporting a model only outside a data-secure environment has become operational with the introduction of HealthGAN. Generated data is competitive in resemblance with other methods, while meeting the requirements of privacy preservation and small model footprint. Our methodology includes novel metrics, based on nearest neighbor adversarial accuracy, for defining the resemblance and privacy of synthetic data generated from real data. We evaluated these metrics as well as utility and footprint on six methods using the MIMIC-III mortality data. HealthGAN was the only effective method that maintained privacy and that allowed model export. HealthGAN was then tested further in case studies in education and research. In an education case study, the synthetic data showed high levels of utility as students attempted a mortality prediction challenge. In research, three different papers using MIMIC data were replicated with synthetic data generated from HealthGAN. This workflow can be used to address the vital need to create datasets for health education and research without undergoing obfuscation, which can be both costly and risky and lose information. In addition, our proposed metrics will continue to be useful to monitor progress in synthetic data generation. All data, code, and the mortality prediction challenge that we organized are publicly available.[5]

Future work on this research is to use the HealthGAN method on medical datasets that extend beyond the ICU setting of MIMIC-III. This comparison will be done in the same style as the MIMIC-III paper replications, but with much more varied datasets. These synthetic datasets can then be used in curricula to teach students including creating challenges for them to solve health care problems on more diverse synthetic datasets. Beyond replicating papers, new datasets can be used to generate synthetic data and help create workflows for solving real research problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Andrew Yale: Conceptualization, Software, Methodology, Writing - original draft, Writing - review & editing. **Saloni Dash:** Software, Validation. **Ritik Dutta:** Software, Validation. **Isabelle Guyon:** Conceptualization, Methodology, Supervision. **Adrien Pavao:** Software, Validation. **Kristin P. Bennett:** Writing - review & editing, Conceptualization, Funding acquisition, Supervision.

References

[1] The Health Insurance Portability and Accountability Act of 1996, 110 Stat. §1936 (1996).
[2] G.J. Annas, et al., Hipaa regulations – a new era of medical-record privacy? N. Engl. J. Med. 348 (15) (2003) 1486–1490.
[3] Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive), L119, 4 May 2016, p. 1–88, (2016).

[5] https://github.com/yknot/ESANN2019

[4] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, Mimic-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.

[5] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, J. Am. Med. Inf. Assoc. 25 (3) (2018) 230–238, doi:10.1093/jamia/ocx079.

[6] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, in: Proceedings of the 31st IEEE Computer Security Foundations Symposium, CSF 2018, 2018, pp. 268–282, doi:10.1109/CSF.2018.00027. 1709.01604. Oxford, United Kingdom

[7] A.A.H. Khatri, Preventing Overfitting in Deep Learning Using Differential Privacy, State University of New York at Buffalo, 2017 Ph.D. thesis.

[8] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, The reusable holdout: Preserving validity in adaptive data analysis, Science 349 (6248) (2015) 636–638.

[9] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, K.P. Bennett, Privacy preserving synthetic health data, in: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2019, 2019. Bruges, Belgium

[10] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Machine Learning for Healthcare Conference, 2017, pp. 286–305.

[11] R. Mendes, J.P. Vilela, Privacy-preserving data mining: methods, metrics, and applications, IEEE Access 5 (2017) 10562–10582.

[12] C. Dwork, Differential privacy, Autom. Lang. Program. 4052 (2006) 1–12.

[13] C. Dwork, Differential privacy: A survey of results, in: Proceedings of the International Conference on Theory and Applications of Models of Computation, Springer, 2008, pp. 1–19.

[14] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2017.

[15] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[17] S. Dash, R. Dutta, I. Guyon, A. Pavao, A. Yale, K. P. Bennett, Synthetic event time series health data generation, ML4H, Machine Learning for Health (2019).

[18] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.

[19] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076, doi:10.1214/aoms/1177704472.

[20] P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models, in: Proceedings of the Advances in Neural Information Processing Systems, 2009, pp. 689–696.

[21] F. Prasser, J. Eicher, R. Bild, H. Spengler, K.A. Kuhn, A tool for optimizing de-identified health data for use in statistical classification, in: Proceedings of the 30th IEEE International Symposium on Computer-Based Medical Systems, 2017.

[22] K.A.K. Raffael Bildraffael, F. Prasser, Safepub: A truthful data anonymization algorithm with strong privacy guarantees, Proc. Priv. Enhan. Technol. 2018 (1) (2018) 67–87.

[23] N. Patki, R. Wedge, K. Veeramachaneni, The synthetic data vault, in: Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2016, pp. 399–410.

[24] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graph. Stat. 5 (3) (1996) 299–314.

[25] M.L. Mundkur, F.M. Callaghan, S. Abhyankar, C.J. McDonald, Use of electronic health record data to evaluate the impact of race on 30-day mortality in patients admitted to the intensive care unit, J. Rac. Ethn. Health Disparit. 4 (4) (2017) 539–548.

[26] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care IIii (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.

[27] L. Fuchs, C.E. Chronaki, S. Park, V. Novack, Y. Baumfeld, D. Scott, S. McLennan, D. Talmor, L. Celi, ICU admission characteristics and mortality rates among elderly and very elderly patients, Intens. Care Med. 38 (10) (2012) 1654–1661.

[28] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), JAMA 315 (8) (2016) 801–810.

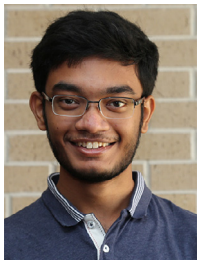
[29] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Med. Care 36 (1) (1998) 8–27.

[30] L.A.G. Celi, R.J. Tang, M.C. Villarroel, G.A. Davidzon, W.T. Lester, H.C. Chueh, A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury, J. Healthc. Eng. 2 (1) (2011) 97–110.

[31] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study, JAMA 270 (24) (1993) 2957–2963.

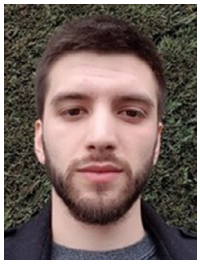
Andrew Yale is a Computer Science Ph.D. candidate at Rensselaer Polytechnic Institute. His research is in machine learning and specifically studying the generation of privacy preserving synthetic data and its application to health informatics.

Saloni Dash is an undergraduate student at BITS Pilani – Goa, majoring in Computer Science and Mathematics. Her research interests lie in leveraging Machine Learning tools for solving critical problems in Healthcare and Sustainability.



Ritik Dutta is an undergraduate student in the B.Tech in Computer Science and Engineering program at the Indian Institute of Technology Gandhinagar, India.

Isabelle Guyon is professor of data science at Université Paris-Saclay (UPSud/INRIA, Orsay), specialized in statistical data analysis, pattern recognition and machine learning. Her areas of expertise include computer vision, bioinformatics, and power systems. Her recent interest is in applications of machine learning to the discovery of causal relationships. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organized many challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, DARPA, and the European Commission, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. Isabelle Guyon holds a Ph.D. degree in Physical Sciences of the University Pierre and Marie Curie, Paris, France. She is president of Chalearn, a non-profit dedicated to organizing challenges, action editor of the Journal of Machine Learning Research, editor of the Springer series of Challenges in Machine Learning, and served recently as program co-chair of NIPS 2016 and general co-chair of NIPS 2017.



Adrien Pavao studied at Université Paris-Saclay (UPSud/INRIA, Orsay) and is now a Data Scientist Consultant working on organization of “Automatic Deep Learning” challenges (AutoDL series) with Google, Chalearn and INRIA as well as research on medical applications of generative models.

Kristin P. Bennett is the Associate Director of the Institute for Data Exploration and Application and a Professor in the Mathematical Sciences and Computer Science Departments and at Rensselaer Polytechnic Institute. Her research focuses on extracting information from data using novel predictive or descriptive mathematical models and data visualizations, and the applications of these methods to support decision making and to accelerate discovery in science, engineering, public health and business. She has 25 years of experience and over 100 publications in these areas. As an active member of the machine learning, data mining, and operations research communities, she has served as present or past associate or guest editor for ACM Transactions on Knowledge Discovery from Data, SIAM Journal on Optimization, Naval Research Logistics, Machine Learning Journal, IEEE Transactions on Neural Networks, and Journal on Machine Learning Research. She served as program chair of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. She has a Ph.D. in Computer Sciences from the University of Wisconsin-Madison. She founded and directs the NIH sponsored "TB-Insight" project which provided molecular epidemiology tools and methods to help track and control tuberculosis. She also founded and directs the "Data Analytics Throughout Undergraduate Mathematics" or DATUM which is pioneering highly effective new approaches for data analytics undergraduate education. She also founded the Data Interdisciplinary Challenges Intelligent Technology Exploration Laboratory (Data INCITE Lab.) In the Data INCITE Lab, undergraduate and graduate students tackle open applied data analytics problems contributed by industry, foundations, and researchers.