Additional plots are below.

- Team Sage

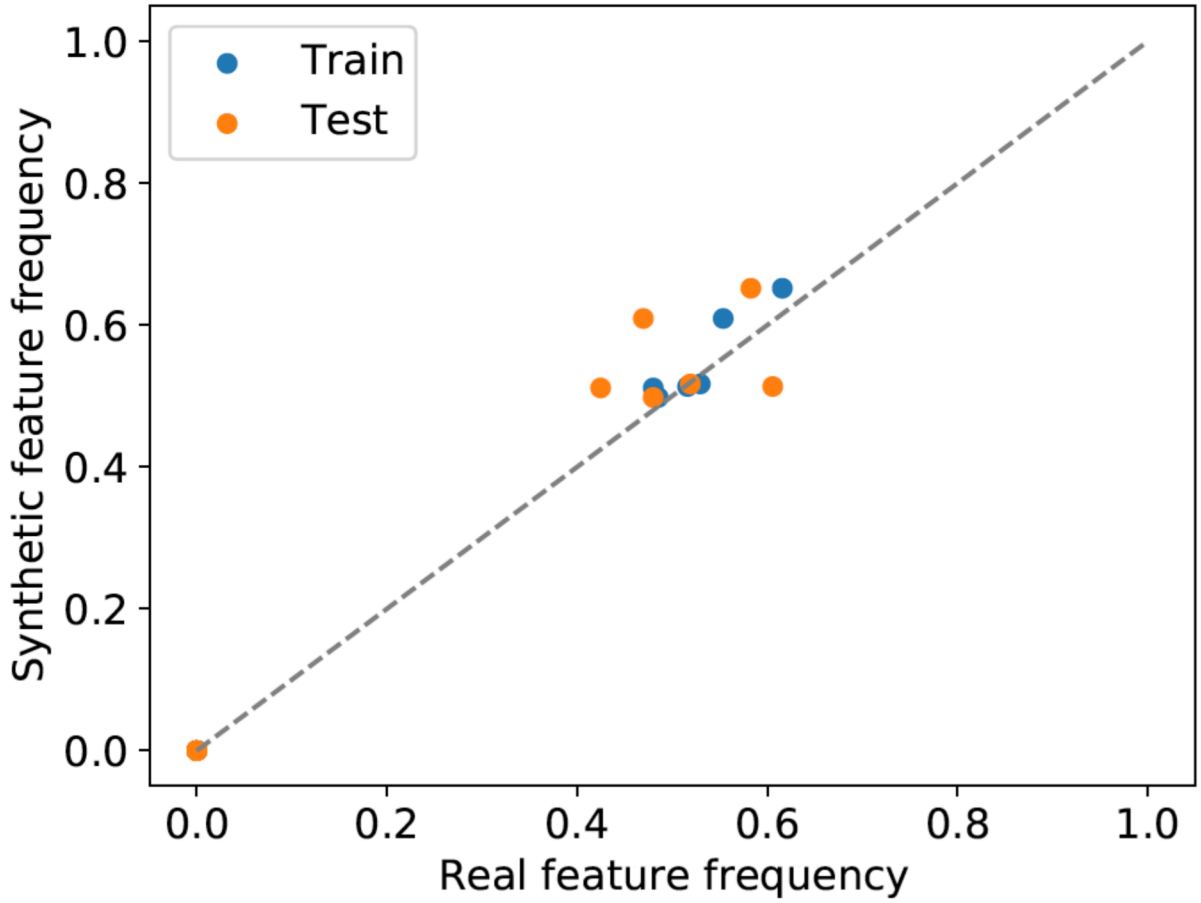

Figure 1: Correlation between respective features in the real and synthetic datasets. Train and test refer to subsets of the real data used to train the synthetic data generator and a separate held-out test set not used to produce the synthetic data.

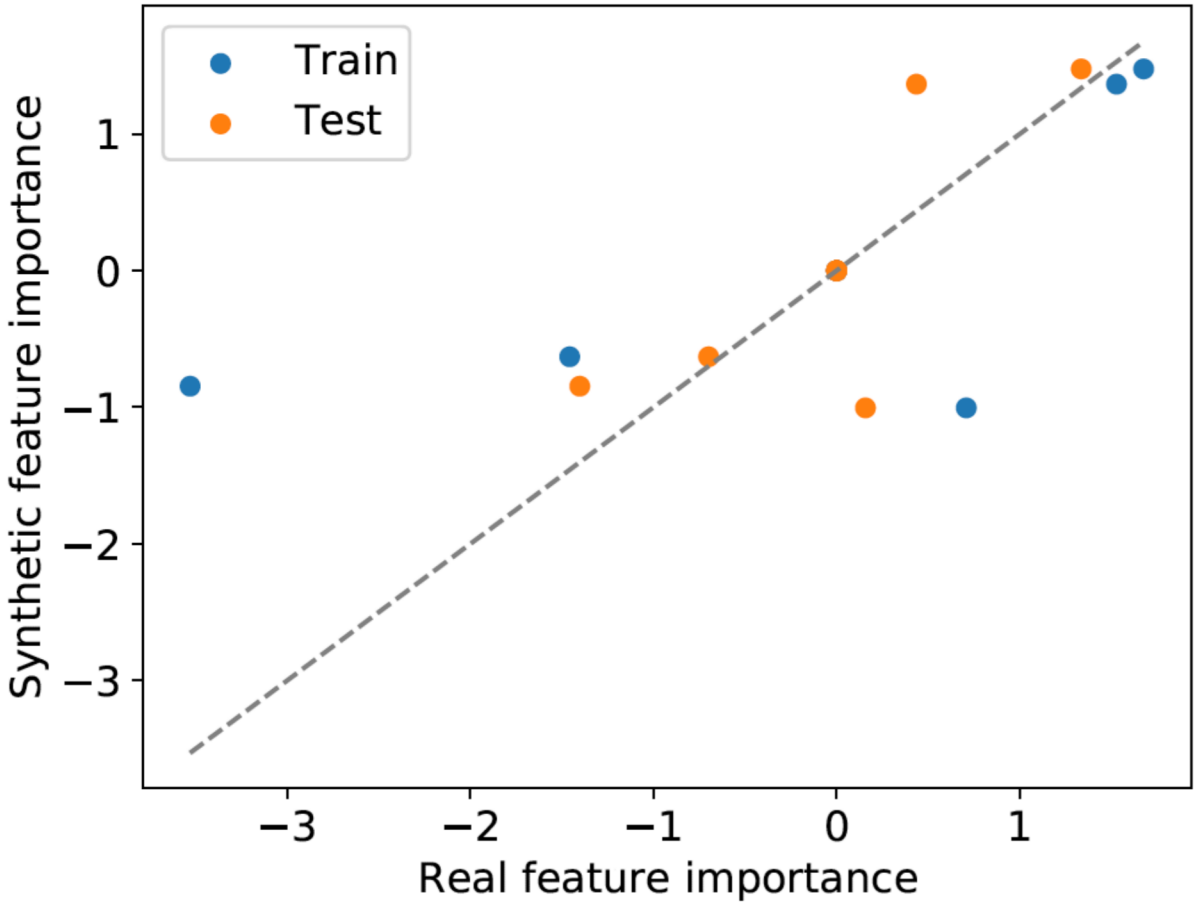

Figure 2: Correlation between real and synthetic feature importance after training a logistic regression model. Train and test refer to subsets of the real data used to train the synthetic data generator and a separate held-out test set not used to produce the synthetic data. Each subset of real data and the synthetic dataset was used to train a separate logistic regression model.

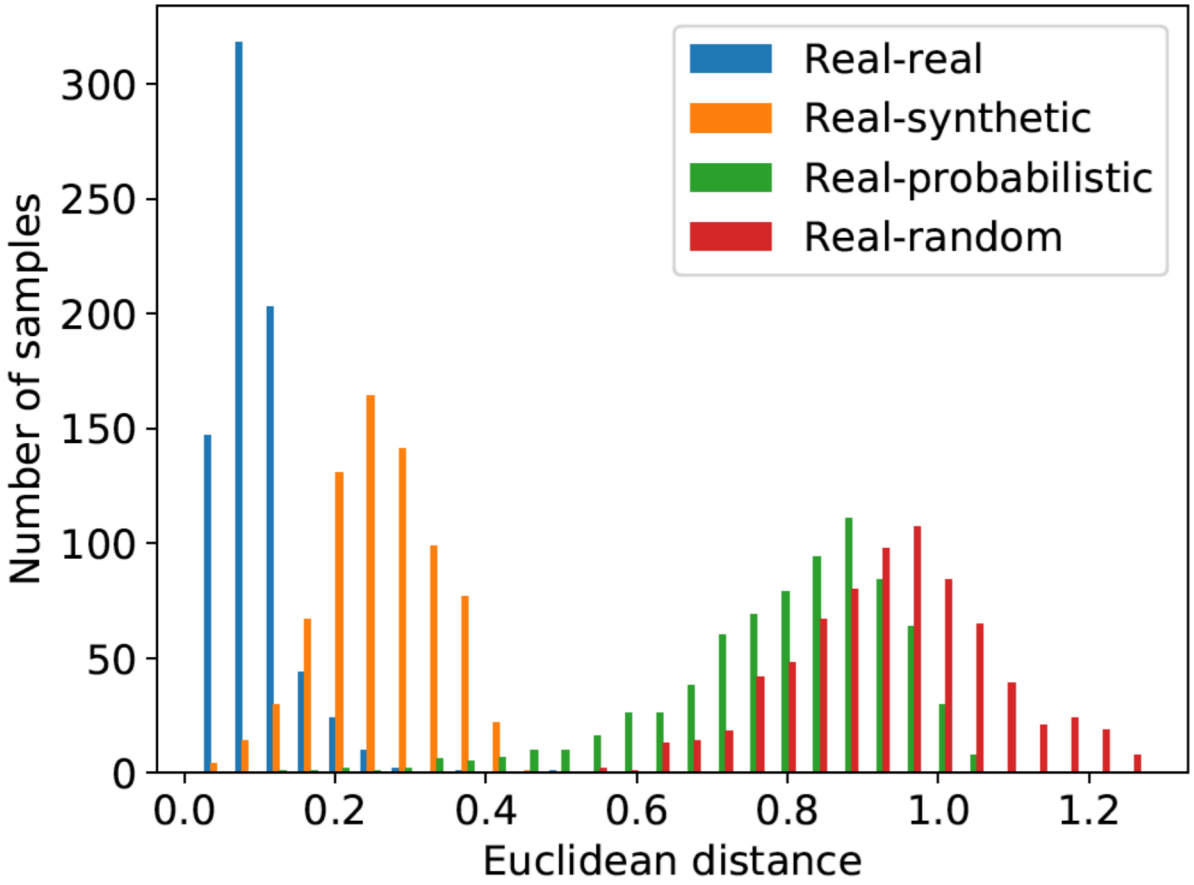

Figure 4: Distribution of pairwise distances between pairs of samples between datasets. The synthetic data is more different from the real dataset than the real dataset to itself but closer to the real dataset than randomly sampled data.

- Real-real: distance between randomly selected pairs of real samples
- Real-synthetic: distance between pairs of real and synthetic samples
- Real-probabilistic: distance between a real sample and sampled binary vector where each column is sampled from a binomial where the frequency equals that in the real training set
- Real-random: distance between a real sample and a randomly sampled binary vector

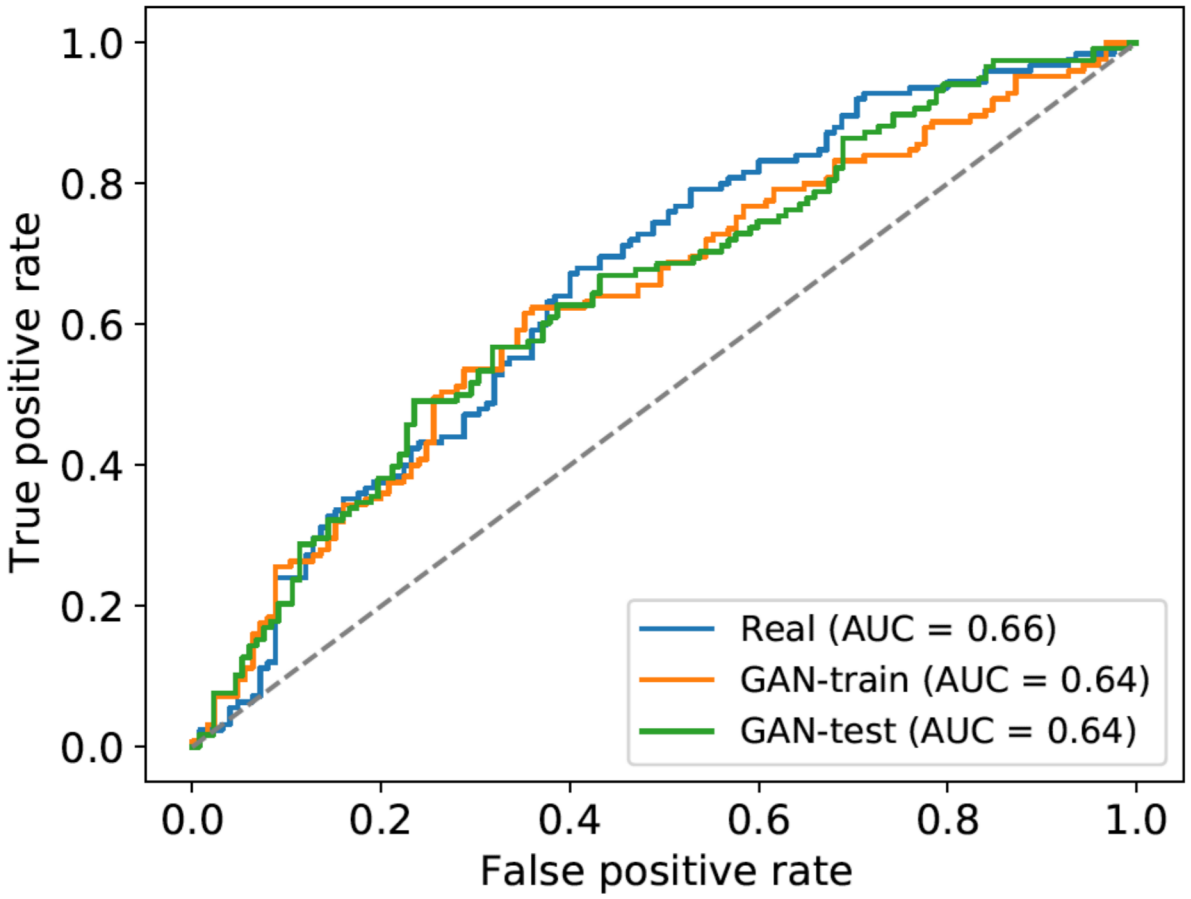

Figure 4: Using the synthetic dataset to train or test a logistic regression model and comparing performance with training and testing on real data. Performance is similar, indicating the ability of synthetic data to both train and assess a predictive model.

- Real: use real dataset to train predictive model and test on a separate real dataset
- GAN-train: use synthetic dataset to train predictive model and test on a real dataset
- GAN-test: use real dataset to train predictive model and test on the synthetic dataset

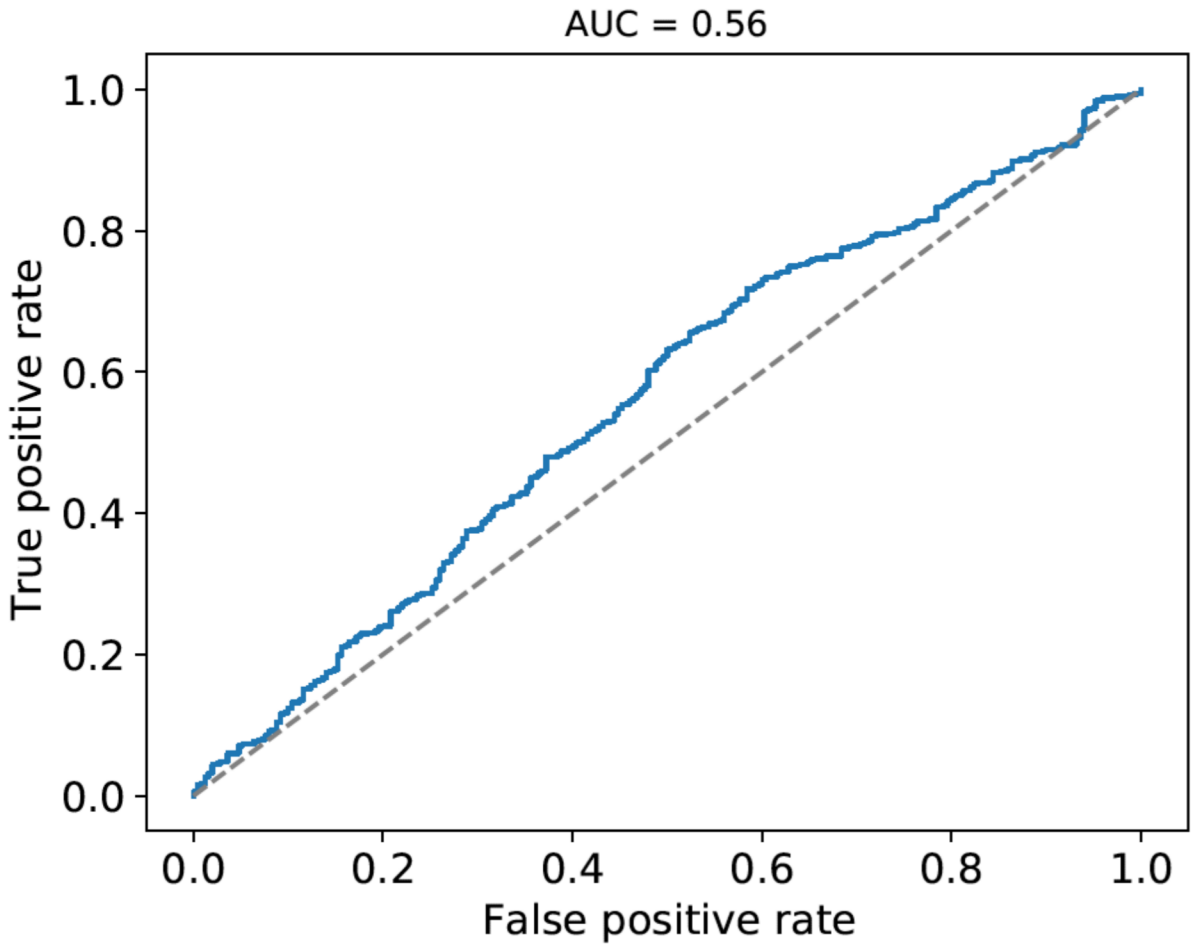

Figure 5: Membership inference risk assessment. Performance of a logistic regression classifier trained on the synthetic dataset to differentiate between real data used to train the classifier and a held out test set. Performance is, reassuringly, low.