Assessing Privacy and Quality of Synthetic Health Data

Andrew Yale
yalea@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York

Saloni Dash
salonidash77@gmail.com
BITS Pilani, Department of CSIS, Goa Campus
Goa, India

Ritik Dutta
dutta.ritik@iitgn.ac.in
IIT Gandhinagar
Gandhinagar, India

Isabelle Guyon
guyon@clopinet.com
UPSud/INRIA U. Paris-Saclay
Paris-Saclay, France

Adrien Pavao
adrien.pavao@gmail.com
UPSud/INRIA U. Paris-Saclay
Paris-Saclay, France

Kristin P. Bennett
bennek@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York

ABSTRACT

This paper builds on the results of the ESANN 2019 conference paper "Privacy Preserving Synthetic Health Data" [16], which develops metrics for assessing privacy and utility of synthetic data and models. The metrics laid out in the initial paper show that utility can still be achieved in synthetic data while maintaining both privacy of the model and the data being generated. Specifically, we focused on the success of the Wasserstein GAN method, renamed HealthGAN, in comparison to other data generating methods.

In this paper, we provide additional novel metrics to quantify the susceptibility of these generative models to membership inference attacks [14]. We also introduce Discriminator Testing, a new method of determining whether the different generators overfit on the training data, potentially resulting in privacy losses.

These privacy issues are of high importance as we prepare a final workflow for generating synthetic data based on real data in a secure environment. The results of these tests complement the initial tests as they show that the Parzen windows method, while having a low privacy loss in adversarial accuracy metrics, fails to preserve privacy in the membership inference attack. Only HealthGAN shows both an optimal value for privacy loss and the membership inference attack. The discriminator testing adds to the confidence as HealthGAN retains resemblance to the training data, without reproducing the training data.

ACM Reference Format:
Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2019. Assessing Privacy and Quality of Synthetic Health Data. In *Artificial Intelligence for Data Discovery and Reuse 2019 (AIDR '19), May 13–15, 2019, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3359115.3359124

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

AIDR '19, May 13–15, 2019, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-7184-1/19/05…$15.00
https://doi.org/10.1145/3359115.3359124

1 INTRODUCTION

Privacy concerns frequently prevent dissemination of datasets containing sensitive data such as personal health data. In the United States, laws such as HIPAA prevent sharing of electronic health records (EHR) to protect the privacy of patients. Synthetic data generation methods provide an attractive alternative for making data available for research and education purposes without violating privacy. This paper focuses on new ways to empirically assess if a synthetic dataset is truly private while still retaining its utility (usefulness to solve a given problem) and resemblance (how close the distribution of synthetic data distribution matches the original data distribution). "Privacy Preserving Synthetic Health Data" [16] introduced novel metrics for measuring the quality of the synthetic data generators, and investigated multiple synthetic data generative methods with respect to these metrics. These generative methods were a Wasserstein GAN [1, 6] method being called HealthGAN, a Gaussian Multivarate[3] method, a Parzen Windows[10] method, an Additive Noise Model[8], a Differential Privacy preserving data obfuscation[4, 12] method, and simply "copying the training data" method. Of these HealthGAN and Additive Noise Model are novel and the others are taken as baselines to compare against since several have obvious privacy, utility, and resemblance characteristics. For example, "copying the training data" and over-fit Parzen Windows have excellent utility and resemblance but unacceptable privacy, while the Gaussian Multivariate method has high privacy but poor utility and resemblance. The Differential Privacy method only protects information for the 7 quasi-identifier columns[1] and leaves other columns as real data therefore having unacceptable privacy. Additionally model privacy was another characteristic of the different methods. This evaluates whether the model contains any of the real data and therefore doesn't retain privacy. It was determined that only HealthGAN and Gaussian Multivariate methods preserve model privacy. While not all the methods were viable as a final method choice due to privacy concerns, the different types of methods showed different pros and cons to styles of generators.

To test these methods, we developed the concept of *nearest neighbor adversarial accuracy* and *privacy loss*. Nearest neighbor adversarial accuracy, shown in Equation 1, compares the distance from one point in a target distribution T, to the nearest point in a source distribution S, defined as $d_{TS}(i) = \min_j \|\mathbf{x}_T^i - \mathbf{x}_S^j\|$, to the distance to the next nearest point in the target distribution, defined as

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

$d_{TT}(i) = \min_{j,j\neq i} ||\mathbf{x}_T^i - \mathbf{x}_T^j||$. By comparing this across all points, it gives us the adversarial accuracy. This metric can be interpreted much like balanced accuracy where the value is an average of the accuracy for each class. Therefore we are striving for a value of 0.5 where the synthetic and real data cannot be distinguished. If that is achieved on both the training and test datasets, then privacy is said to be preserved.

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

Privacy loss is defined in Equation 2 as the difference between the adversarial accuracy on the test set and the adversarial accuracy on the training set. As the ideal value for both of these is 0.5, the privacy loss should be 0.0 when privacy is completely conserved. In the case where the model is exposing data, the value of the training adversarial accuracy will be lower than 0.5, and therefore even if the test adversarial accuracy is 0.5, the loss will increase.

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

In this paper, the focus is on improving the methods for measuring the quality of the data, both in terms of the privacy and the utility. The first improvement is by measuring the effect of membership inference attacks. This is a novel way to measure how well privacy is maintained across the different methods. Second, we develop a new measure of how well the synthetic data resembles the original data. Specifically, we exploit the discriminator that distinguishes between real and synthetic data created in the training of HealthGAN. We assess how well synthetic points are predicted to be synthetic points by HealthGAN. These results are compared with previous privacy loss and utility results on these same methods.

2 MEMBERSHIP INFERENCE ATTACKS

In a membership inference attack scenario, an attacker attempts to determine whether a given record was used to train a model [15]. In this scenario the attacker also has black-box access to the model, meaning they have the ability to feed data into the model and observe the output of the model [13, 14]. The original scenario doesn't exactly match what would happen with HealthGAN because the input to HeathGAN generator network is random noise, rather than real data. Therefore in HealthGAN setting the model the attacker has access to is just the generator and cannot train the model, only feed it random noise in order to generate data. Therefore, instead we show how using the synthetic data generated from the network and a variant of nearest neighbor accuracy can be used to assess vulnerability to this kind of attack. Membership inference attacks are important to prevent because if an attacker can infer the membership of a patient in a training set for a model then they can infer other information about the patient. For instance, if the cohort is all diabetic patients and membership can be inferred then the attacker knows that the patient is diabetic. Even more importantly, membership inference attacks lead to additional attacks such as profiling and property inference[7]. The following attack is an adapted formulation of the original membership inference attack specifically for testing a GAN generator.

In the attack scenario we are considering, an attacker has access to some real data R with incomplete records for each patient. Specifically, we can assume, without loss of generality, that the attacker has access to columns $[c_1 \dots c_k]$, but not to columns $[c_{k+1} \dots c_N]$. Simultaneously, the attacker has access to a synthetic (artificial) dataset A for which all columns $[c_1 \dots c_N]$ are given, which allows him/her to create a predictor of columns $[c_{k+1} \dots c_N]$ from columns $[c_1 \dots c_k]$. Subsequently, this could allow him/her to predict the missing columns in real data, which could constitute a breach of privacy. This violation of privacy can be quantified in the membership attack scenario context by evaluating the fraction of real data records that can be identified after completing the missing data in R.

We place ourselves in the worst case scenario, in which the attacker has available a large fraction of the columns in R, making the attack simpler. We consider the limit case in which all the columns are available, and determine how easy it is to identify which real data records were used for training our data generative model. We construct R to be a random shuffle of the training and non-training data, and attempt to sort out if each point is from training or not, using a nearest neighbor classifier. We compute the distance from a sample in R to it's nearest neighbor in A, then measure the AUC of prediction { training *vs.* non-training sample } using the measured nearest neighbor distance as a ranking measure. If the AUC is greater than 0.5 (chance level), then the model may be exposing private data by allowing the attacker to figure out which records are in training.

This experiment was conducted on the same dataset from Medical Information Mart for Intensive Care (MIMIC)[9] used in the previous paper. The dataset includes categorical demographic data such as religion, language, and ICU admission type. It also includes vital signs from the first 48 hours in the ICU in the form of continuous values, such as heart rate, blood pressure, and temperature. Finally, there is binary data in the form of diagnosis categories as defined by the Clinical Classification Software (CCS)[2] codes and mortality of the patient after 48 hours. The task is to predict if a patient died or not using the other available features. Note since HealthGAN requires numeric features, categorical features are converted into values between 0 and 1 using the synthetic data vault formulation [11]. The results of the experiment for each method are shown in Figure 1 and Table 1. In addition to the membership inference results the privacy loss and utility are included from the previous metrics. The utility is now measured using the area under the curve value when predicting mortality instead of balanced accuracy on data with the proportion of mortality matching the original dataset.

The membership inference test results show that the worst methods are the Training Data and the Additive Noise Model. This makes intuitive sense because the training data being passed of as synthetic are the most over-fit model that can be created. The Additive Noise Model is also an over-fitting model and therefore it isn't surprising to have low privacy. The next worst is the Differential Privacy method, which only obscures some of the data and therefore also has the same issue as the training data in revealing real

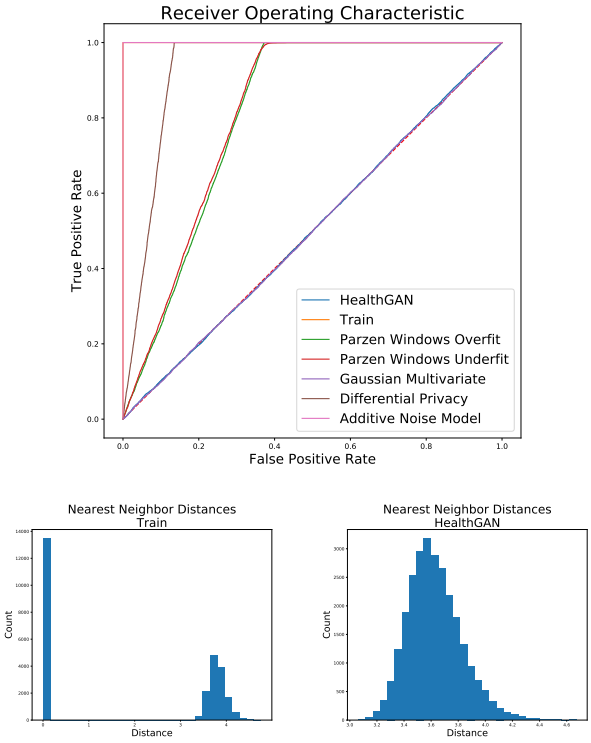

Figure 1: Results from the membership inference attack. The top plot shows the ROC curves and AUC values for each method. The optimal methods are HealthGAN and Gaussian Multivariate, which follow the diagonal and have the an AUC of 0.5. The Additive Noise Model and Training data follow a right angle to the upper left and show the worst case scenario with an AUC of 1.0. The bottom plots show the distribution of distances between the synthetic data and the real data being tested. For the left plot treating the Training data as synthetic data, distances between the training and non training data are easily separable because the distances to the training data are always zero. In the right plot, distances for HealthGAN synthetic data distances are all very similar, and making it hard to distinguish the training data from the non training data.

data. The Parzen Windows methods actually perform similarly, whether on the last iteration or the first. The initial window size was 0.28, and under-fits the training data in general, and by the last iteration the window is down to 0.00028, where the data overfits the model. Despite these differences, they still have approximately the same exposure in a membership inference attack. Finally, the best methods are HealthGAN and Gaussian Multivariate which have the optimal result of 0.5. In the membership inference attack scenario, these two methods don't reveal any information about the use of a specific data point in the training set.

Comparing these results to the privacy and utility results of the last paper, we can see that the membership inference attacks are

Table 1: Results from the membership inference attack compared with privacy loss and utility metrics. Membership inference is measured in area under the curve. A value of 0.5 shows that the ability to distinguish data used for training from data that wasn't is no better than random. A value of 1.0 shows that those two classes can be perfectly separated, and thus the method may be vulnerable to a membership attack. PW = Parzen Windows, GM = Gaussian Multivariate, DP = Differential Privacy, ANM = Additive Noise Model.

Method	Mem. Inf.	Priv. Loss	Utility
Train	1.00	0.50	0.88
ANM	1.00	0.50	0.74
DP	0.93	0.47	0.87
PW Over-fit	0.81	0.00	0.87
PW Under-fit	0.82	0.00	0.77
GM	0.50	0.02	0.62
HealthGAN	0.50	0.00	0.66

exposing new privacy concerns with both of the Parzen Windows methods. In the adversarial accuracy metrics, the privacy loss of those methods was 0.0 and seemed to indicate that they had good privacy, but with the results of the membership inference attack we can see that while the data itself might seem private it is exposing information in the form of membership inference information. With the other methods we can see that the membership inference results confirm the privacy evaluation of the previous metrics. HealthGAN and Gaussian Multivariate methods both score well on both privacy metrics. Additionally the Training, Differential Privacy, and Additive Noise Model methods all score poorly on both metrics as is expected since these models over-fit.

3 DISCRIMINATOR TESTING

In addition to the nearest-neighbor adversarial accuracy and utility metrics, there are other ways to measure the quality of the synthetic data. As part of training HealthGAN, there is a discriminator network and a generator network[5]. The discriminator network measures the Wasserstein distance[1, 6] from each record in a batch of data to the modeled distribution of the real data. The farther this distance is from zero, the more likely the data is to be synthetic. Therefore, another way to test the quality of the synthetic data, as well as whether the discriminator itself seems to be functioning as expected, is to look at how that discriminator distinguishes synthetic datasets generated by other methods.

The discriminator network was tested on several generative methods and datasets and the results are shown in Table 2. The first dataset is the training data used by the model. The result of the training data is predictably the closest average distance to zero of all the methods, as the discriminator is modeled based on that data. The test data was also tested with the discriminator and had a mean of -0.065. This mean is also small but far larger than that of the training data. This indicates that there is a potential that the discriminator might be over-fitting to the training data. If a model over-fits the training data this could indicate a lack of privacy in the generated data. It also shows that a mean distance of data less than the testing data might indicate over-fitting in the

generative method. This happens with the over-fit version of the Parzen Windows method and the Differential Privacy method. The Parzen Windows method has been trained for too many iterations and is generating data too similar to the original data. This is different from the result of the membership inference attack, where the under-fit and over-fit models were similar. The difference in the results shows the importance of using different metrics and comparing them. The Differential Privacy method only obscures some of the features in the original data and therefore still has a smaller distance to the training data than even the test data has. On the other end of the spectrum is the under-fit Parzen Windows method, which has the farthest distance from the real data distribution. This shows how poorly the data fits the original data. Finally, the last three methods, Additive Noise Model, Gaussian Multivariate, and HealthGAN, show distances farther than the test data but close enough to still retain utility.

Table 2: Results from the discriminator test compared with the membership inference, privacy loss, and utility metrics. The discriminator results are measured as mean Wasserstein distance away from the modeled training distribution. PW = Parzen Windows, GM = Gaussian Multivariate, DP = Differential Privacy, ANM = Additive Noise Model.

Method	Disc. Mean	Mem. Inf.	Priv. Loss	Utility
Train	-0.013	1.00	0.50	0.88
PW Over-fit	-0.013	0.81	0.00	0.87
DP	-0.031	0.93	0.47	0.87
ANM	-0.107	1.00	0.50	0.74
HealthGAN	-0.120	0.50	0.00	0.66
GM	-0.136	0.50	0.02	0.62
PW Under-fit	-0.266	0.82	0.00	0.77

Combining these results with the membership inference results from the previous section and the privacy loss and utility results from the previous paper, we can make conclusions on the overall privacy and utility of the data. As expected, copying the training data shows predictable results across the board; it has the worst privacy values and best utility values. The Differential Privacy method, which only has slight changes to obscure the quasi-identifiers in the data, performs similarly to the training data. The Additive Noise Model isn't obviously over-fitting in the discriminator mean, but in the membership inference and privacy loss, we can see that it has poor privacy retention. The Parzen Windows methods seemed to preserve privacy and have high utility based on the previous metrics, but in both the membership inference results and the discriminator testing we can see that they expose information about the real data. In a membership inference attack, they have a poor value of AUC and in the discriminator mean we can see the over-fit model is too close to the training data, closer than the test data, and the under-fit model is the farthest away of all of the methods. Finally, HealthGAN and Gaussian Multivariate perform very similarly, but with HealthGAN edging out the Gaussian Multivariate in the privacy loss, utility, and discriminator mean. For both of these methods, the privacy values are close to the best possible, and the utility values, while lower than the methods with poor privacy, are still showing utility in the mortality task.

4 CONCLUSION

Due to the critical privacy concerns with medical data, it is important to ensure that not only does no real data get generated by the model, but no information about the real dataset is exposed by the synthetic data. By combining the results from the membership inference attack experiment with the discriminator testing experiment, we can see a clearer picture of how our methods balance privacy and utility. First, we see the simplistic methods fail at the privacy metrics, while succeeding at resemblance and utility. Second, we see some of the methods having worse privacy than was clear in the initial metrics. Finally, we see that two of our methods that preserve model privacy are validated with the new tests.

These additional metrics validate HealthGAN's privacy retention, but still show gaps in the utility of the generated data. In general, data generated by GANs does not model outliers well, but within the field of health data that type of data can be very common. Therefore, an open question is how to improve HealthGAN in order to model and synthesize outlier and rare data without compromising on high levels of privacy.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

[2] Healthcare Cost, Utilization Project, et al. 2016. Clinical classifications software (CCS) for ICD-9-CM. *last modified October* 7 (2016).

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

[4] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.

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

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

[7] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2017. LOGAN: Membership Inference Attacks Against Generative Models. *arXiv preprint arXiv:1705.07663* (2017).

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

[9] 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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.

[10] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33, 3 (1962), 1065–1076.

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

[12] Fabian Prasser, Johanna Eicher, Raffael Bild, Helmut Spengler, and Klaus A Kuhn. 2017. A tool for optimizing de-identified health data for use in statistical classification. In *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*. IEEE, 169–174.

[13] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[15] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Transactions on Services Computing* (2019).

[16] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2019. Privacy Preserving Synthetic Health Data. In *Proceedings of the 27. European Symposium on Artificial Neural Networks ESANN*. 465–470.