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'.

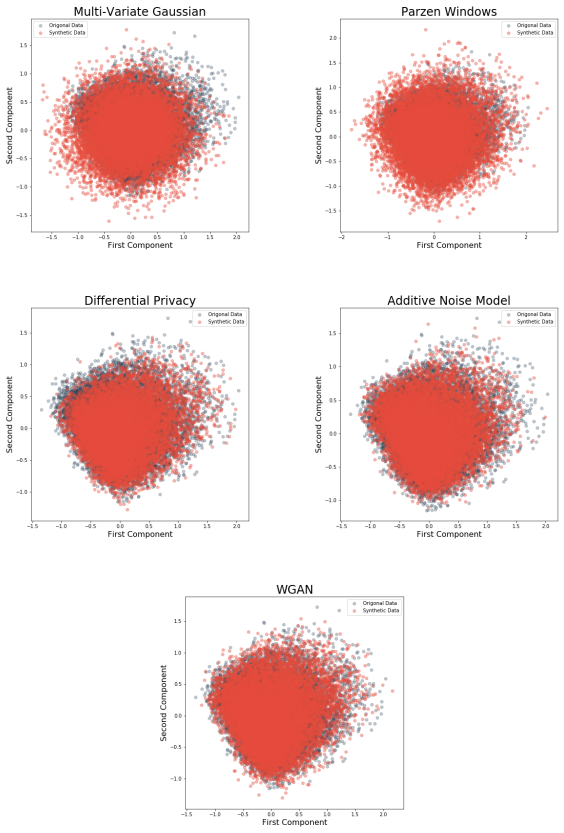

Fig. 4: Comparison of generative methods using PCA projection created using the real data