Evaluation of Synthetic Data

An Assessment of Quality and Privacy in Synthetic Data

Andreas Persson, Data Scientist (AI Sweden)

Table of Contents

Background

In this appendix, we present all the results regarding the quality and the privacy of all synthetic data generated by all the models evaluated in this study. For each evaluated model, ten individual datasets were generated (each dataset with the same size of the population as the original dataset). The results presented in the in this report are the average result for each evaluated model (and for each of the original datasets used in this study, i.e., dataset "discharged" and dataset "recovered"). The results, presented in this appendix, are for the following evaluated models:

- **CTGAN:** the official implementation of Conditional Tabular Generative Adversarial Network[1], which is publicly available through the Synthetic Data Vault (SDV)[2].
- **TVAE:** the official implementation of Tabular Variational Auto Encoder[1]. Presented together with the CTGAN model (for comparison) and which is, likewise, publicly available through SDV.
- **PD-CTGAN:** an extension of the CTGAN model that utilizes Differential Privacy[3] to improve individual privacy. The implementation of this model is

1 Xu, Lei , et al. “Modeling Tabular data using Conditional GAN.” Advances in Neural Information Processing Systems (NeurIPS) (2019).

2 https://sdv.dev/

publicly available through the SmartNoise library as part of Open Differential Privacy[4].

- **PATE-CTGAN:** another extension of the CTGAN model that, instead, uses Private Aggregation of Teacher Ensembles[5] to improve privacy. This implementation is, likewise, available through the SmartNoise library.
- **Syndata – Model 1:** the initial model from Syndata. This model is based on the same CTGAN implementation available through SDV.
- **Syndata – Model 2:** the second model from Syndata. This model is, fundamentally, a probabilistic model based on Gaussian Copula functions[6], which is also available trough SDV.
- **Syndata – Model 3:** the third model from Syndata. This model is an improvement of the initial model from Syndata. Hence, this model is yet another CTGAN based model.

Quality measures for dataset:

	"discharged"		"recovered"	
Model	**CS Test**	**KS Test**	**CS Test**	**KS Test**
CTGAN	0.9977	0.8836	0.9937	0.8889
TVAE	0.9969	0.8842	0.9933	0.8886
DP-CTGAN	0.9224	0.6323	0.9172	0.6564
PATE-CTGAN	0.9548	0.7575	0.9529	0.7684
Syndata - Model 1	0.9937	0.8257	0.9500	0.8446
Syndata - Model 2	0.9815	0.8920	0.9708	0.8911
Syndata - Model 3	0.9711	0.7837	-	-

Table 1: Quality measures estimated by the Chi-Squared distribution test (CS) for categorical attributes and the Kolmogorov-Smirnov distribution test (KS) for numerical attributes, respectively. Each distribution test gives a combined percentage estimation of how the distributions of the synthetic data correlates to the corresponding distributions of the original data.

3 Abadi, Martin, et al. "Deep Learning with Differential Privacy." Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016).

4 https://opendp.org/

5 Papernot, Nicolas, et al. “Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data.” International Conference on Learning Representations (ICLR) (2017).

6 Schmidt, Thorsten. "Coping with copulas." Copulas-From theory to application in finance 3 (2007): 34.

The results, seen in Table 1, indicate excellent quality regarding categorical attributes. For categorical attributes, the overall best distributions can be found in the synthetic data generated by the CTGAN model, which have distributions that correlate to 99.77 % for the original dataset "discharged" and 99.37 % for the original dataset "recovered", respectively. However, the quality of numerical attributes is not as prominent as that of categorical attributes. Nonetheless, Syndata - Model 2 indicates the overall best distributions for numerical attributes with distributions that correlate to 89.20% for dataset "discharged" and 89.11% for dataset "recovered", respectively. It should also be noted, there is only a marginal difference in the quality of the synthetic data generated by the CTGAN model, TVAE model, and Syndata - Model 2.

Summary: the results indicates that the distribution for categorical attributes are best preserved in synthetic data generated by the CTGAN model, while the distribution for numerical attributes are best preserved in synthetic data generated by Syndata - Model 2.

For a further discussion regarding distributions and a detailed evaluation of Syndata - Model 1, 2 & 3, see Appendix 1.

Privacy measures for dataset:

	"discharged"		"recovered"	
Model	**Gen. CAP**	**RF**	**Gen. CAP**	**RF**
CTGAN	0.9931	0.9924	0.9921	0.9909
TVAE	0.9928	0.9918	0.9930	0.9921
DP-CTGAN	0.9956	0.9964	0.9963	0.9970
PATE-CTGAN	0.9944	0.9943	0.9939	0.9945
Syndata - Model 1	0.9930	0.9937	0.9944	0.9948
Syndata - Model 2	0.9929	0.9922	0.9928	0.9937
Syndata - Model 3	0.9945	0.9956	-	-

Table 2: Privacy measures estimated by categorical Generalized Correct Attribute Probability (Gen. CAP) and categorical Random Forest (RF). As implied, both methods assume that privacy is estimated based on categorical attributes. Both methods are, reversely, based on the synthetic dataset, while the original dataset is, subsequently, used for validating each method and thereby estimating a measure of privacy.

The methods for privacy measures assume that both key attributes (i.e., attributes that can identify an individual), as well as sensitive attributes (i.e., attributes that can infer the key attributes), are initially identified. The privacy is, subsequently, estimated based on the sensitive attributes and with respect to the key attributes. The original datasets used in this study were already anonymized to the extent that the only personal data available was age and gender. Age and gender were, therefore, used as key attributes, while categorical attributes consisting of medical notes were used as sensitive attributes (e.g., notes based on the Richmond Agitation-Sedation Scale, notes about transfers within the hospital, etc.).

The results, seen in Table 2, indicate an overall high level of privacy with a measure above 99% privacy for all evaluated models. However, as expected, the best privacy measures were found in the synthetic dataset generated by the DP-CTGAN model, i.e., a CTGAN model trained with extra emphasis on differential privacy. In a worst-case scenario, the synthetic data generated by the CTGAN model have a privacy measure of 99.56% for dataset "discharged" and 99.63% for dataset "recovered" (for privacy estimated by Gen. CAP, which can be considered as "brute force" search for similarity between the synthetic dataset and the original dataset). As privacy is measured as the probability for false prediction during validation based on the original data, the results can, reversely, be interpreted as 0.44% and 0.37% probability (for dataset "discharged" and dataset "recovered", respectively) to correctly identify an individual patient in the original data based on the synthetic data.

Based on the results (presented in Table 2), it is also evident that there is not a tremendous difference in privacy measures for the DP-CTGAN model, Syndata – Model 1, and Syndata – Model 3. By comparing with the results of quality measures (seen in Table 1), it is further apparent that there is a correlation between quality and privacy, i.e., synthetic data with higher quality have, in general, lower privacy measures, and vice versa. However, it should also be noted, the gain in privacy for synthetic data with low quality is considerably lower than the gain in quality for synthetic data with low privacy.

Summary: as expected, the best measure of privacy was found in the synthetic data generated by the privacy-preserving DP-CTGAN model. However, all the evaluated models indicated a generally high level of privacy.

Kernel density estimates of personal data for dataset:

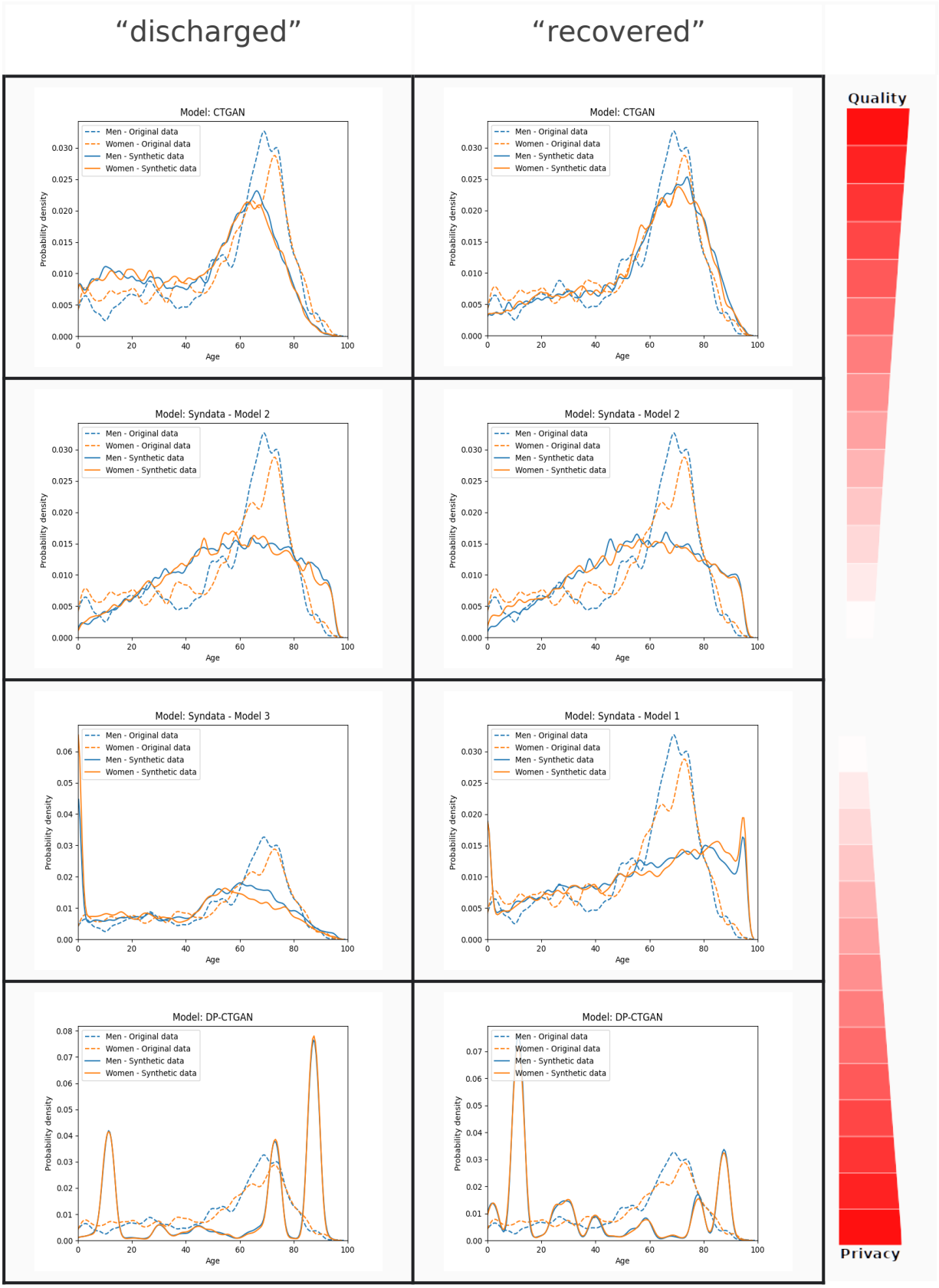

Table 3: Examples of Kernel Density Estimates (KDE) of personal data for synthetic datasets compared to the original dataset. From top to bottom: a declining level of quality in the datasets that are exemplified. From bottom to top: reversely, a declining level of privacy in the datasets that are exemplified.

For a further illustration of the correlation between quality and privacy in synthetic data, consider the graphs in Table 3. Each figure represent the probability density of the age for men and women in the synthetic data compared to the probability density of the age for men and women in the original data. Viewing the figures from top to bottom, there is a declining level of quality in the synthetic datasets exemplified in Table 3. Reversely, from bottom to top, there is, instead, a declining leave of privacy in the synthetic datasets.

Based on the graphs, presented in Table 3, it is seen that synthetic data of high quality, e.g., synthetic data generated by the CTGAN model and Syndata - Model 2, also have a probability kernel that generally follows the probability kernel of the original data. Opposite, it is also seen that synthetic data of high privacy, e.g., synthetic data generated by the DP-CTGAN model and Syndata - Model 1 & 3, have unexpected peaks in probability kernel, which does not correlate with the probability kernel of the original data.

Summary: synthetic data of high quality have a probability kernel that generally follows the probability kernel of the original data. This means that synthetic data of high quality have a population with attribute values that are similar to that of the original data. It is, therefore, also a higher probability that there are individuals in the synthetic data with almost the same characteristic as individuals in the original data, which might result in a compromised privacy.