# Impacts of Data Synthesis: A Metric for Quantifiable Data Standards and Performances

Gunjan Chandra

gunjan.chandra@oulu.fi
Biomimetics and Intelligent Systems Group, P.O. BOX 4500, FI-90014, University of Oulu, Oulu, Finland

## Introduction

Medical data analyses unfold a wide range of information that can improve overall quality of life by enhancing existing procedures for medical prognoses, diagnoses, and treatments. Such data contain sensitive information, and data anonymisation is a standard step to overcome the risk of disclosure. While data anonymisation has failed many times, not sharing data hinders innovative opportunities. This study evaluates the performance of tool *Synthpop*, which produces synthetic data, an alternative and secure approach toward data anonymisation.

## Methodology

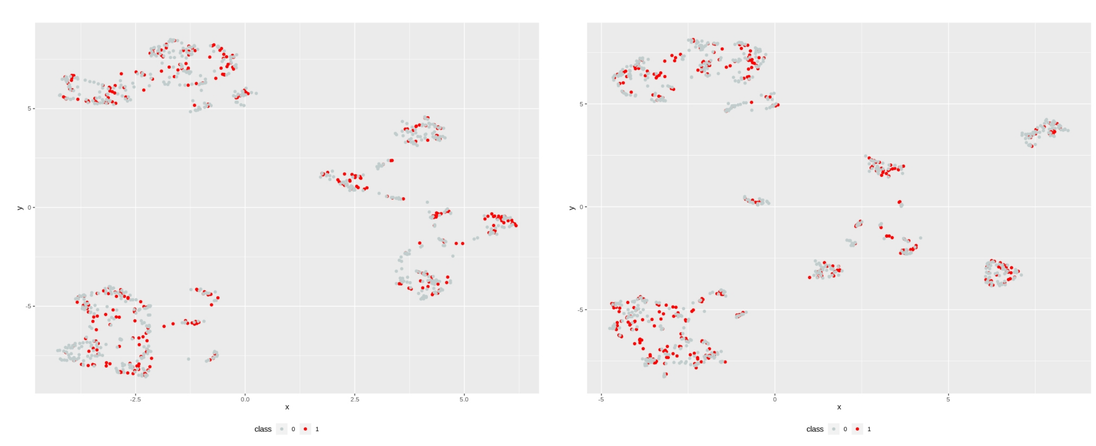
The study establishes data standards based on original data analyses and measures variations in the synthetic data to evaluate the performance of *Synthpop*. *Synthpop* replaces some or all observed values by sampling from an appropriate probability distribution, conditional on a variable to be synthesised, values from all previously synthesised columns of original data, and fitted parameters of a conditional distribution or posterior predictive distribution while retaining statistical properties of data and relationships between the variables. Synthetic data assessment can be divided into general utility, specific utility, and quality of information. The general utility considers whether synthetic data have overall similarities in the statistical properties and multivariate relationships with original data by analysing the correlation between data variables, visualisation, distributions, and similarity. The specific utility evaluates the similarity in performance of a fitted machine learning model on synthetic data to the original data. To estimate the quality of information, the concepts of information theory, such as evaluating change in entropy and estimating mutual information between variables, will help quantify the level of distortion and information loss caused by data synthesis. The experiments are based on two data sets: the Wisconsin Diagnostic Breast Cancer (WDBC) and the Type 1 Diabetes Prediction and Prevention (DIPP).

## Experiments and Results

Synthetic data must resemble all properties of original data with a statistically non-significant difference. Hypotheses will be as follows: Let $D$ denote original data, and $S_i$ denote synthetic data, where $i$ indicates index for synthetic data produced with the different synthesising methods. Let $t$ denote a vector of tests returning a statistic, and $C^*$ be a comparison function producing a $p-value$. Finally, comparing output of $C^*$ with $\alpha$, a threshold value for the level of significance, set to 0.05.

$$H_o : C^*\{t(D), t(S_i)\} \geqslant \alpha, \qquad \text{for all } t \in [0, \tau]$$
$$H_a : C^*\{t(D), t(S_i)\} < \alpha, \qquad \text{for any } t \in [0, \tau]$$

The original data sets were synthesised numerous times using diverse methods of synthesis. Figure 1 and 2, and Table 1 reveal that the synthetic data showed no signs of variation in data utility and succeeded at all performed tests with statistically non-significant differences from the original data. Furthermore, the complexity of the data from the information quality perspective is also preserved.




(a) Original (b) Synthetic

Figure 1:Uniform Manifold Approximation and Projection of DIPP data showing similar local and global structures between original and synthetic data.

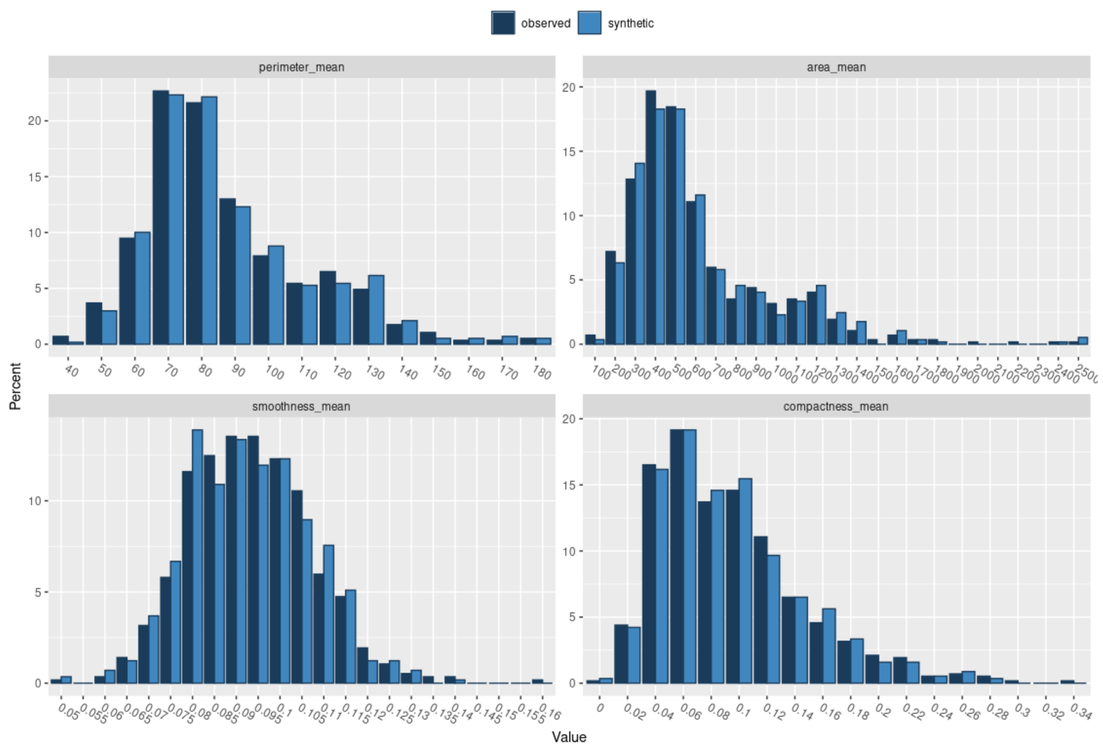



Figure 2:Relative frequency distribution of a few original (dark blue) and synthetic (light blue) WDBC data variables showing similar data distribution.

Table 1:Original and synthetic (SynW1 and SynW2) WDBC data and their performance over Random Forest models

| Test set | Model | Confusion Matrix | | | Evaluations Parameters | | |
|---|---|---|---|---|---|---|---|
| | | | Predicted labels | | F1 score | Area Under ROI | Accuracy |
| | | | Negative | Positive | | | |
| **Original** | **Original** | Negative | 113 | 2 | 0.98 | 0.96 | 0.97 |
| | | Positive | 3 | 53 | 0.95 | | |
| **SynW1** | **SynW1** | Negative | 111 | 2 | 0.98 | 0.97 | 0.98 |
| | | Positive | 2 | 56 | 0.97 | | |
| **Original** | **SynW2** | Negative | 102 | 5 | 0.94 | 0.92 | 0.93 |
| | | Positive | 7 | 57 | 0.90 | | |

## Conclusion

The article was inspired by the benefits of open healthcare databases and aimed to solve perpetual hindrances in data sharing caused by the risk of disclosure from shortcomings of current data anonymisation. Synthpop fulfils all the necessities for data sharing and hence allows various opportunities in the research community, including easy data sharing, more significant collaborations, and information protection. In conclusion, the tool performed remarkably and exceeded the expectations of its intended purpose.