Synthetic Data Evaluation Report

This report details the quality, privacy and utility evaluation metrics gained from the synthetic data, and visualisations to help interpret them.

Metrics Summary

Privacy Metrics	Score	Quality Metrics	Score	Utility Metrics	Score
Exact Matches	1.0	Boundary Adherence	1.0	Statistic Similarity	0.89
Detection	0.0	Coverage	0.94	Correlation	0.78
Inference Protection	0.17	Complement	0.72	ML Efficacy	0.18
Singling Risk	0.17	nan	nan	nan	nan
Linkability Risk	0.06	nan	nan	nan	nan
Inference Risk	0.18	nan	nan	nan	nan

Synthetic Data Categorisation Level: Correlated Synthetic Data

Correlated Synthetic Data is categorised as the highest risk due to capturing information about relationships and patterns between variables. Therefore, the privacy metrics should be evaluated carefully to ensure individuals arent at risk of being identifiable.

What are the Synthetic Data Categorisation Levels?

Synthetic data exists on a spectrum of low to high fidelity depending on how similar it is to the real data. We can categorise these levels of fidelity as below:

Level 0 (Random Synthetic Data)
Values which are randomly generated, without using the real data or metadata.

Level 1 (Metadata Synthetic Data)
Random data which is generated based on real metadata such as data types, value ranges etc.

Level 2 (Structural Synthetic Data)
Generated by a synthesiser to mimic the structure of the real data, but without capturing the relationships between attributes.

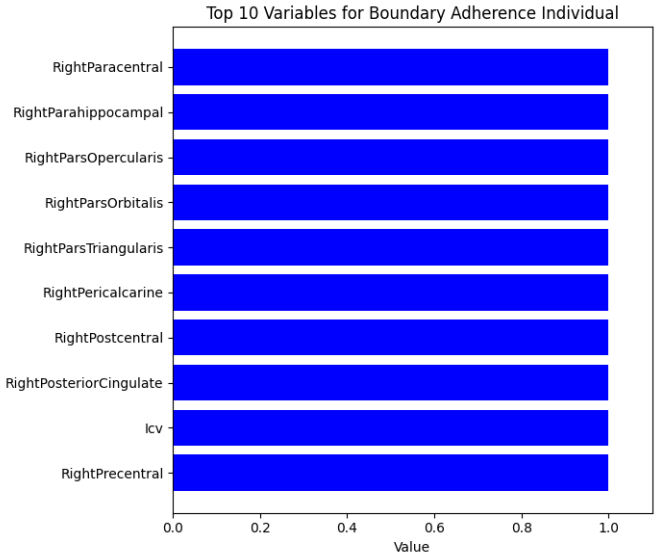
Level 3 (Statistical Synthetic Data)
Roughly keeps the same summary statistics such as mean, median, standard deviation etc.

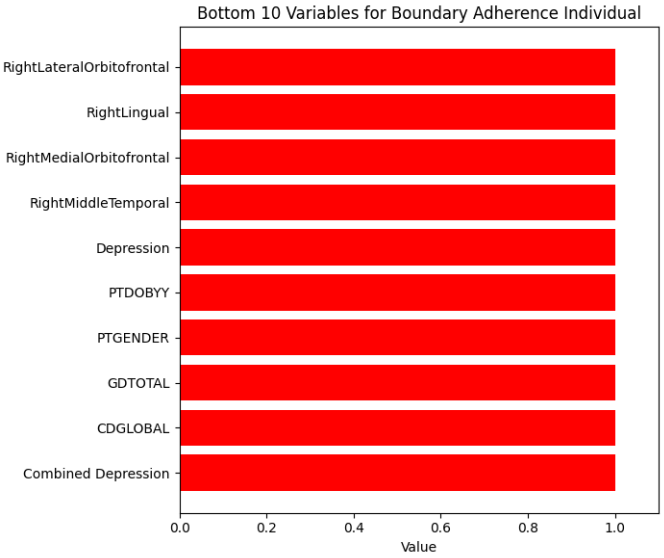
Level 4 (Correlated Synthetic Data)
Relationships and correlations between variables are captured.

Level 5 (Augmented Data)
Real data is adjusted or nearest neighbours is used to create highly realistic data points which are just slightly different to the real data.

Boundary Adherence Scores

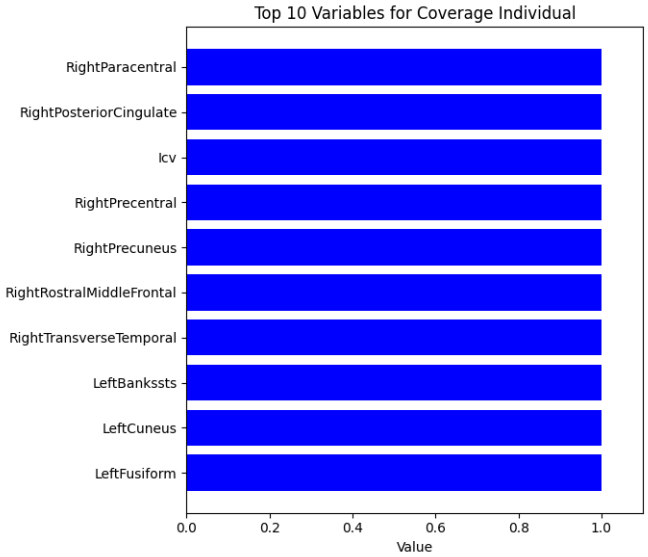
Boundary adherence measures whether values stay within the original min/max ranges of the data. (0.0: means none of the attributes have the same min/max ranges, 1.0: means all attributes have the same min/max ranges)

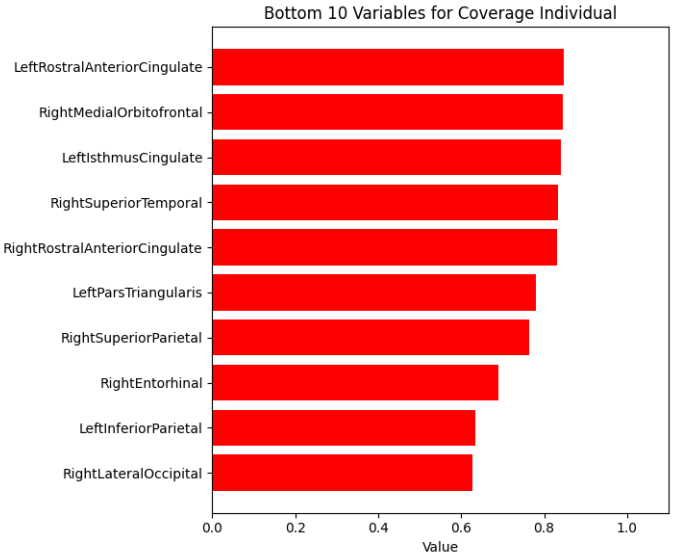




Coverage Scores

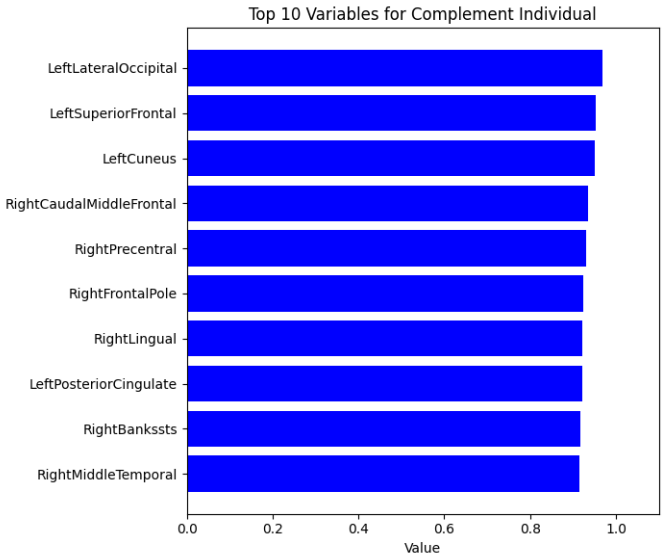
Coverage measures whether the whole range of values are represented. (0.0: means none of the values are represented, 1.0: means all values are represented)

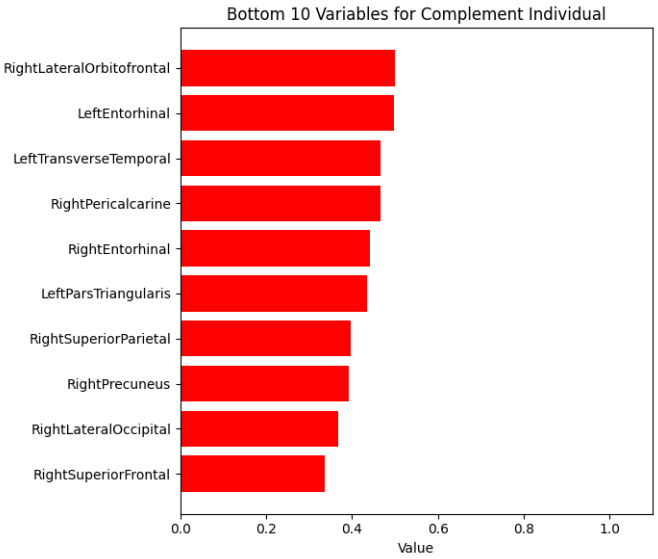




Complement Scores

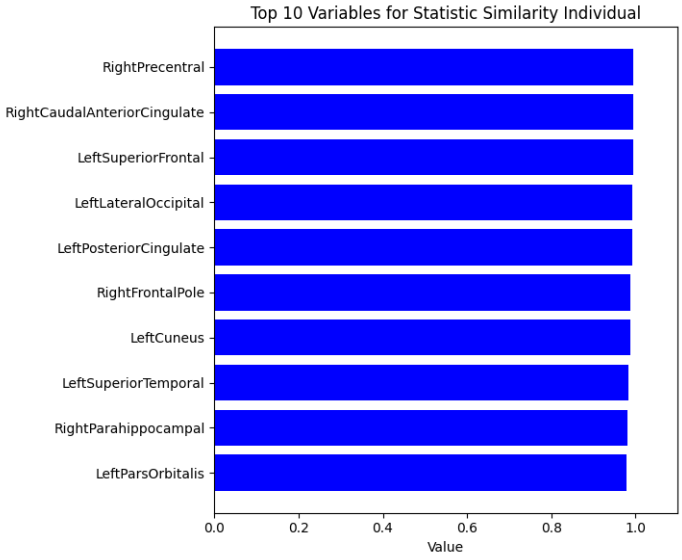
Complement measures whether the distributions look the same. (0.0: means the distributions are as different as they can be, 1.0: means the distributions are exactly the same)

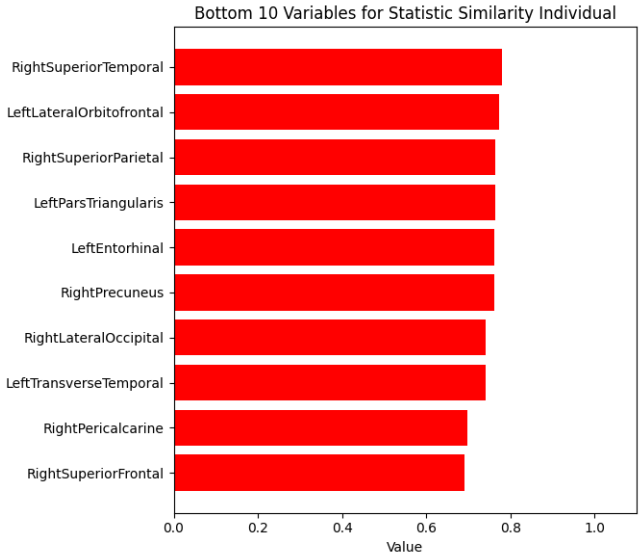




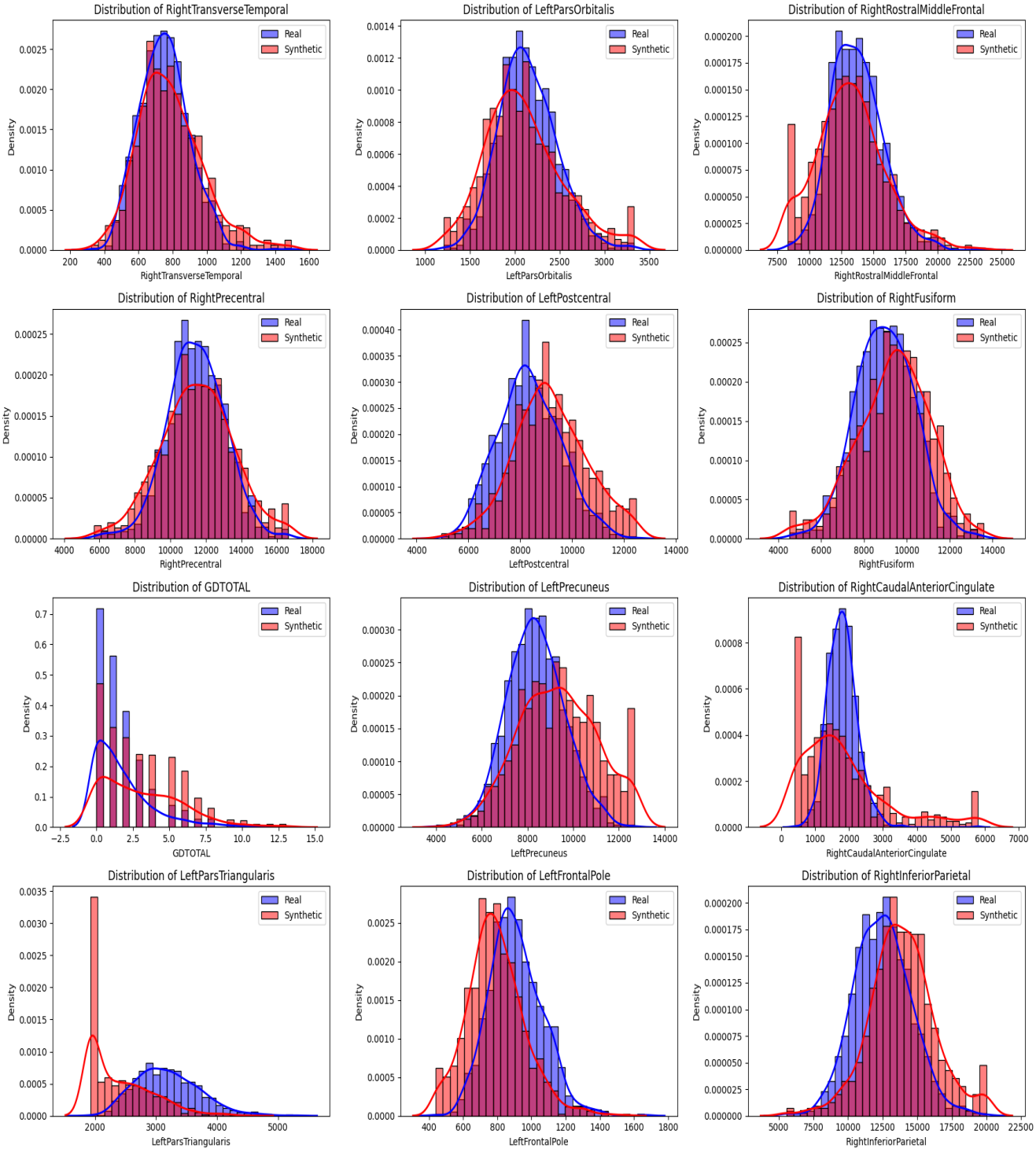
Similarity Scores

Statistic similarity measures how similar the summary statistics are such as mean and standard deviation. (0.0: means the summary statistics are extremely different to each other, 1.0: means the summary statistics are exactly the same)

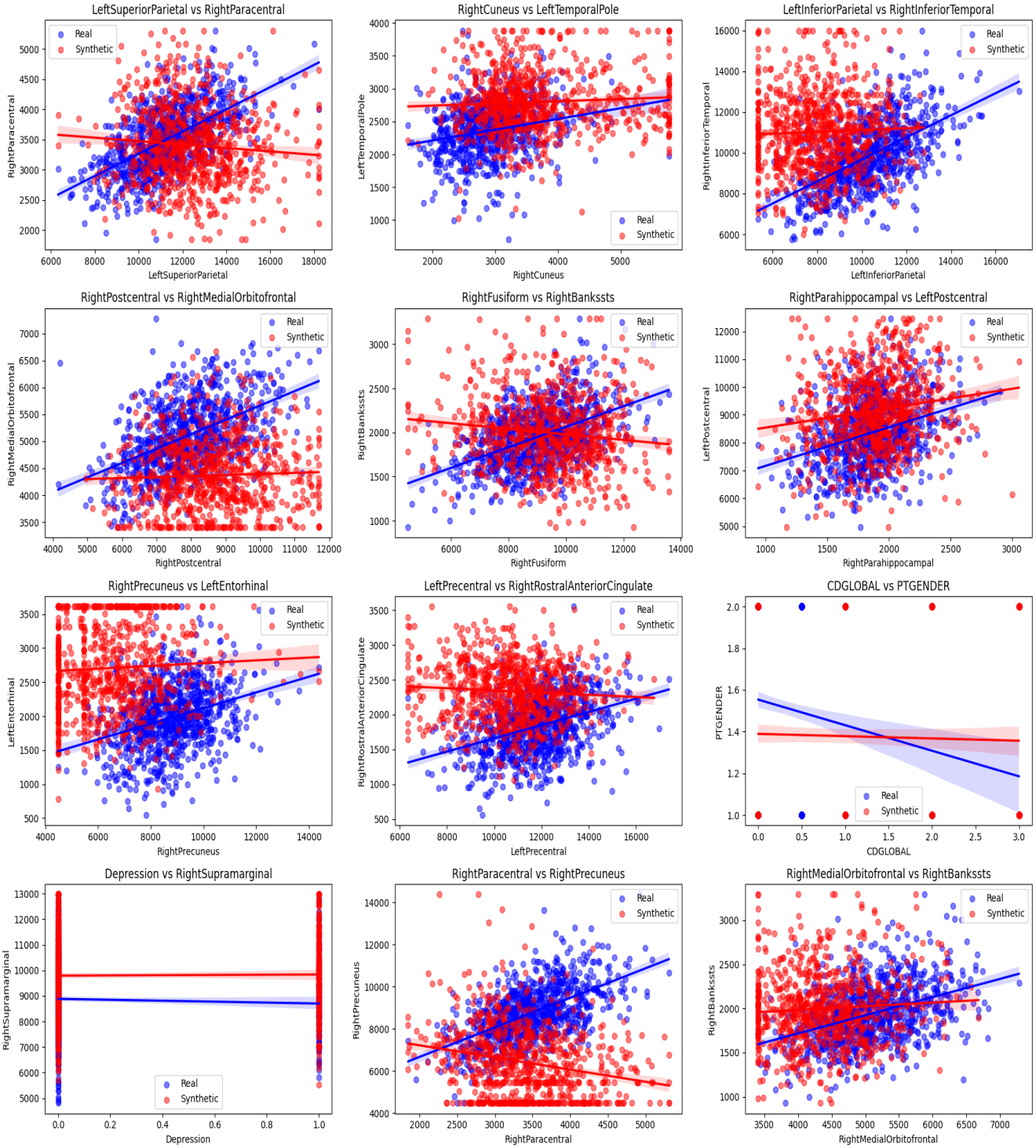




Example Distribution Comparisons



Example Correlation Comparisons



Meaning of Metrics

(Privacy) Exact Matches

This metric measures whether each row in the synthetic data is novel, or whether it exactly matches an original row in the real data.

(best) 1.0: The rows in the synthetic data are all new. There are no matches with the real data.
(worst) 0.0: All the rows in the synthetic data are copies of rows in the real data.

(Privacy) Detection

This metric calculate how difficult it is to tell apart the real data from the synthetic data using machine learning techniques.A score of 1 may indicate high quality but it could also be a clue that the synthetic data is leaking privacy (for example, if the synthetic data is copying the rows in the real data).

(worst) 1.0: The machine learning model cannot identify the synthetic data apart from the real data.
(best) 0.0: The machine learning model can perfectly identify synthetic data apart from the real data.

(Privacy) Inference Protection

This metric calculates the risk of an attacker being able to infer real, sensitive values. It is assumed that an attacker already possess a few columns of real data; they will combine it with the synthetic data to make educated guesses.

(best) 1.0: The real data is 100% safe from the attack. The attacker is not able to correctly guess any of the sensitive values by applying the chosen attack algorithm.
(worst) 0.0: The real data is not at all safe from the attack. The attacker is able to correctly guess every sensitive value by applying the chosen attack algorithm.

(Privacy) Singling Out Risk

This metric measures how much the synthetic data can help an attacker finding a combination of attributes that single out records in the training data. This attack evaluates the robustness of the synthetic data to finding unique values of some attribute which single out an individual.

(worst) 1.0: There is a high risk that an individual can be singled out by a unique combination of their attributes.
(best) 0.0: There is a low risk that an individual can be singled out by a unique combination of their attributes.

(Privacy) Linkability Risk

This metric measures how much the synthetic data will help an adversary who tries to link two other datasets based on a subset of attributes. For example, suppose that the adversary finds dataset A containing, among other fields, information about the profession and education of people, and dataset B containing some demographic and health related information. Can the attacker use the synthetic dataset to link these two datasets? (It is assumed the attacker knows the key fields of the individual and that they are split across different datasets.)

(worst) 1.0: There is a high risk that attributes can be linked to identify an individual.
(best) 0.0: There is a low risk that attributes can be linked to identify an individual.

(Privacy) Inference Risk

This metric measures the inference risk. It does so by measuring the success of an attacker that tries to discover the value of some secret attribute for a set of target records on which some auxiliary knowledge is available. (Again, like the linkability risk, it assumes the attacker knows the key fields)

(worst) 1.0: There is a high risk that sensitive attributes can be inferred.
(best) 0.0: There is a low risk that sensitive attributes can be inferred.

(Quality) Boundary Adherence

This metric measures whether a synthetic column respects the minimum and maximum values of the real column. It returns the percentage of synthetic rows that adhere to the real boundaries.

(best) 1.0: All values in the synthetic data respect the min/max boundaries of the real data.
(worst) 0.0: No value in the synthetic data is in between the min and max value of the real data.

(Quality) Coverage

This metric measures whether a synthetic column covers the full range of values that are present in a real column.

(best) 1.0: The synthetic column covers the range of values present in the real column.
(worst) 0.0: The synthetic column does not overlap at all with the range of values in the real column.

(Quality) Complement

This metric computes the similarity of a real column vs. a synthetic column in terms of the column shapes -- aka the marginal distribution or 1D histogram of the column.

(best) 1.0: The synthetic distribution shapes are exactly the same to the real data.
(worst) 0.0: The synthetic distribution shapes are nothing like the real data.

(Utility) Statistic Similarity

This metric measures the similarity between a real column and a synthetic column by comparing a summary statistic (mean, median, standard deviation).

(best) 1.0: The summary statistics are exactly the same.
(worst) 0.0: The summary statistics are completely different.

(Utility) Correlation Similarity

This metric measures the correlation between a pair of numerical columns and computes the similarity between the real and synthetic data -- aka it compares the trends of 2D distributions.

(best) 1.0: The pairwise correlations of the real and synthetic data are exactly the same.
(worst) 0.0: The pairwise correlations are as different as they can possibly be.

(Utility) ML Efficacy

This metric calculates the success of using synthetic data to perform an ML prediction task.

(best) 1.0: Given the synthetic training data, you will be able to perform ML tasks with 100% accuracy on the real data
(worst) 0.0: Given the synthetic training data, you will not be able to predict any of the real data correctly.