Synthesizing Quality Open Data Assets from Private Health Research Studies

Andrew Yale, Saloni Dash, Karan Bhanot, Isabelle Guyon, John S Erickson, Kristin P Bennett

Synthesizing Quality Open Data Assets from Private Health Research Studies

Andrew Yale[1,4], Saloni Dash[2], Karan Bhanot[1], Isabelle Guyon[3], John S. Erickson[1], and Kristin P. Bennett[1]

[1] Rensselaer Polytechnic Institute, Troy, NY
[2] BITS Pilani, Goa Campus, Goa, India
[3] UPSud/INRIA U. Paris-Saclay, Paris-Saclay, France
[4] OptumLabs Visiting Fellow

Abstract. Generating synthetic data represents an attractive solution for creating open data, enabling health research and education while preserving patient privacy. We reproduce the research outcomes obtained on two previously published studies, which used private health data, using synthetic data generated with a method that we developed, called HealthGAN. We demonstrate the value of our methodology for generating and evaluating the quality and privacy of synthetic health data. The dataset are from OptumLabs® Data Warehouse (OLDW). The OLDW is accessed within a secure environment and doesn't allow exporting of patient level data of any type of data, real or synthetic, therefore the HealthGAN exports a privacy-preserving generator model instead. The studies examine questions related to comorbidites of Autism Spectrum Disorder (ASD) using medical records of children with ASD and matched patients without ASD. HealthGAN generates high quality synthetic data that produce similar results while preserving patient privacy. By creating synthetic versions of these datasets that maintain privacy and achieve a high level of resemblance and utility, we create valuable open health data assets for future research and education efforts.

1 Introduction

The inability to share private health data can stifle research and education activities. For example, studies based on unpublished electronic medical record (EMR) data cannot be reproduced, thus future researchers are not able to use them to develop and compare new research. This contributes to the reproducibility crisis in biomedical research [3]. Making open data available for research can spur innovation and research. The public Medical Information Mart for Intensive Care datasets, MIMIC-II and MIMIC-III, are widely used with over 2000 citations reported in Google Scholar in March 2020 [7,10]. But since MIMIC-II and MIMIC-III focus on Intensive Care Unit patients in Boston hospitals, the resulting research may be biased and have limited generalization. Also since MIMIC requires users to undergo a training/approval process, it is not well suited for classroom use. The cost and time required, along with re-identification risk concerns make de-identification only a partial solution to this problem.

Recent synthetic data generation methods provide an attractive alternative for making data available for research and education purposes without violating privacy. Deep learning approaches for synthetic data specifically show significant promise [1,6,8] In the future, synthetic data generation methods combined with automatic machine learning methods could enable synthetic versions of data to be released when research papers are published. Results could be reproduced and novel methods and analysis could be developed without compromising patient privacy. Figure 1 illustrates one scenario for use of synthetic data. To accomplish this, synthetic data assets must have 1) privacy: how well does the synthetic generation data method preserve anonymity, 2) resemblance : whether the distribution of synthetic data is indistinguishable from the distribution of real data. 3) utility: can research studies be reproduced successfully with synthetic data and 4) efficiency: how practical is the training and generation pipeline.

In this paper, we report our experience of generating synthetic data using the process in Figure 1 for two published research studies [14,15] performed in the OptumLabs Data Warehouse (OLDW). The studies focus on the same cohort of children with ASD. The first focuses on the difference between gastrointestinal symptoms and oral antibiotic use in children with ASD and children without ASD. The second study investigates how different groups of children with ASD can be clustered based on their comorbid medical conditions (CMCs), and what that means about the different clusters. We utilize novel enhancements of the prior HealthGAN [4,16–18] synthetic data approach, and then evaluate the privacy, resemblance, efficiency, and utility of the synthetic data.

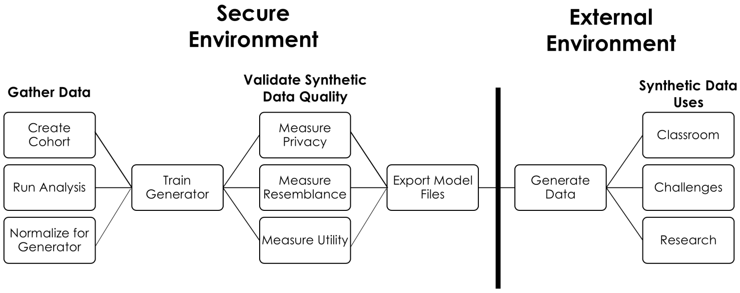

Fig. 1: Workflow used to generate synthetic data securely. The data is gathered in the same way as the studies did, processed, and used to train the generator model inside the secure environment. The synthetic data is then validated for privacy, resemblance, and utility. After being certified as private by the OptumLabs staff, the model is exported. Finally, data is generated using the model and used for many types of applications.

This paper focuses on documenting and assessing the HealthGAN synthetic data generation method on real medical datasets in a secure environment and exporting the models outside of the environment. We used the OptumLabs Data Warehouse as the source of the medical data. We report on the process of exporting the data from the OLDW[5]. We demonstrate how to verify privacy and resemblance using recently published metric. We make the tools for generating and evaluating synthetic readily available in a Python package. [6] The next section goes over synthetic data generation and evaluation methods. Following that the selected generation and evaluation methods are used to reproduce two studies in OLDW. The synthetic data for each of these datasets are evaluated for privacy, resemblance, and utility. Finally, we conclude and discuss future work.

2 Methods

2.1 Generation

We use HealthGAN [4, 16–18], a generative adversarial network (GAN) method we developed to generate synthetic data for datasets containing categorical, continuous, binary, and time series data. This method is based on the Wasserstein Generative Adversarial Network [2, 5] and uses a novel variant of the categorical encoding method from "The Synthetic data vault" (SDV) [9]. We added the ability to encode ordinal data to the original HealthGAN. By creating the SDV mapping using the inherent ordinal order and including values that might not exist in the original dataset, we improved resemblance of ordinal variables.

Part of the constraints for our generative model is the ability to export the model from the secure environment instead of synthetic data. To fulfill this requirement we needed to ensure that the model itself did not contain or require any real data to generate synthetic data. In "Privacy Preserving Synthetic Health Data" [17], after comparing multiple methods, we found that HealthGAN method best satisfies these constraints due to the fact that a neural network model does not store real data or require real data to run. For HealthGAN specifically, we export just the generator network and never export any actual data.

2.2 Evaluation

To test the synthetic data generated by HealthGAN we investigate privacy, resemblance, and utility. In "Privacy Preserving Synthetic Health Data" [17], 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 $d_{TT}(i) = \min_{j,j \neq i} ||\mathbf{x}_T^i - \mathbf{x}_T^j||$. 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

[5] As of 4/27, we successfully exported the model of the first dataset but approval on the second dataset is ongoing

[6] github.com/TheRensselaerIDEA/synthetic_data

0.5 where the synthetic and real data cannot be distinguished. If that is achieved then we can say that the synthetic data and real data have high resemblance.

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

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.

$$\begin{aligned} \textbf{TrResemblLoss}\ (\mathit{Train}\ \mathcal{A}\mathit{dversarial}\ \mathcal{A}\mathit{cc.}) &= E[\mathcal{AA}_{RtrA_1}] \\ \textbf{TrResemblLoss}\ (\mathit{Test}\ \mathcal{A}\mathit{dversarial}\ \mathcal{A}\mathit{cc.}) &= E[\mathcal{AA}_{RteA_2}] \\ \textbf{PrivacyLoss} &= \mathit{Test}\ \mathcal{AA}\ -\ \mathit{Train}\ \mathcal{AA} \end{aligned} \tag{2}$$

Beyond nearest neighbor adversarial accuracy, we test the privacy of the synthetic data using a membership inference attack scenario. In this scenario an attacker attempts to determine whether a given record was used to train a model [13]. The attacker 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 [11, 12]. The original scenario doesn't exactly match what would happen with HealthGAN because the input to HealthGAN generator network is random noise, rather than real data. 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.

In the attack scenario we are considering, an attacker has access to some real data R with incomplete records for each patient. We 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.

In the worst case scenario, 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 its 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 know which records are in training.

Finally, once the privacy and resemblance have been tested the utility of the synthetic data must be evaluated. The method for this varies based on what the real dataset was intended for, but in the case where we are replicating a published study we reproduce the analysis done on the original data on the synthetic data. For the two studies being reproduced in this paper we will use Cox regression and k-means clustering to analyze the performance of the synthetic data.

3 Research Studies Reproduced

We examine creating synthetic open-data for two research papers, including the synthetic generation method, and the approaches for validating the quality of the two generated datasets. Both of the original studies were performed in the OLDW. OptumLabs is an open, collaborative research and innovation center founded in 2013 as a partnership between Optum and Mayo Clinic with its core linked data assets in the OLDW. The database contains de-identified, longitudinal health information on enrollees and patients, representing a diverse mixture of ages, ethnicities and geographical regions across the United States. The claims data in OLDW includes medical and pharmacy claims, laboratory results and enrollment records for commercial and Medicare Advantage enrollees. The EMR-derived data includes a subset of EMR data that has been normalized and standardized into a single database. Access to the OptumLabs database is, justifiably, tightly controlled. Access is limited to partner researchers who are granted access to certified de-identified research data views. All work is completed in the secure environment in which OLDW is hosted. Any release of data or derived products like graphs and tables must be reviewed and approved by OptumLabs on a case by case basis. Data for the subjects in the studies are typically not released, which means the OptumLabs datasets used in published studies are typically not available.

3.1 Study 1: Gastrointestinal Symptoms and Oral Antibiotic Use in Children with Autism Spectrum Disorder

In "Gastrointestinal Symptoms and Oral Antibiotic Use in Children with Autism Spectrum Disorder: Retrospective Analysis of a Privately Insured U.S. Population" [15] the authors look at the relationship between taking oral antibiotics in early childhood and occurrences of later gastrointestinal (GI) diagnosis in children diagnosed with autism spectrum disorder (ASD). Previous studies have shown different rates of GI symptoms in children with ASD, but at least one study claims that the estimated odds of having a general GI complaint are 4.4 times greater for children with ASD than for children without ASD. One confounding factor in this comparison is the presence of oral antibiotics. Oral antibiotics in early childhood may cause long-term disruption in the gut microbiome's composition, leading to an increased number of GI symptoms in early childhood. Several studies have shown that on average children diagnosed with ASD consume more oral antibiotics than children without ASD. Therefore, this work looks at the hazard ratios of demographics, oral antibiotics, and ASD diagnoses. The authors found children with ASD had a greater rate of GI diagnoses than children without ASD. Examining hazard ratios, greater numbers of oral antibiotics significantly increased risks GI-related diagnoses for both groups.

EMR Data The study looks at OptumLabs claims data from 1/1/2000 to 9/30/2015. The patients that are used in the dataset must have at least five years of un-interrupted data called continuous enrollment to be included. From there they are included in the ASD cohort if the patient has at least two different ASD related diagnoses in the time period.

In order to measure the effect of oral antibiotics, the study defines two periods for observation: early enrollment and late enrollment. Early enrollment is the first three years of the five year period, while late enrollment is the last two years. The number of GI related diagnoses are counted for each period and examined. This type of analysis separates separate short term GI symptoms related to oral antibiotics from longer term conditions.

The ASD cohort collected has 3,278 patients while the non-ASD cohort has 279,428 patients. Within that cohort 37% of the ASD cohort has GI related diagnoses and 20% of the non-ASD cohort has GI related diagnoses. The final demographics of the study population are tabulated in Table 2 of their paper [15]. Due to slight variations in the view we are using in OptumLabs, we are not able to get the exact same data. Specifically, our view is a zoomed out version of the census divisions, which means our data has four different regions instead of the nine in this study. However, the data is almost exactly the same despite this change. The nine regions in the original study can be cleanly aggregated into the four regions in our data, reflected in the similar cumulative counts for the four regions across both datasets. Our dataset contains a total of 283,462 patients, which is very similar to the overall total of 281,623 patients in the original study.

Synthetic Data The first measure of quality for the synthetic data is our nearest neighbor adversarial accuracy metric [17]. The values for the *TrainResemblanceLoss* and the *TestResemblanceLoss* should be close to 0.5 to ensure that the resemblance is good, but most importantly the privacy loss, which is defined as *PrivacyLoss* = *TestResemblanceLoss* - *TrainResemblanceLoss*, should be as close to 0 as possible [17]. On our candidate model the *TrainResemblanceLoss* is 0.5236 and the *TestResemblanceLoss* is 0.5272. For the *PrivacyLoss*, we get 0.0036. Overall, these metrics indicate we are preserving privacy with a low privacy loss, but also maintaining resemblance with resemblance loss values close to 0.5.

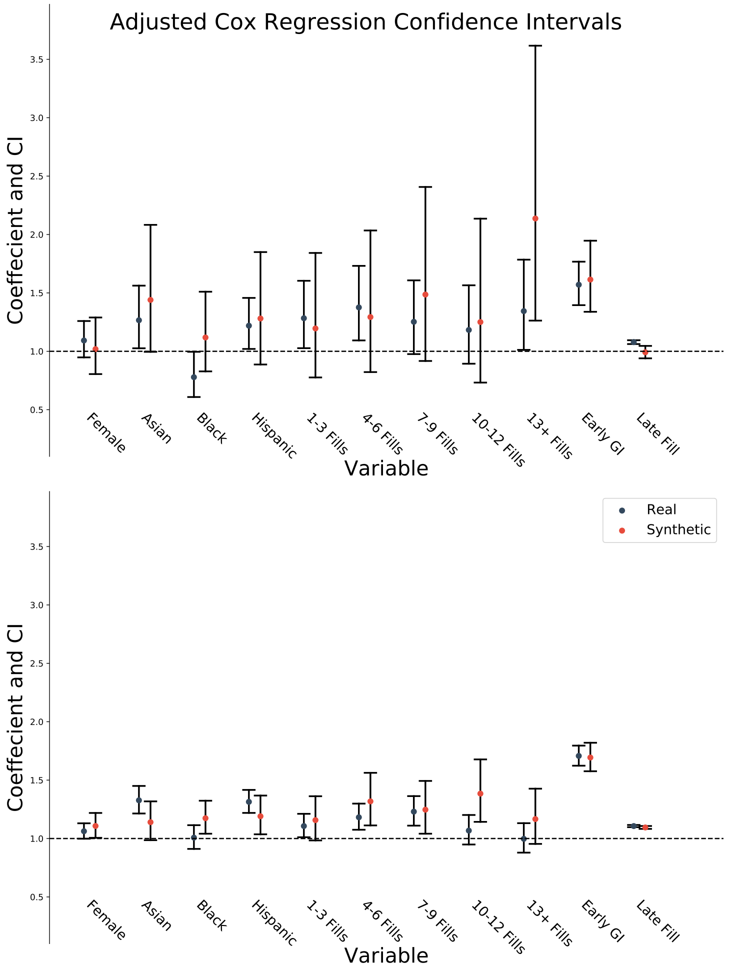

Fig. 2: Adjusted Cox Regression Hazard Ratio Confidence Intervals (CI) (95%) by different covariates observed on **ASD Cohort** (top) and **Population Cohort** (bottom) for real and synthetic data.

Another measure of privacy is robustness against a membership inference attack [16]. This measured uses the area under the curve (AUC). Any AUC value above 0.50 in this test indicates a potential loss of privacy in a membership inference attack. To verify the baseline for this metric, we run the membership inference attack using the real data and obtain an AUC of 1.00 as expected. When computing this measure for the synthetic data, we get an optimal value of 0.50. This result further affirms previous privacy metrics, indicating privacy is also preserved in the membership inference attack scenario.

Finally, we assess utility by reproducing the study on the synthetic data. In the original study, the analysis used Cox regression to estimate adjusted and unadjusted hazard ratios (HRs). The original results from the Cox regression model show that GI diagnoses in late enrollment are more likely to happen if oral antibiotics are taken in early enrollment, regardless of an ASD diagnosis. In addition, the similarity in the HR of the two groups indicates that the effect is not exclusive to the ASD population. Another covariate was a diagnosis for GI symptoms in early enrollment. Understandably, this variable has a high HR as it likely indicates systemic issues with the GI tract.

We run the same Cox regression analysis on the synthetic data in order to calculate the hazard ratios of the same covariates. In Figure 2 the ASD Cohort is on top and the Control Cohort is on the bottom, the real 95% confidence intervals are shown in black compared to the synthetic confidence intervals (CI) in red. If the CI overlap, then there is no statistically significant difference between the real and significant results. In the top of Figure 2, the only variable that does not overlap in confidence intervals is whether the patient had any late enrollment prescription fills. In the bottom of Figure 2, all the variables have overlapping confidence intervals, thus display no significant differences. In addition, the findings of the original paper state that the ASD and POP cohorts have similar hazard ratios for 7-9 fills. Our results verify this claim in both our real data with an ASD value of 1.25 CI (0.98, 1.61), and a POP value of 1.24 CI (1.11, 1.36), and in our synthetic data with an ASD value of 1.48 CI (0.92, 2.41), and a POP value of 1.24 CI (1.04, 1.49). These metrics indicate the synthetic data has high utility in terms of providing the same relationships demonstrated in prior work. Overall, the metrics computed demonstrate that the synthetic data created using our end-to-process result in data that retains privacy while maintaining high levels of resemblance and utility.

3.2 Study 2: Clustering of co-occuring conditions in ASD

In the paper "Clustering of co-occurring conditions in autism spectrum disorder during early childhood: A retrospective analysis of medical claims data" [14] the authors analyze patterns in diagnoses of comorbid medical conditions (CMCs) in patients with ASD. The study uses the same cohort of patients as the previous study. Based on the data, they are able to separate the patients with ASD diagnoses into three different clusters. The first cluster was 23.7% (n = 776) of the patients and encompassed a high rate of CMCs. The second cluster was 26.5% (n = 870) and contained patients with a higher rate of developmental delays. The third cluster, making up 49.8% (n = 1,632) of the data, contained low numbers of CMCs. Evaluating the data over time shows that the same patterns persist within these three clusters. The goal of this work was to help inform future treatment protocols for patients with ASD related to CMCs.

EMR Data Membership in the cohort of patients ASD was determined with the same criteria as the previous paper, but with the addition of data regarding the

CMCs. Seven different categories of CMCs were identified: auditory disorders, development delays, gastrointestinal symptoms, immune related conditions, psychiatric disorders, seizure disorders, and sleep disorders. These conditions were measured over time in six-month windows. For each patient, ten six-month windows were constructed for each of the categories, resulting in a vector of length 70 per individual with binary values indicating whether there was a diagnosis in the time window.

To analyze the CMC dataset the authors used principal component analysis (PCA) on the 70 CMC columns to transform the data to a continuous dataset, and then clustered the transformed data. This analysis was only done on the sub-cohort of patients with ASD diagnoses. Through k-means clustering they found the three different clusters of children with ASD defined previously.

Synthetic Data Following the same process as the first paper, we train a new HealthGAN model on this dataset. We compute the privacy methods for the data and we evaluate the nearest neighbor adversarial accuracy metric. For our candidate model the performance on *TrainResmeblanceLoss* is 0.5416 and our *TestResemblanceLoss* is 0.5430. This value results in a *PrivacyLoss* of 0.0014. These metrics demonstrate that this synthetic data retains privacy as well as maintaining resemblance values close to 0.5. We also examine membership inference attacks using AUC. After verifying the baseline for the real data, the synthetic data had an AUC of 0.50, meaning no privacy loss in a membership inference attack scenario as well.

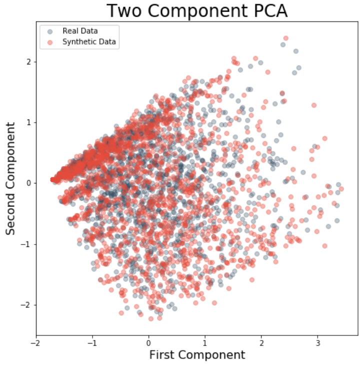

Fig. 3: PCA Plot of Real vs Synthetic Data for Study 2

Next we examine resemblance of the synthetic data compared to the real data. For Study 2, we look at the population cohort, ASD cohort, and each of the three specific clusters. In addition to the demographic data, we look at the average number of CMC categories diagnosed per cohort. Visualizing the first 2 components of PCA for the real data and the synthetic data can further help understand any differences. In Figure 3, PCA plots of the real and synthetic data (colored by cluster see below) show very similar distributions. We see that we have a good level of coverage over the real data. This is shown by the fact that we do not have many outlier values being generated in the synthetic data as well as the synthetic data not missing that many portions where the real data exists.

After measuring privacy and resemblance, we examine the utility of the synthetic data. Instead of Cox regression, the analysis of the original dataset is done with k-means clustering. To verify the utility of the synthetic data, this cohort is put through the same clustering method as the one used on the real data. The three emerging clusters are compared using characteristics of the CMCs for patients in each cluster.

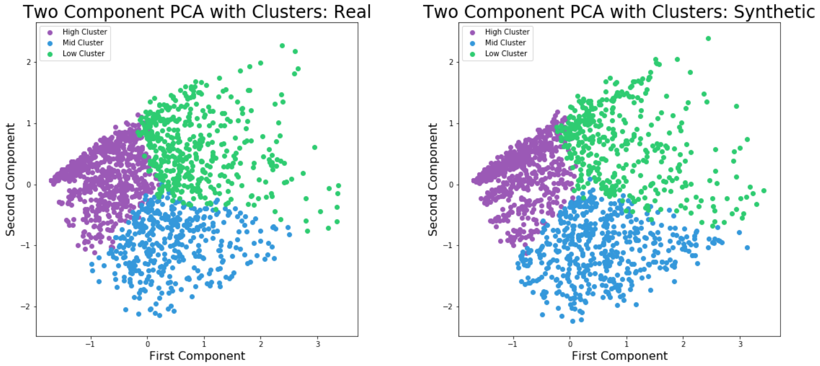

Fig. 4: PCA Plot of Real Data and Synthetic Data with Marked Clusters

Using the PCA plot that was created to check the resemblance of the data, we can also check the utility by looking at how well the k-means clustering method works on the synthetic data. In Figure 4, we have a PCA of the real data on the left and colored according to the clusters. On the right we have the PCA of the synthetic data colored by clusters found using the same k-means model found on real data. Visually we see that the clusters are very similar in both the real and synthetic plots which indicates a high level of resemblance and therefore utility.

In addition to the pure size and shape of the clusters, we can also look at the average number of CMCs over time in the clusters. In Figure 5, we can see a comparison of the average number of CMCs in each group over the time, in years, of the study. We see that there is a similar, if less smooth, trend in the groups. The high and mid clusters have a very similar relationship where they cross each other around the three year mark. The ASD population as a whole stays consistent in the middle of the graph, and the low and control populations have

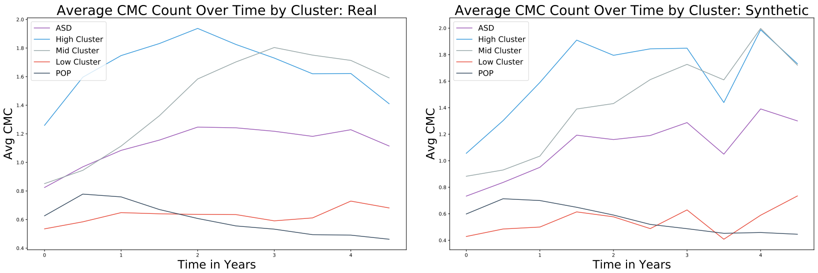

Fig. 5: Study 2: CMC Over Time by Different Clusters, Real and Synthetic

similar numbers of CMCs over time. Overall by replicating these various results from the original paper we can see that we achieve a high level of resemblance and utility while maintaining a high level of privacy[7].

4 Conclusions and Future Work

We demonstrate how to create synthetic datasets from real and assess their resemblance, privacy, and utility using a process that obtained approval from a commercial health EMR data provider. HealthGAN creates a GAN model in a secure environment, which is then exported to generate synthetic data outside the secure environment Recreating the analysis of two studies show that the method produces high quality for health informatics education. New research models and algorithms can be created on the synthetic data to create and evaluate new approaches and evaluate without compromising patient privacy.

Synthetic data generation provides an approach to make private data assets public. We recommend that empirical metrics for assessing privacy, resemblance and utility be utilized whenever synthetic data are generated, even if the underlying algorithms have theoretical guarantees of privacy. If synthetic data generated by automatic machine learning methods became a routine part of the publication process, scientific discovery and reproducibility would be accelerated and improved. Approaches could be developed on the synthetic data, and then evaluation on the real data in the secure environment. Since HealthGAN is designed to duplicate the underlying multivariate distributions while preserving individual privacy, it may not not necessarily protect proprietary information represented in the data, such as business process and patterns.

References

1. Alzantot, M., Chakraborty, S., Srivastava, M.: Sensegen: A deep learning architecture for synthetic sensor data generation. In: 2017 IEEE International Conference

[7] Further analysis of the synthetic data for both studies is included in the supplemental material https://git.io/Jf3mK

on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 188–193. IEEE (2017)

2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
3. Begley, C.G., Ioannidis, J.P.: Reproducibility in science: improving the standard for basic and preclinical research. Circulation research **116**(1), 116–126 (2015)
4. Dash, S., Dutta, R., Guyon, I., Pavao, A., Yale, A., Bennett, K.P.: Synthetic event time series health data generation. arXiv preprint arXiv:1911.06411 (2019)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777 (2017)
6. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2315–2324 (2016)
7. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
8. Krishnan, P., Jawahar, C.: Generating synthetic data for text recognition. arXiv preprint arXiv:1608.04224 (2016)
9. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on. pp. 399–410. IEEE (2016)
10. Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., Mark, R.G.: Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Critical care medicine **39**(5), 952 (2011)
11. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint arXiv:1806.01246 (2018)
12. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18. IEEE (2017)
13. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing (2019)
14. Vargason, T., Frye, R.E., McGuinness, D.L., Hahn, J.: Clustering of co-occurring conditions in autism spectrum disorder during early childhood: A retrospective analysis of medical claims data. Autism Research **12**(8), 1272–1285 (2019)
15. Vargason, T., McGuinness, D.L., Hahn, J.: Gastrointestinal symptoms and oral antibiotic use in children with autism spectrum disorder: Retrospective analysis of a privately insured us population. Journal of autism and developmental disorders pp. 1–13 (2018)
16. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.P.: Assessing privacy and quality of synthetic health data. In: Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse. pp. 1–4 (2019)
17. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.P.: Privacy preserving synthetic health data. In: Proceedings of the 27. European Symposium on Artificial Neural Networks ESANN. pp. 465–470 (2019)
18. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.P.: Generation and evaluation of privacy preserving synthetic health data. Neurocomputing (April 2020)