Privacy Preserving Synthetic Health Data

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, Kristin P Bennett

Privacy Preserving Synthetic Health Data

Andrew Yale[1], Saloni Dash[3], Ritik Dutta[4], Isabelle Guyon[2],
Adrien Pavao[2], Kristin P. Bennett[1] *

1- Rensselaer Polytechnic Inst., New York, 2- UPSud/INRIA U. Paris-Saclay, France,
3- BITS Pilani, Department of CSIS, Goa Campus, India, 4- IIT Gandhinagar, India

Abstract. We examine the feasibility of using synthetic medical data generated by GANs in the classroom, to teach data science in health informatics. We present an end-to-end methodology to retain **instructional utility**, while preserving **privacy** to a level, which meets regulatory requirements: (1) a GAN is trained by a certified medical-data security-aware agent, inside a secure environment; (2) the final GAN model is used outside of the secure environment by external users (instructors or researchers) 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, and/or cause loss of data fidelity. We benchmark our proposed GAN versus various baseline methods using a novel set of metrics. At equal levels of privacy and utility, GANs provide small footprint models, meeting the desired specifications of our application domain. Data, code, and a challenge that we organized for educational purposes are available.

1 Introduction

Teaching data analysis with actual medical data such as electronic healthcare records (EHR) is greatly restrained by laws protecting patients' privacy, such as HIPAA in the United States. While beneficial, these laws severely limit access to medical data thus stagnating innovation and limiting educational opportunities. The process of obfuscation of medical data is costly and time consuming with high penalties for accidental release. Health histories recovered from obfuscated data may result in discrimination. Research and education using EHR are highly skewed to a few shareable datasets such as MIMIC-III (Medical Information Mart for Intensive Care), which consists of de-identified ICU (intensive care unit) data from 2001 to 2012 [1]. While MIMIC-III is extremely useful, it is limited to ICU data, therefore does not give access to the entire medical history of patients, hence limiting the types of problems that can be studied. This paper addresses the problem of making a wider variety of medical datasets available to medical students and researchers by create synthetic data retaining utility for teaching purposes, and ideally even for research, while definitively preserving privacy. Our proposed workflow (Figure 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 develop a novel Wasserstein GAN and conduct a benchmark study on MIMIC data comparing it to 5 other approaches using a battery of metrics of utility, resemblance and privacy.

*This work is supported by the United Health Foundation. It was initiated with the collaboration of Thomas Gerspacher.

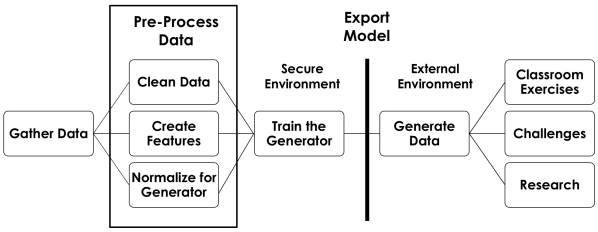

Fig. 1: Workflow used to generate synthetic data securely.

2 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 $\mathcal{S}_T = \{(\mathbf{x}_T^1, y_T^1), \cdots (\mathbf{x}_T^n, y_T^n)\}$ from P_T and $\mathcal{S}_S = \{(\mathbf{x}_S^1, y_S^1), \cdots (\mathbf{x}_S^n, y_S^n)\}$ from P_S. We assume that in all cases $\mathbf{x}$ variables belong to a **common metric space** *e.g.,* $\mathbb{R}^d$ and y is a categorical or continuous variable (*i.e.,* defining classification or regression tasks). We also assume that all variables have been **standardized**, e.g. by subtracting the mean and dividing by the standard deviation[1].

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 $\mathbf{x}_T^i \in \mathcal{S}_T$ and its nearest neighbor in $\mathcal{S}_S$. 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:

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

where the indicator function $\mathbf{1}(.)$ takes value 1 if its argument is true and 0 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 the 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 G. We denote by $E(.)$ the expected value

[1] To be perfectly correct machine-learning-wise, the mean and standard deviation must be estimated with training data and the same value applied to test data (when appropriate).

over the randomness of A_i and define 3 kinds of losses:

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

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 0. 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 0. 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 0 (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). Figure 2 provides a 2 dimensional synthetic example of these three cases in which red is the real data and blue is the artificial data. We generate train and test data ($n = 50$ from a $\frac{1}{2}$ circle plus Gaussian noise then standardized). We generate 2 artificial datasets A_1 and A_2 of the same size with the Parzen Windows density estimator, using a Gaussian kernel of varying width 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 show for each case. For the same example, Figure 3 provides curves representing Train $\mathcal{AA}$, Test $\mathcal{AA}$, and the Privacy Loss for decreasing Parzen Windows kernel widths.

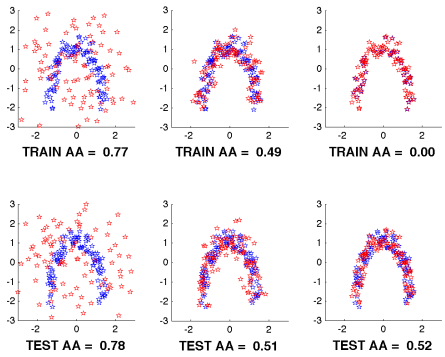

Fig. 2: **Parzen Windows, toy example**. Blue markers represent "real" 2-d data samples (R_{tr} and R_{te}) and red markers artificially generated data with Parzen Windows(A_1 and A_2). *Top row:* R_{tr} and A_1. *Bottom row:* R_{te} and A_2. *Form left to right:* Large kernel ⇒ underfitting; optimized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

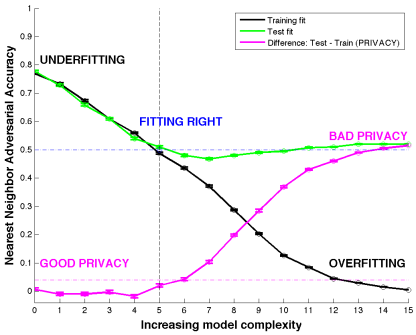

Fig. 3: **Parzen Windows, learning curves**. The Train $\mathcal{AA}$ keeps decreasing, but not the test $\mathcal{AA}$. The privacy is good when the difference Test $\mathcal{AA}$ - Train $\mathcal{AA}$ is small. The best compromise is attained around the point where the black curve crosses the dashed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

3 Privacy-Preserving Data Generation Methods

We performed a comparison of 6 data generative methods[2] on the MIMIC-III mortality problem: (1) **Gaussian Multivariate** [2], (2) **Wasserstein GAN** (WGAN) [3, 4], (3) Parzen Windows [5], (4) Additive Noise Model (ANM) [6], (5) Differential Privacy preserving data obfuscation (DP) [7], and (6) Copy the original data (CP). All methods are described in some detail in supplemental material. Briefly: Method (1), **Gaussian Multivariate**, uses only low order distribution statistics, mean and covariance matrix. It is expected to underfit data. Method (2), **WGAN**, is our advocated method, it fits data well while protecting privacy. Method (3), Parzen Windows, is a well-known classical statistics density estimation method, which is a good baseline method; when the kernel width is fitted, it approximates well the original distribution, but has the draw-back that all samples must be stored. Method (4), ANM, models all columns of the data matrix as a function of all others plus Gaussian noise; it is a good baseline method, but also requires storing predicted columns. This violates our **footprint requirements** which are no export of actual data, and size much smaller that the original real dataset. Method (5), DP, allegedly protects privacy, but in fact does so only for quasi-identifiers (a small subset of the columns of the datasets); we find that this provides insufficient privacy protection. Additionally, exporting data in any form is forbidden in our protocol. Method (6), CP, is only here for comparison purposes. Methods fulfilling our footprint requirements are emphasized in **bold**. The use of WGAN and ANM in this context are novel, to the best of our knowledge.

4 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 indicators whether they died in the ICU. It includes demographics, vital signs, diagnoses, and procedures performed. We generated the data, evaluated the different approaches using the proposed metrics, and then deployed the WGAN synthetic data in an online challenge used in courses at Rensselaer Polytechnic Institute[3] The challenges proved to be highly effective for students but here we focus on the quality of the synthetic data.

We compared the adversarial accuracy (Equation (3)) in terms of TrainResemblanceLoss, TestResemblanceLoss, and PrivacyLoss = TestResemblanceLoss - TrainResemblanceLoss. Gaussian Multivariate preserves privacy, but suffers from high testing adversarial accuracy (0.55). A well fitted Parzen Window (optimized kernel width) and WGAN both perform well with respect to resemblance and privacy. But the footprint of Parzen Windows rules it out for this purpose. 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

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

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

few and deep trees to illustrate a case of overfitting: indeed the ANM overfits the data badly, completely exposing the original 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.

		Methods					
		Wasserstein GAN (*)	**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	Balanced Accuracy	0.58	0.50	0.52	0.51	0.62	0.63
Footprint	Up-to-specs	yes	yes	no	no	no	no

Table 1: **Comparison of models with respect to various metrics**. Bold: Excellent; Plain: Good; Italics: 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, as opposed to data (real or synthetic).

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 on the real test dataset. Using a logistic regression model and comparing the balanced accuracy 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 method is WGAN, which performs well for balanced accuracy and still retains privacy. Gaussian Multivariate, Parzen Windows, and ANM perform the worst. The utility metric is important to consider, because it roughly captures usefulness of the synthetic data in the classroom setting.

5 Discussion, Conclusions and Future Work

GANs have been very popular in the recent years; when we started this work, their effectiveness was not clear to us. We first experimented with medGAN [8] with mixed results. WGAN made us make a big leap forward. The workflow that we presented for generating synthetic data from real data an exporting a model only outside a data-secure environment has become operational with the introduction of WGAN. 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 6 methods on the MIMIC-III mortality data. WGAN was the only effective method that maintained privacy and that allowed model export. 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.

References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[2] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[4] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[5] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076, 09 1962.

[6] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[7] C. Dwork. Differential privacy. *Automata, Languages and Program- ming*, 4052:1–12, 2006.

[8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *Machine Learning for Healthcare*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Fabian Prasser, Johanna Eicher, Raffael Bild, Helmut Spengler, and Klaus A. Kuhn. A tool for optimizing de-identified health data for use in statistical classification. *30th IEEE International Symposium on Computer-Based Medical Systems*, June,2017.

[11] Klaus A. Kuhn Raffael Bildraffael and Fabian Prasser. Safepub: A truthful data anonymization algorithm with strong privacy guarantees. *Proceedings on Privacy Enhancing Technologies*, Volume 2018: Issue 1:67–87, 2018.

[12] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 399–410. IEEE, 2016.

6 Supplemental Material

6.1 Data Generation Methods

We describe the 6 data generation methods in more detail.

- **Gaussian Multivariate:** This method simply consists in 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 original data and does not directly represent any sample (provided that the means are not actual data points).

- **Wasserstein GAN:** We developed a Wasserstein GAN (WGAN). We found WGAN to be much more effective on mixed continuous and categorical data, such as MIMIC-III, than the prior medGAN [8] model. The WGAN uses the earth mover's distance or Wasserstein distance versus Kullback-Leibler (KL) divergence [9] used in medGAN. WGAN represents an attractive black box method with a very compact footprint (parameters of the model) since the bottleneck in WGAN is constructed to prevent memorization.

- **Additive Noise Model:** Inspired from 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.

- **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 bold 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 Original 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 [7]. We generated a ϵ, δ Differentially Private version of the MIMIC-III dataset by creating generalization hierarchies for the 7 quasi-identifier attributes[4] using ARX, an open source anonymization tool for medical data [10] based on the SafePub Algorithm [11].The footprint of this method is unacceptable because it requires export of most of the original data and privacy is limited to quasi-identifiable fields.

6.2 Data Preprocessing

Data transformation was essential for the success of the WGAN. Recall MIMIC-III contains a mix of categorical and discrete variables. We adapted data transformation strategies used in the Synthetic Data Vault (SDV) [12]. We map all features to range between 0 and 1, 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 0 to 1 into sections based on the cumulative probability for each category. Finally, lining up each category with its section on the interval from 0 to 1, we take a sample from that section using a truncated Gaussian distribution. The reverse transformation maps the synthetic data to the original categories.

6.3 PCA Plots

We found 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) the generated test data in (red dots). Here we can see that Gaussian Multivariate and Parzen Windows span a larger space than the original data, which aligns with the fact that those methods create differences in the data in both directions uniformly. The Differential Privacy PCA spans a smaller space, which represents fact that the quasi-identifiers are changed enough to not reveal outlier data. Both the real and synthetic data distributions of WGAN 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.

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

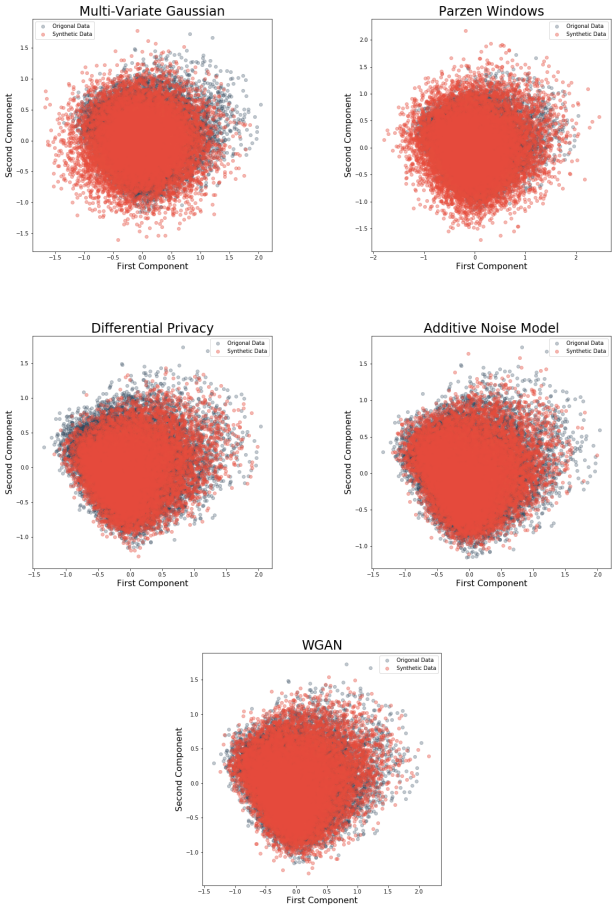

Fig. 4: Comparison of generative methods using PCA projection created using the real data