Article published online: 2023-01-09

Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions

Mikel Hernadez[1] Gorka Epelde[1,2] Ane Alberdi[3] Rodrigo Cilla[1] Debbie Rankin[4]

[1] Digital Health and Biomedical Technologies Department, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastian, Spain
[2] eHealth Group, Biodonostia Health Research Institute, Donostia-San Sebastian, Spain
[3] Biomedical Engineering Department, Mondragon Unibertsitatea, Arrasate-Mondragón, Spain
[4] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry-Londonderry, United Kingdom

Address for correspondence Mikel Hernandez, PhD Student, Digital Health and Biomedical Technologies Department, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastian, Spain (e-mail: mhernandez@vicomtech.org).

Methods Inf Med 2023;62:e19–e38.

Abstract

Background Synthetic tabular data generation is a potentially valuable technology with great promise for data augmentation and privacy preservation. However, prior to adoption, an empirical assessment of generated synthetic tabular data is required across dimensions relevant to the target application to determine its efficacy. A lack of standardized and objective evaluation and benchmarking strategy for synthetic tabular data in the health domain has been found in the literature.

Objective The aim of this paper is to identify key dimensions, per dimension metrics, and methods for evaluating synthetic tabular data generated with different techniques and configurations for health domain application development and to provide a strategy to orchestrate them.

Methods Based on the literature, the resemblance, utility, and privacy dimensions have been prioritized, and a collection of metrics and methods for their evaluation are orchestrated into a complete evaluation pipeline. This way, a guided and comparative assessment of generated synthetic tabular data can be done, categorizing its quality into three categories ("*Excellent,*" "*Good,*" and "*Poor*"). Six health care-related datasets and four synthetic tabular data generation approaches have been chosen to conduct an analysis and evaluation to verify the utility of the proposed evaluation pipeline.

Results The synthetic tabular data generated with the four selected approaches has maintained resemblance, utility, and privacy for most datasets and synthetic tabular data generation approach combination. In several datasets, some approaches have outperformed others, while in other datasets, more than one approach has yielded the same performance.

Conclusion The results have shown that the proposed pipeline can effectively be used to evaluate and benchmark the synthetic tabular data generated by various synthetic tabular data generation approaches. Therefore, this pipeline can support the scientific community in selecting the most suitable synthetic tabular data generation approaches for their data and application of interest.

Keywords

- synthetic tabular data generation
- synthetic tabular data evaluation
- resemblance evaluation
- utility evaluation
- privacy evaluation

received
June 13, 2022
accepted after revision
October 29, 2022
article published online
January 9, 2023

DOI https://doi.org/10.1055/s-0042-1760247.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Introduction

Background

We are in a digital era where the amount of data generated is growing exponentially, leading to a paradigm shift from traditional and manual processes toward artificial intelligence (AI) applications in different contexts. However, many AI developments are being slowed down by data-protection laws and imbalanced data. Therefore, synthetic data generation (SDG) research has gained importance in recent years. Synthetic data (SD) was first proposed and defined by Rubin[1] and Little[2] in 1993 as datasets consisting of records of individual synthetic values instead of real values. Nowadays, the term SD has been extended and can be defined as artificial data generated by a model trained or built to replicate real data (RD) based on its distributions (i.e., shape and variance) and structure (i.e., correlations among the attributes).[3] As claimed by Hernandez et al,[4] Hu,[5] Reiter[6] and Taub et al,[7] fully SD does not contain data from the original real dataset and less information is lost compared to other anonymization techniques (i.e., Parzen Window,[8] Additional Noise Model,[8] Random Noise[9] and Independent Sampling[9]) or data balancing techniques (i.e., SMOTE,[10] ADASYN,[11] ROS[12]). For this reason, SDG has shown great promise for (1) augmenting RD by balancing datasets or supplementing the available data to train predictive models[13–21] and (2) preserving privacy to enable secure and private data sharing.[8,22–35] In the health domain, SDG has been researched for different types of data, such as biomedical signals,[14,27] medical images,[15–18] electronic health records (EHR) free-text content,[36] time-series smart-home and living labs activity data,[19,20,23,28,37] and tabular data from EHR.[8,22,29,30,33–35]

This study focused on synthetic tabular data generation (STDG), as it is the predominant type of data used to develop machine learning (ML) models to aid health care decision-making. Therefore, tabular health data potentially offers the most valuable opportunities to develop AI-based health care systems. Most of the research studies on STDG in the health domain have focused on proposing new STDG approaches[8,29,30,32–35,38] and evaluating and benchmarking different STDG approaches over various datasets.[39–44] There are also studies that have highlighted the potential value of STDG for secure data exchange that ensures patient data privacy. In the workflows proposed by Rankin et al[22] and Hernandez et al,[23] STDG techniques are incorporated into complete data processing workflows that can be used to accelerate research on ML model development for health care decision-making.

All previously referenced studies have demonstrated that although STDG is a potentially valuable technology for health care applications, prior to its adoption, an empirical assessment of the synthetic tabular data (STD) generated with various approaches is required across different dimensions. These dimensions include resemblance, utility, privacy, computational cost, veracity, diversity, and generalization. The resemblance dimension evaluates how well SD represents RD (covering aspects related to data distribution and correlations between attributes). The utility dimension evaluates the usability of statistical conclusions drawn from SD or the results from ML models trained with SD. Finally, the privacy dimension can measure how private SD is in terms of the disclosure risk of private or sensitive data. The results of such an assessment will support developers in selecting the ideal STDG approach for their particular application requirements.

The study developed by Dankar et al[41] classified STD evaluation metrics and methods into univariate fidelity, bivariate fidelity, population fidelity, and analysis-specific measures. However, similar to the studies developed by Hittmeir et al,[39] Giles et al,[40] and Alaa et al,[44] this classification only covers resemblance and utility dimensions. Hittmeir et al[42] also proposed and gathered different STD evaluation metrics and methods for the utility and privacy dimensions without including resemblance. In contrast, the evaluation proposed by Platzer and Reutterer[43] only covers resemblance and privacy evaluation.

Since our work aims to provide a collection of different metrics and methods for evaluating different STD generation techniques for targeted health domain applications, the focus has been placed on the resemblance, utility, and privacy dimensions. This manuscript's approach has not included other dimensions such as generation performance (in terms of time and required computation resources).

All abbreviations used throughout this manuscript are gathered in ►Appendix A.

Related Work

This section presents some existing STDG approaches for tabular health care data and the most common metrics and methods for evaluating the proposed STD dimensions.

STDG Approaches

To generate STD in the health care context, many approaches can be found in the literature. The simplest STDG approaches include Gaussian Multivariate (GM),[8,35] Bayesian Networks,[22,29,45] Categorical maximum entropy model,[46] and Movement-based kernel density estimation.[47] These approaches employ a statistical model to learn the multivariate distributions of the RD to sample a set of STD. They are typically used for small amounts of data and are not very scalable.

Due to the efficiency and popularity of generative models for SDG, specifically generative adversarial networks (GAN), in other areas and applications of health care, there is an interest in exploring whether they have the potential to generate high-quality STD for health care. GANs are composed of two neural networks (a generator and a discriminator) that learn to generate high-quality STD through an adversarial training process.[48] Several authors have used these generative models, presenting improvements, tuning hyperparameters, or adding new features.[8,9,13,21,24,25,33,34,38,49] While ehrGAN[13,21] and medGAN[8,9,13,33,34] were proposed to synthesize mainly numerical and binary data, Wasserstein GAN (WGAN) with Gradient Penalty,[8,13,33,34,38] healthGAN[8] and Conditional Tabular GAN (CTGAN)[24,25,49] synthesize numerical, binary, and categorical data efficiently.

Furthermore, different open-source and commercial packages for STDG have also been released. For example, the Synthetic Data Vault (SDV) is an ecosystem of STDG approaches for tabular data and time-series data composed of ensemble approaches that combine several probabilistic graphical modeling and Deep Learning (DL)-based techniques.[50,51] This ecosystem has been incorporated into the controlled data processing workflow proposed by Hernandez et al for secure data exchange.[23] Other open-source and commercial packages for STDG are SYNTHO,[52] the Medkit-Learning environment,[53] and YData.[54]

STD Evaluation Metrics and Methods for Resemblance, Utility, and Privacy Dimensions

The metrics and methods used to evaluate the resemblance, utility, and privacy of STD in the literature are diverse. Most studies related to STDG in the health care context evaluate the resemblance and utility dimensions, but only a few evaluate the privacy dimension. The most relevant metrics and methods for STD reported in the literature are presented below.

Resemblance Evaluation

The first step in resemblance evaluation is to analyze whether the distribution of STD attributes is equivalent to the distribution of the RD. Che et al,[21] Chin-Cheong et al,[38] and Bourou et al[25] compared the distributions of the attributes of RD against STD. Yang et al[13] compared the frequency of the attributes. Additionally, Choi et al,[9] Wang et al,[31] Abay et al,[45] Baowaly et al,[34] and Yale et al[8] compared the dimensional probability or probability distributions of RD and STD.

For distributions comparison, Yang et al[13] and Rashidian et al[32] analyzed the mean absolute error between the mean and standard deviation values of RD and STD. Some authors also use statistical tests to analyze the univariate resemblance of STD. Baowaly et al,[34] Bourou et al,[25] and Dankar et al[41] used Kolmogorov–Smirnov (KS) tests to compare distributions, Dash et al[37] applied Welsch *t*-tests, and Yoon et al[33] performed the Student *t*-test to compare mean values of the attributes. Yoon et al[23] and Bourou et al[25] used Chi-square tests to compare the independence of categorical attributes. In these studies, they analyzed the *p-values* obtained from the statistical tests to determine whether the null hypothesis is accepted to assure that the STD attributes preserve the properties of RD attributes analyzed with the statistical tests.

Additionally, other authors have used distance metrics to evaluate the resemblance of STD. Hittmeir et al[39] measured the distance between RD and SD, computing the nearest neighbors row-by-row.

In evaluating multivariate relationships, Rankin et al,[22] Yale et al,[8] Wang et al,[29] Bourou et al,[25] Hittmeir et al,[39] Dankar et al,[41] and Rashidian et al[32] visually compared the Pairwise Pearson Correlation (PPC) matrices to assess whether correlations between attributes of RD are maintained in STD. Additionally, principal component analysis transformation has been used by Yale et al[8] to compare the dimensional properties of STD and RD.

To analyze whether the semantics or significance of RD is maintained in STD, Choi et al,[9] Wang et al,[29] Beaulieu-Jones et al,[30] and Lee et al[55] asked some clinical experts to evaluate the STD qualitatively, scoring between 1 and 10. This score indicated how real the STD records appeared to them, where a score of 10 was most realistic. Another method that can be used if access to clinical experts is unavailable is to train some ML classifiers to label records as real or synthetic, as Lee et al[55] proposed in their study. Bourou et al[25] also proposed using ML classifiers to analyze how difficult it is to differentiate between SD and RD samples.

Utility Evaluation

The evaluation of the utility dimension has mainly been performed using STD in ML models by training and analyzing the performance of these models.

Train on Real Test on Real (TRTR) and Train on Synthetic Test on Real (TSTR) methods were used by Park et al,[56] Wang et al,[31] Beaulieu-Jones et al,[30] Chin-Cheong et al,[38] Baowaly et al,[34] Kotal et al,[24] Bourou et al,[25] Hittmeir et al,[39] Giles et al,[40] Dankar et al,[41] and Rashidian et al.[32] These authors trained ML models with RD and STD separately and then tested them with held-out RD not used for training. They use different classification metrics (e.g., Accuracy, F1-score, ROC, and AUC-ROC) to evaluate and analyze the differences in the models' performance when training the models with RD and STD.

On the other hand, Che et al,[21] Wang et al,[31] and Yang et al[13] augmented the training set of RD with STD. The authors compared the performance ML models trained only with RD and models trained with RD and STD combined.

Furthermore, in the study developed by Giles et al,[40] they analyzed the utility of the STD based on how well the feature importance from RD is represented in STD.

Privacy Evaluation

The few metrics and methods authors have used for privacy assessment of STD are based on distance and similarity metrics and re-identification risk evaluation.

Regarding distance and similarity-based metrics, Park et al[56] and Platzer and Reutterer[43] used distance to the closest record, computing the pairwise Euclidean distance between real and synthetic records where the closer the mean distance value is to 0, the more the privacy is preserved. According to Norgaard et al,[28] the maximum real to synthetic similarity value, computed by the cosine similarity, indicates whether the model has memorized and stored RD and is really generating data and not copying it. Other distance-based metrics used by Yoon et al[33] are the Jensen-Shannon divergence (JSD) and the Wasserstein distance. The authors used them to compute the balance between the identifiability and quality of STD.

Several disclosure attack simulations have been proposed in the literature to assess the re-identification of RD disclosure risk through STD. Choi et al,[9] Park et al,[56] Yale et al,[8] Hittmeir et al,[42] and Mendelevitch et al[57] simulated membership inference attacks to analyze the disclosure risk of a complete record in RD by computing distance metrics between RD and

STD records and using accuracy and precision metrics to quantify the membership risk. In contrast, Choi et al[9] and Mendelevitch and Lesh[57] additionally simulated attribute inference attacks to quantify the disclosure risk of some attributes of the dataset. Defining attributes considered quasi-identifiers (QID) and training some ML models with STD to predict the rest of the attributes, they analyze how accurately an attacker could predict some RD attributes if they obtained access to the STD.

Objectives

Published studies that evaluate STD generated using one or more STDG approaches use different metrics and methods to evaluate STD quality. Although some published studies propose or categorize STD evaluation metrics and methods into different groups, they have only focused on evaluating one or two of the previously defined dimensions. Furthermore, as concluded in the review by Ghosheh et al,[58] a universal and standardized evaluation methodology for developing reliable STDG approaches for STD has not yet been defined.

A complete, guided, and objective evaluation and benchmarking strategy for STD over the dimensions of resemblance, utility, and privacy does not exist in the literature. Thus, an organized pipeline or process is required to assess STD in these three dimensions to enable the selection of the best approach or approaches to generate the desired STD for individual use cases.

The aim of this paper is to identify key dimensions, per dimension metrics, and methods for evaluating STD generated with different techniques and configurations for health domain applications development and to provide a strategy to orchestrate them. We propose and assess a collection of standardized metrics and methods to evaluate the resemblance, utility, and privacy dimensions of STD for use in the health domain. The proposed metrics and methods were selected from the literature and orchestrated into a complete evaluation pipeline to enable a guided and comparable evaluation of STD. Our contributions can be summarized as follows:

1. We propose a set of metrics and methods to evaluate the resemblance, utility, and privacy dimensions of STD and present a meaningful orchestration of them. Different universal metrics and methods are suggested for independent use for each dimension. Although the metrics and methods proposed are not new, their orchestration in an organized way and the calculation of overall scores for each dimension are novel.
2. To the best of our knowledge, this work is the first attempt to propose and use a complete and universal STD evaluation pipeline covering the resemblance, utility, and privacy dimensions. The pipeline is generalizable to any kind of STD since the metrics and methods have been selected according to the most commonly used STD evaluation metrics and methods reported in the literature. Additionally, we present a methodology to categorize the performance of each STDG approach in each dimension as "*Excellent,*" "*Good,*" and "*Poor.*"
3. We analyze and evaluate the suggested pipeline using six different health care-related open-source datasets and four STDG approaches. The proposed evaluation metrics and methods are used to evaluate the quality of the STD generated for each dataset and STDG approach combination.
4. Based on the evaluation results, we benchmark the STDG approaches used to generate STD and discuss the veracity and efficiency of the proposed STD evaluation metrics and methods. We demonstrate that the presented pipeline can effectively be used to evaluate and benchmark different approaches for STDG, helping the scientific community select the most suitable approaches for their data and application of interest.

Methods

Synthetic Tabular Data Evaluation Metrics and Methods

The proposed metrics and methods for evaluating STD can be clustered into three dimensions: resemblance, utility, and privacy. Different metrics and methods from the literature have been selected and configured in an organized way within each dimension. The complete taxonomy of the selected methods, which can be used within the defined pipeline to evaluate STD generated with one STDG approach or compare the STD generated by different STDG approaches, is depicted in ►Fig. 1.

Resemblance Evaluation

In the resemblance dimension, the capacity of STD to represent RD is evaluated. Statistical, distribution, and interpretability characteristics are analyzed using four methods: univariate resemblance analysis (URA), multivariate relationships analysis (MRA), and data labeling analysis (DLA).

Two other metrics have been proposed for resemblance evaluation, i.e., visual analysis of attributes and dimensional reduction analysis (DRA). However, due to the difficulties associated with quantitatively analyzing these, they have been excluded from the final methodology and included in **Appendix B** and **Supplementary Material 1**, available in online version only.

Univariate Resemblance Analysis

This analysis examines the attributes of RD and STD independently to determine whether the univariate statistical characteristics of RD are preserved in STD. Statistical tests, distance calculation, and visual comparisons are proposed.

Statistical tests can be used to compare the attributes from RD and STD. They should be performed independently for each attribute with a proposed significance level of $\alpha = 0.05$, meaning that if the *p-value* obtained from the test is higher than this value, the null hypothesis (*h0*) is accepted. Otherwise, the alternative hypothesis (*h1*) is accepted. The properties analyzed in each test are preserved in STD if *h0* is accepted. For numerical attributes, the following tests are proposed:

- Student T-test for the comparison of means.
  - o *h0*: Means of RD feature and STD attribute are equal.
  - o *h1*: Means of RD feature and STD attribute are different.

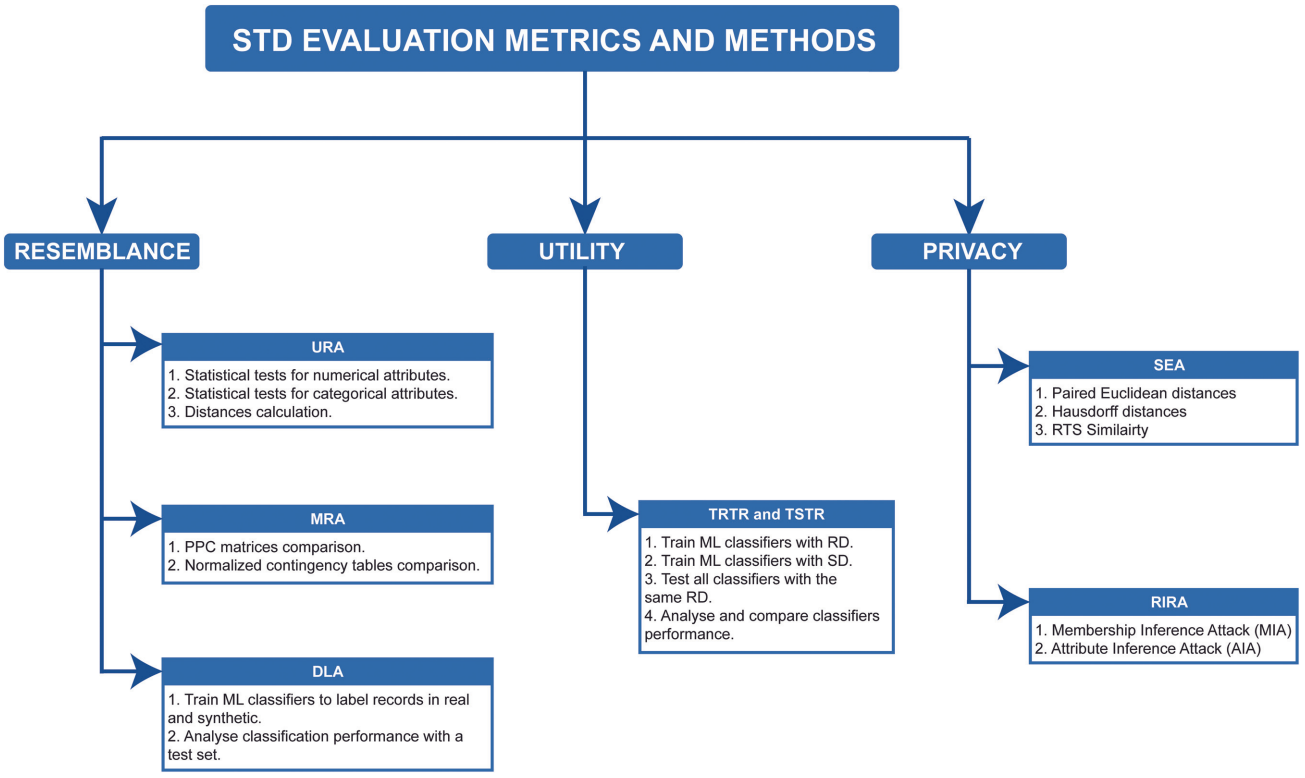

Fig. 1 Taxonomy of the proposed pipeline of metrics and methods to evaluate STD in three dimensions: resemblance, utility, and privacy. STD, synthetic tabular data.

- Mann Whitney U-test for population comparison.
  o *h0*: RD feature and STD attribute come from the same population.
  o *h1*: RD feature and STD attribute do not come from the same population.
- Kolmogorov–Smirnov test for distributions comparison.
  o *h0*: RD feature distribution and STD attribute distribution are equal.
  o *h1*: RD feature distribution and STD attribute distribution are not equal.

For categorical features, the Chi-square ($\chi 2$) test is proposed to analyze the feature independence between real and synthetic categorical attributes. If h0 is accepted, the statistical properties are not preserved in this case. *h0* and *h1* are defined as:

- *h0*: No statistical relationship exists between the real categorical variable and the synthetic categorical variable.
- *h1*: There is a statistical relationship between the real and synthetic categorical variables.

Some distance metrics can also be computed between the RD and STD attributes for URA. The lower the distance values are, the better the univariate resemblance is preserved in STD. Three distance metrics are proposed: cosine distance, Jensen-Shannon distance, and Wasserstein distance. Before computing all distances, RD and STD need to be scaled. In the following (**Equations 1 to 5**), r is the attribute of RD, and s is the attribute of STD.

Cosine distance is defined as the complement of cosine similarity, which is the cosine of the angle between two vectors in n-dimensional space; i.e., the dot product of the two vectors is divided by the product of the two vectors' lengths (Equation 1). As cosine similarity is defined positive, the cosine distance is bounded between 0 and 1, indicating the distance between two sets of values. The lower the value is, the higher the resemblance between the two sets of values will be. Therefore, a threshold of 0.3 has been experimentally set based on the exploratory analysis of the results to indicate that the STD attribute significantly resembles the RD attribute. A value higher than this threshold would represent no substantial resemblance between the STD attribute and RD attribute.

$$\cos\_dist(r,s) = 1 - \frac{\sum_{i=1}^{n} r_i \cdot s_i}{\sqrt{\sum_{i=1}^{n} r_i{}^2} \cdot \sqrt{\sum_{i=1}^{n} s_i{}^2}}$$

Equation 1: Cosine distance

Jensen-Shannon distance is the square root of the JSD, which measures the similarity between two probability distributions (Equation 2); m is the pointwise mean of p and q, and D is the Kullback-Leibler divergence, defined in Equation 3. The probability distributions of the features have been used to compute these distances: p is the probability distribution of the RD attribute, and q is the probability distribution of the STD attribute. A value lower than 0.1 represents a perfect resemblance since higher values would indicate that the differences in distributions are significantly higher.

$$js\_dist(p,q) = \sqrt{\frac{D(p||m) + D(q||m)}{2}}$$

Equation 2: Jensen-Shannon Distance

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Equation 3: Kullback-Leibler Divergence

Wasserstein distance can be seen as the minimum cost required to transform a vector (r) into another vector (s), where the cost is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved (Equation 4); R and S are the cumulative distribution function of the RD, and STD attributes, respectively. As in cosine distance, a threshold of 0.3 is proposed to assure the resemblance of the attribute.

$$was\_dist(r,s) = \int_{-\infty}^{+\infty} |R - S|$$

Equation 4: Wasserstein Distance

For statistical tests and the distance calculation, the number of STD attributes that fulfill the requirements to resemble RD attributes has been considered to categorize the STDG approach performance. If more than half of the attributes maintain resemblance, the approach is classified as "*Excellent.*" If less than half maintained resemblance, it is classified as "*Good*" and if none of the attributes maintained resemblance, it is classified as "*Poor.*"

To get an overall score of URA, the performances for all previously presented metrics can be calculated according to their results. First, the categorization is translated into a numerical value ("*Excellent*" = 3, "*Good*" = 2 and "*Poor*" = 1), and then the same weight (33.33%) is given for the three methods presented (statistical tests of numerical attributes, statistical tests of categorical attributes, and distance calculations for numerical attributes). The resulting score, rounded to the nearest integer, gives a value between 1 and 3 that indicates the URA score of STD.

Multivariate Relationship Analysis

This analysis involves assessing whether the multivariate relationships of RD are preserved in STD or not. To do that, the computation of two correlation matrices has been defined for RD and STD: PPC matrices for numerical variables (Equation 5) for each pair of features x and y; $\hat{x}$ and $\hat{y}$ are the mean value of the features and normalized contingency tables for categorical variables. The correlation matrices of RD can be visually compared with the matrices of STD using heatmaps.

$$PPC(x,y) = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2}}$$

Equation 5: PPC

Additionally, for each matrix, the differences between the correlations of RD and STD are calculated to compute the percentage of relationships maintained in the STD (those that have a different value lower than 0.1). This way, two values between 0 and 1 are obtained, one expressing the percentage of numerical attribute relationships maintained and the other the percentage of categorical attribute relationships maintained. If the values are higher than 0.6, the STDG approach is categorized as "*Excellent*" since more than half of the relationships are preserved in RD. If they are equal or between 0.4 and 0.6, it is categorized as "*Good,*" representing that half of the relationships are preserved. Finally, if the values are lower than 0.4, the performance of the STDG approach is "*Poor*" as it has preserved less than half of the relationships. After completing this categorization for each matrix, a total MRA performance is obtained by giving equal weight (50%) to both analyses.

Data Labelling Analysis (DLA)

The final step proposed for resemblance evaluation can be used to evaluate the semantics of STD. This method is proposed to analyze the performance of some classifiers when labeling records as real or synthetic through the following steps:

1. Combine and label real and synthetic datasets (0 for RD and 1 for STD) in a single dataset.
2. Split the combined dataset into train and test sets, i.e., 80:20 split.
3. Pre-process the train and test data, i.e., standardize numerical attributes and one-hot encode categorical attributes.
4. Train (with training data) and evaluate (with test data) some ML classifiers to analyze their performance in labelling records as real or synthetic.

The commonly used and diverse ML classifiers proposed for this analysis are Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machines (SVM) and Multilayer Perceptron (MLP). For our analysis, these classifiers have been implemented using Python's Scikit-Learn 0.24.2 ML library with the listed parameters:

- The RF model has been implemented using *RandomForestClassifier* with *n_estimators = 100, n_jobs = 3, random_state = 9* and all other parameters set to their defaults.
- The KNN model has been implemented using *KNeighborsClassifier* with *n_neighbors = 10, n jobs = 3* and all other parameters set to their defaults.
- The DT model has been implemented using *DecisionTreeClassifier* with *random_state = 9* and all other parameters set to their defaults.
- The SVM model has been implemented using *SVC* with *C = 100, max_iter = 300, kernel = "linear," probability = True, random state = 9* and all other parameters set to their defaults.
- The MLP model has been implemented using MLPClassifier with *hidden_layer_sizes = (128,64,32), max_iteration = 300, random_state = 9* and all other parameters set to their defaults.

After training and testing the models, classification performance metrics (accuracy, precision, recall, and F1-score) can be

analyzed visually with box plots. To indicate that the semantics of RD are preserved in STD, a classifier should not distinguish if a record is synthetic or real. Thus, the classification metrics should be lower than or equal to 0.6 for "*Excellent*" resemblance, meaning that the models have classified most of the synthetic records as real. Obtaining metric values higher than 0.6 and lower than 0.8 indicate a "*Good*" resemblance, while values equal to or greater than 0.8 indicate a "*Poor*" resemblance.

Total Resemblance Performance

After completing the four analyses proposed for resemblance evaluation, a weighted average can be computed to obtain a total resemblance score. This allows us to assign higher or lower importance to each result based on the researchers' STDG goals. An example weighting applied in STD Evaluation (section 4) has been URA = 0.4, MRA = 0.4, and DLA = 0.2. These weights were experimentally set to prioritize the ability of STD to resemble univariate and multivariate patterns over the ability of a classifier to distinguish between RD and STD for this evaluation. Finally, by applying these weights, the total score is obtained, and after rounding it to the nearest integer, the total score is categorized as "*Excellent*" if the resulting value is 3, "*Good*" if the resulting value is 2, or "*Poor*" if the resulting value is 1.

Utility Evaluation

In the utility dimension, the ability of STD, instead of RD, to train ML models is analyzed to determine if ML models trained with STD produce similar results to ML models trained with RD. To do that, TRTR and TSTR analyses are proposed. ML classifiers should be trained with RD and then separately with STD. The same ML classifiers described for DLA analysis are proposed for this evaluation due to their simplicity, scalability, and training efficiency.

The data must be pre-processed before training and testing the models, i.e., by standardizing numerical attributes and one-hot encoding categorical attributes. All trained models should be tested with the same RD (20% of the real dataset held out before training the STDG approaches). To analyze the classification results, accuracy, precision, recall, and F1-score classification metrics are proposed, and their absolute differences when TRTR and TSTR. To assure that STD utility is "*Excellent,*" the metrics differences should not exceed a proposed threshold of 0.2. If differences between 0.2 and 0.8 are obtained, the performance in the utility dimension should be categorized as "*Good*" and if they are higher than 0.8, the performance is considered "*Poor.*"

Privacy Evaluation

For the privacy dimension, it is proposed to evaluate the similarity of RD and STD and the re-identification risk of real patients or records.

Similarity Evaluation Analysis (SEA)

In the SEA, it is proposed to evaluate how private STD is compared to RD in terms of similarity between real and synthetic records. Based on the distance and similarity metrics for privacy evaluation used by other authors and mentioned in Section 1.2.2, three metrics are proposed for this: Euclidean distance between each pair of records, Hausdorff distance between STD and RD, and Synthetic To Real (STR) similarity.

The Euclidean distance is the square root of the sum of squares of differences between RD and STD features, as defined in Equation 6. In this case, the Euclidean distance can be computed for each pair of records. Then, the mean and standard deviation of all distances should be analyzed. The higher the mean distance and the lower the standard deviation are, the more the privacy is preserved. Thus, mean values higher than 0.8 and standard deviation values lower than or equal to 0.3 indicate privacy is preserved.

$$euc\_dist(r,s) = \sqrt{\sum_{i=1}^{n}(r_i - s_i)^2}$$

Equation 6: Euclidean Distance

The cosine similarity metric is proposed to compute the STR similarity, which computes similarity as the normalized dot product of two datasets (Equation 7; *R* is a record from RD and *S* is a record from STD). The pairwise similarity value can be computed for each pair of records, and the mean and maximum values of those pairwise similarity values should be analyzed. If the mean value is higher than 0.5, the STD is very close to the RD, so privacy is not preserved. In all other cases, it can be said that privacy is preserved.

$$\cos\_sim(S,R) = \frac{\sum_{i=1}^{n} r_i \cdot s_i}{\sqrt{\sum_{i=1}^{n} {r_i}^2} \cdot \sqrt{\sum_{i=1}^{n} {s_i}^2}}$$

Equation 7: Cosine Similarity

The Hausdorff distance measures how far two subsets of a metric space are from each other as it is the greatest of all the distances from a point in one set to the closest point in the other set (**Equation 8**; *R* is the real dataset; *S* is the synthetic dataset). Two sets are close in the Hausdorff distance if every point of either set is close to some point in the other set. Thus, the higher this distance value is, the better the privacy is preserved in STD, as a high value indicates that the STD is far from the RD. Since this metric is not bounded between 0 and 1, a value higher than 1 has been considered to assure that privacy is preserved.

$$haus\_dist(S,R) = max\{h(S,R), h(R,S)\}$$

Equation 8: Hausdorff Distance

For these three distance-based metrics, if all three metrics fulfill the condition to preserve privacy, the categorization for privacy preservation is "*Excellent.*" If one or two metrics fulfill the condition, it is "*Good*"; otherwise, it is "*Poor.*"

Re-Identification Risk Analysis (RIRA)

This analysis proposes to evaluate the level of disclosure risk if an attacker or an adversarial obtains access to STD and a subset of RD. For this, two simulations are proposed: (1)

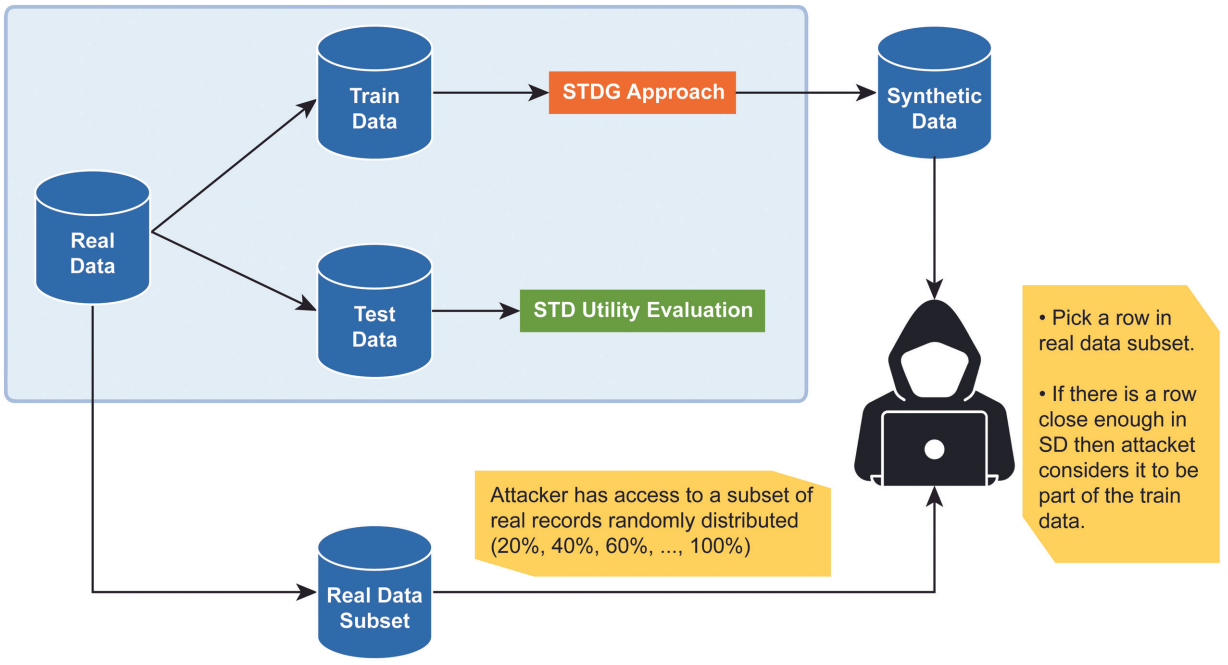

Fig. 2 Simulated MIA. MIA, membership inference attack.

membership inference attack (MIA) when an attacker tries to identify if real patient records have been used to train the STDG approach (►**Fig. 2**), and (2) attribute inference attack (AIA) when an attacker has access to some attributes of the RD and tries to guess the value of an unknown attribute of a patient from the STD (►**Fig. 3**).[57]

In an MIA, if the attacker determines that real records were used to train the STDG approach, it could be said that they have re-identified the patient from the STD.[57] ►**Fig. 2** illustrates this attack where a hypothetical attacker has access to all records of the STD and a subset of the RD randomly distributed. Using a patient record (r) from the RD subset, the attacker will try to identify the closest records in the STD with a distance metric calculation. If there is any distance lower than some threshold, the attacker determines that there is at least one row close enough to RD in the STD, meaning that r has been used to generate STD. This process is depicted in **Algorithm 1** and ►**Fig. 2**.

For this analysis, the attacker's success rate is proposed to be evaluated by simulating this kind of attack by calculating

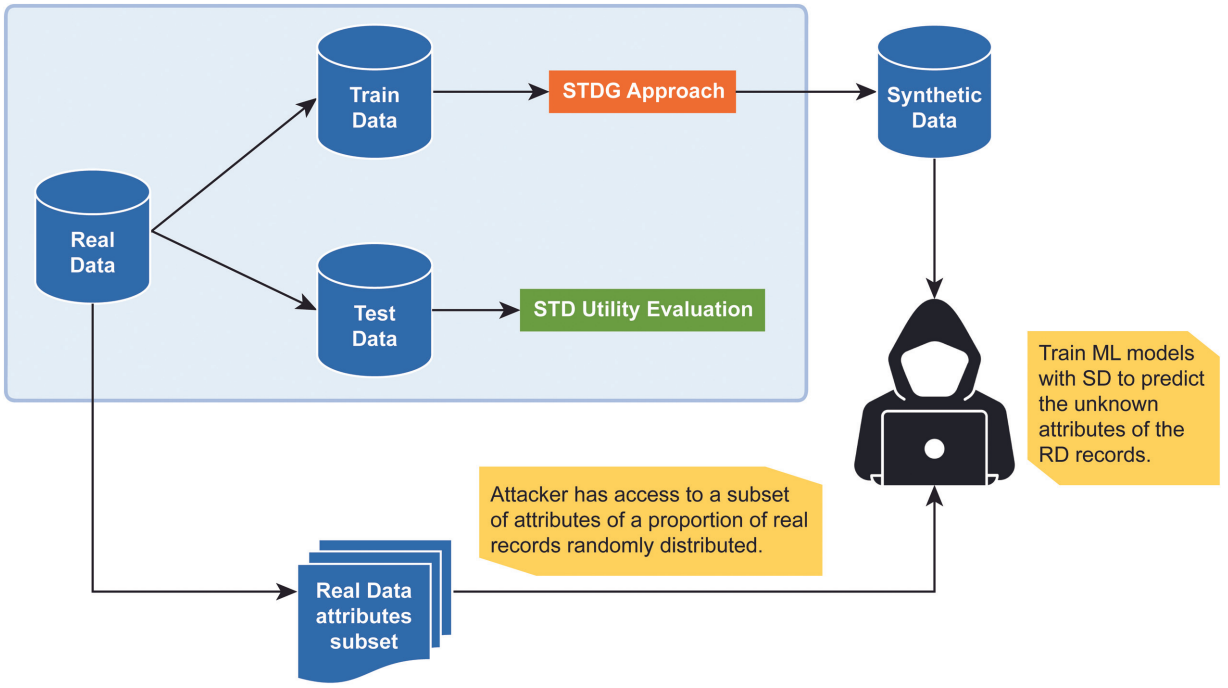

Fig. 3 Simulated AIA. AIA, attribute inference attack.

Algorithm 1: Simulation of a MIA

```
attacker_data ← real set of data known by the attacker
synthetic_data ← synthetic set of data that is publicly
available
for row in attacker_data
  dist ← compute_distances(row, synthetic_data)
  if any(dist) < thr then
    record_identified ← True
  else
    record_identified ← False
```

the Hamming distance (Equation 9), which represents the proportion of non-equal attributes between two records, between each row of the RD subset and STD rows.

$$haus\_dist(S,R) = max\{h(S,R),h(R,S)\}$$

Equation 9: Hamming distance

Reasonable thresholds to assure that r is close enough to a record in STD are 0.4, 0.3, 0.2, and 0.1 because a lower Hamming distance represents more similarity in the records. Since it is known if r belongs to the training data or not, accuracy (proportion of correct predictions made by the attacker) and precision (proportion of records used for training the STDG approach identified by the attacker) values have been calculated. To obtain an "Excellent" privacy preservation categorization, accuracy and precision values should be 0.5 or lower for all thresholds. Any value above 0.5 indicates increasing levels of disclosure risk, obtaining a "*Good*" or "*Poor*" privacy preservation categorization depending on the number of thresholds where the values are higher than 0.5. The accuracy and precision values have been plotted to analyze and interpret these results as a function of the proportion of records in the STD present in the RD subset known by the attacker for each threshold.

In an AIA, an attacker has access to STD and a subset of attributes for some RD records, generally QIDs such as age, gender, height, weight, etc.[57] As shown in ►Fig. 3 and Algorithm 2 the attacker will use ML models trained with STD to predict the values of rest of the attributes of the RD records.

The success of this attack can be measured by defining the QID of each dataset and then using them to train ML models, i.e., DT models, with STD to predict the rest of the attributes. Next, 50% of the RD (randomly distributed) can be used to evaluate the performance of the models, generating batches of data with each QID combination. This way, the predictions made by the models trained on STD for each data batch combination can be evaluated using accuracy for categorical attributes and root-mean-squared-error (RMSE) for numerical attributes. Higher accuracy values (close to 1) indicate higher disclosure risk, while lower RMSE values (close to 0) reflect higher disclosure risk. The metric values are obtained for each QID combination, and all attributes are evaluated visually with a boxplot for each risk attribute. Additionally, the percentage of correctly predicted attributes has been calculated to categorize the re-identification risk. The mode of the analyzed metrics has been considered to determine if an attribute has been predicted; for categorical attributes, accuracy mode is equal to 1, and for numerical attributes, RMSE mode is equal to 0. If a high percentage (higher than 0.6) is obtained, as more than half of the attributes have been re-identified, the AIA results are categorized as "*Poor*." For a percentage between 0.4 and 0.6, the result is categorized as "*Good*" since approximately half of the attributes have been re-identified. A percentage lower than 0.4 indicates that less than half of the attributes have been re-identified, demonstrating "*Excellent*" privacy preservation for this attack.

Algorithm 2: Simulation of an AIA

```
attacker_data ← real data set of QID known by the attacker
synthetic_data ← synthetic set of data that is publicly
available
x_train ← QID of the synthetic set of data
attributes ← List of non QID attributes of the set of data
y_train ← Non QID attributes of the synthetic set of data
for attribute in attributes
  ml_model.train(x_train, y_train[attribute])
  inferred_values ← ml_model.predict(attacker_data)
```

Total Privacy Performance

After computing the results for the three privacy evaluation methods proposed, a total privacy score should be calculated, weighting the results from all the methods. For example, in the STD evaluation, the following weights have been experimentally given to the three analyses for the privacy dimension: SEA = 0.4 and RIRA = 0.6 (MIA = 0.3 and AIA = 0.3). A higher weight was assigned to RIRA to prioritize the metrics analyzing whether RD could be inferred from STD over the metric and investigating how similar STD is to RD. Finally, once this weighting is applied and the result is rounded to the nearest integer (i.e., assigned to 1, 2 or 3), the privacy of STD is categorized as "*Excellent*" (3), "*Good*" (2), or "*Poor*" (1) based on the total score obtained when applying these weights.

Synthetic Tabular Data Evaluation

To prove and trust the efficiency and usability of the proposed pipeline for STD evaluation, several different datasets have been selected and then synthesized with different STDG approaches. The STD generated for the selected datasets is then evaluated and benchmarked. All code and example notebooks are available in a Github repository.[59]

Selected Data

Six open-source health care-related datasets have been selected for synthesis. A brief description of these datasets and the number of attributes and records are presented in ►Table 1. Each dataset was first pre-processed, deleting missing values and performing a data split into two subsets. Eighty percent of the records have been used for training the STDG approaches, and 20% of the records for utility dimension evaluation and RIRA simulations.

Fully Synthetic Tabular Data Generation Approaches

To generate STD for the datasets presented previously, four open-source STDG approaches have been used, two of which

Table 1 Brief description of the selected health-related datasets for STDG and STD evaluation

ID	Dataset name	Year	Num. attrib.	Num. records
A	Diabetes 130-U.S. hospitals for years 1999–2008 dataset[65]	2014	55	101,766
B	Cardiovascular disease (CVD) dataset[66]	2019	13	70,000
C	Estimation of obesity levels based on eating habits and physical condition dataset[67]	2019	17	2,111
D	Contraceptive Method Choice (CMC) dataset[68]	1997	9	1,473
E	Pima Indians Diabetes (PID) dataset[69]	2016	9	769
F	Indian Liver Patient (ILP) dataset[70]	2012	11	583

are GANs, and the other two are classical approaches. These approaches are as follows:

1. *Gaussian Multivariate (GM)*: A classical STDG approach based on statistical modeling that implements a multivariate distribution using a Gaussian Copula to combine marginal probabilities estimated using univariate distributions. The approach is available at Copulas 0.5.0 documentation.[60]
2. *Synthetic Data Vault (SDV)*: This approach is an STDG ecosystem of libraries that uses several probabilistic graphical modeling and DL-based techniques. To enable a variety of data storage structures, they employ unique hierarchical generative modeling and recursive sampling techniques.[61] It is available at The Synthetic Data Vault.[51]
3. *Conditional Tabular Generative Adversarial Network (CTGAN)*: A STDG approach proposed by Xu et al in 2019[49] defined as a collection of DL models based on GAN models for a single data table. It can learn from RD and generate synthetic clones with high fidelity. It is available at.[62]
4. *Wasserstein Generative Adversarial Network with Gradient Penalty (WGANGP)*: This approach is a GAN proposed by Yale et al in 2020[8] and is composed of a generator and discriminator. The generator learns to generate better STD based on the feedback received by the discriminator and using the Wasserstein distance with gradient penalty as the optimization function. The approach is available at.[63]

Ethical Considerations

This research has not involved data collection from human subjects. Used data is open-source data.

Results

Using the datasets and STDG approaches described in the previous sections, the metrics and methods proposed for STD evaluation in the resemblance, utility, and privacy dimensions have been applied to evaluate the generated STD to evidence and provide trust in their efficiency and usability. Additionally, comparison and benchmarking of the STDG approaches have been performed based on the proposed strategy for STD evaluation.

In the following subsections, the results obtained after applying the defined weighting criteria for each dimension are explained per dataset. A detailed and complete description of the results from all metrics and methods is available in (►Supplementary Material 1, available in online version only). In this supplementary material there are additional metrics that have been excluded from the methodology and explained in the **Appendix B**. Based on the results of each evaluation analysis presented in the pipeline, the results obtained for each synthesized dataset have been summarized and categorized as "*Excellent,*" "*Good,*" or "*Poor*" for each dimension (resemblance, utility, and privacy), based on the final weighted scores.

Resemblance Evaluation

►Table 2 shows the results of applying the proposed resemblance evaluation metrics and methods to the STD synthesized with each STDG approach for each dataset.

The resemblance dimension has been perfectly maintained with GM for the six datasets. SDV has performed excellently on the resemblance dimension for half of the datasets (C, E, F) and performance was good for other datasets (A, B, and D). CTGAN has performed very well in retaining resemblance for one dataset (B), quite well for four datasets (A, C, D, and F), and poorly for one dataset (E). Finally, WGANGP has been the worst approach in generating STD that resembles RD, as it has produced good resemblance for four datasets (A, B, C, and D) and poor resemblance for the other two datasets (E and F). Although the STD generated with the four approaches has maintained the univariate and multivariate resemblance perfectly for most of the datasets, the results from the DLA analyzes propose that the generated STD is easily distinguishable from RD for all approaches and all datasets, except GM. This approach has obtained the best result for DLA analysis.

Despite the poor resemblance for two datasets (E and F), applying the proposed metrics and methods for resemblance evaluation can assure that resemblance has been maintained for most of the datasets using the four STDG approaches.

Utility Evaluation

The results from the utility evaluation for each dataset and STDG approach are shown in ►Table 3. The utility has been perfectly maintained across all approaches for only Dataset A since the difference in the classification metrics when TRTR and TSTR are lower than 0.2 in all cases. For Dataset B, only with SDV a difference in classification metrics higher than 0.2 has been obtained, categorizing this model as having "*Good*"

Table 2 Results of the resemblance evaluation for each dataset

Data ID	STDG approaches	URA (40%)	MRA (40%)	DLA (20%)	Total Resemblance
A	GM	(3) Excellent	(3) Excellent	(2) Good	(3) Excellent
	SDV	(2) Good	(3) Excellent	(1) Poor	(2) Good
	CTGAN	(2) Good	(3) Excellent	(2) Good	(2) Good
	WGANGP	(1) Poor	(2) Good	(2) Good	(2) Good
B	GM	(3) Excellent	(3) Excellent	(2) Good	(3) Excellent
	SDV	(2) Good	(3) Excellent	(2) Good	(2) Good
	CTGAN	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	WGANGP	(2) Good	(2) Good	(1) Poor	(2) Good
C	GM	(3) Excellent	(3) Excellent	(1) Poor	(3) Excellent
	SDV	(3) Excellent	(3) Excellent	(1) Poor	(3) Excellent
	CTGAN	(2) Good	(3) Excellent	(1) Poor	(2) Good
	WGANGP	(2) Good	(2) Good	(1) Poor	(2) Good
D	GM	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	SDV	(2) Good	(3) Excellent	(2) Good	(2) Good
	CTGAN	(2) Good	(1) Poor	(2) Good	(2) Good
	WGANGP	(2) Good	(3) Excellent	(1) Poor	(2) Good
E	GM	(3) Excellent	(3) Excellent	(2) Good	(3) Excellent
	SDV	(3) Excellent	(3) Excellent	(2) Good	(3) Excellent
	CTGAN	(2) Good	(1) Poor	(1) Poor	(1) Poor
	WGANGP	(2) Good	(1) Poor	(1) Poor	(1) Poor
F	GM	(3) Excellent	(2) Good	(3) Excellent	(3) Excellent
	SDV	(3) Excellent	(3) Excellent	(1) Poor	(3) Excellent
	CTGAN	(2) Good	(2) Good	(2) Good	(2) Good
	WGANGP	(2) Good	(1) Poor	(1) Poor	(1) Poor

Table 3 Results of the utility evaluation for each dataset

Data ID	STDG approaches	Acc. diff.	Prec. diff.	Rec. diff.	F1 diff.	Utility score
A	GM	0.05	0.05	0.05	0.05	(3) Excellent
	SDV	0.05	0.10	0.05	0.10	(3) Excellent
	CTGAN	0.05	0.05	0.05	0.10	(3) Excellent
	WGANGP	0.15	0.10	0.10	0.10	(3) Excellent
B	GM	0.05	0.10	0.05	0.15	(3) Excellent
	SDV	0.20	0.20	0.20	0.40	(2) Good
	CTGAN	0.10	0.05	0.10	0.15	(3) Excellent
	WGANGP	0.20	0.20	0.15	0.20	(3) Excellent
C	GM	0.60	0.60	0.60	0.60	(2) Good
	SDV	0.80	0.75	0.75	0.75	(2) Good
	CTGAN	0.90	0.85	0.90	0.90	(1) Poor
	WGANGP	0.60	0.60	0.65	0.70	(2) Good
D	GM	0.15	0.10	0.10	0.15	(3) Excellent
	SDV	0.15	0.20	0.20	0.25	(2) Good
	CTGAN	0.15	0.15	0.15	0.15	(3) Excellent
	WGANGP	0.15	0.20	0.15	0.25	(2) Good
E	GM	0.20	0.20	0.20	0.20	(2) Good
	SDV	0.05	0.05	0.05	0.10	(3) Excellent
	CTGAN	0.35	0.25	0.35	0.40	(2) Good
	WGANGP	0.25	0.20	0.25	0.25	(2) Good
F	GM	0.25	0.15	0.30	0.30	(2) Good
	SDV	0.35	0.20	0.40	0.40	(2) Good
	CTGAN	0.20	0.20	0.20	0.20	(2) Good
	WGANGP	0.40	0.25	0.35	0.40	(2) Good

Note: The first and second columns refer to the dataset and STDG approach. The third to sixth columns indicate the maximum difference in the classification metric when TRTR and TSTR. The last column indicates the final utility categorization based on the metrics differences and applying a threshold of 0.2 for an "Excellent" categorization.

performance and the other as "*Excellent.*" For dataset C, the utility of the STD is poor for CTGAN, which have obtained metrics differences higher than 0.8, while for the other approaches (GM, SDV, and WGANGP), utility has been maintained but not perfectly, with metrics values differences between 0.6 and 0.8. For Dataset D, GM and CTGAN have performed perfectly (classification metrics difference lower than 0.2) while SDV and WGANGP have been categorized as "Good" (classification metrics higher than 0.2 and lower than 0.8). For Dataset E, only SDV has been scored as "Excellent," while the others are "Good." Finally, for dataset F, since all classification metrics are higher than 0.2 but lower than 0.8, the four STDG approaches have been categorized as "Good" for the utility dimension.

Using the proposed methods and threshold for utility evaluation, these results show that the utility of STD has been maintained in all cases except for dataset (C).

Privacy Evaluation

As shown in ►Table 4, the privacy of STD has been quite well maintained with GM for all the datasets. For SDV, privacy has been perfectly maintained in three datasets (C, D, and F) and quite well maintained for the other three (A, B, and E). With CTGAN, privacy has been perfectly maintained for three datasets (A, C, and F) and quite well maintained for the other three (B, D, and E). With WGANGP, privacy has only been perfectly maintained for two datasets (A and D) and quite well maintained for the other four datasets (B, C, E, and F).

Regarding the RIRA, the MIA has yielded better results than the AIA for most cases. This finding indicates that for most of the dataset and STDG approach combinations, the STD is more prone to re-identification in terms of attributes rather than membership.

Applying the proposed privacy dimension evaluation metrics and methods shows that the STD generated with GM has yielded better privacy preservation than for other approaches. However, privacy is still preserved quite well with the other approaches, offering better preservation for certain datasets and STDG approach combinations.

Results Summary

The categorizations made for each dataset and STDG approach combination as a result of applying the proposed metrics and methods are described in each subsection. Since giving equal weights to the three dimensions could undervalue the others, three different weighting scenarios are proposed to compute the final performance score for each STDG approach per dataset. These scenarios have been inspired by El Emam et al.[64] Next, each weighting scenario is presented, and the results obtained from them are explained.

Weighting Scenario 1: No Preference

For this scenario the same weight (33.33%) has been given to the three evaluation dimensions (resemblance, utility, and privacy) since there is no preference on which dimension should have more importance.

►Table 5 shows the results obtained when applying this weighting scenario. From the table it can be concluded that it is not clear which STDG approach is better in general (across all datasets), using the evaluation metrics defined and the applied weights. However, it is possible to determine the best STDG approach(es) for each application (i.e., dataset). GM

Table 4 Results of the privacy evaluation for each dataset

Data ID	STDG approaches	SEA (40%)	MIA (30%)	AIA (30%)	Total resemblance
A	GM	(1) Poor	(3) Excellent	(3) Excellent	(2) Good
	SDV	(2) Good	(3) Excellent	(1) Poor	(2) Good
	CTGAN	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	WGANGP	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
B	GM	(1) Poor	(3) Excellent	(1) Poor	(2) Good
	SDV	(2) Good	(3) Excellent	(1) Poor	(2) Good
	CTGAN	(2) Good	(3) Excellent	(1) Poor	(2) Good
	WGANGP	(2) Good	(3) Excellent	(1) Poor	(2) Good
C	GM	(2) Good	(2) Good	(3) Excellent	(2) Good
	SDV	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	CTGAN	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	WGANGP	(2) Good	(2) Good	(3) Excellent	(2) Good
D	GM	(2) Good	(3) Excellent	(2) Good	(2) Good
	SDV	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	CTGAN	(2) Good	(3) Excellent	(2) Good	(2) Good
	WGANGP	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
E	GM	(2) Good	(3) Excellent	(1) Poor	(2) Good
	SDV	(2) Good	(2) Good	(1) Poor	(2) Good
	CTGAN	(2) Good	(3) Excellent	(1) Poor	(2) Good
	WGANGP	(2) Good	(2) Good	(1) Poor	(2) Good
F	GM	(2) Good	(2) Good	(3) Excellent	(2) Good
	SDV	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	CTGAN	(2) Good	(3) Excellent	(3) Excellent	(3) Excellent
	WGANGP	(2) Good	(2) Good	(3) Excellent	(2) Good

Table 5 Overall results of the STD evaluation for all datasets and STDG approaches combination when applying Weighting Scenario 1 (no preference on which dimension is more important)

ID	Data shape	STDG appr.	Resemblance (33.33%)	Utility (33.33%)	Privacy (33.33%)	Final score
A	Records: 101,766 Attributes: 20	GM* SDV CTGAN* WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (3) Excellent (3) Excellent (3) Excellent	(2) Good (2) Good (3) Excellent (3) Excellent	(3) Excellent (2) Good (3) Excellent (3) Excellent
B	Records: 70,000 Attributes: 13	GM* SDV CTGAN* WGANGP	(3) Excellent (2) Good (3) Excellent (2) Good	(3) Excellent (2) Good (3) Excellent (3) Excellent	(2) Good (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (2) Good
C	Records: 2,111 Attributes: 17	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (2) Good (2) Good	(2) Good (2) Good (1) Poor (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (2) Good
D	Records: 1,473 Attributes: 10	GM* SDV CTGAN WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (3) Excellent	(3) Excellent (2) Good (2) Good (2) Good
E	Records: 769 Attributes: 9	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (1) Poor (1) Poor	(2) Good (3) Excellent (2) Good (2) Good	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (2) Good (2) Good
F	Records: 583 Attributes: 11	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (2) Good (1) Poor	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (2) Good

Note: The STDG approaches that yielded an "Excellent" score are marked with *.

and SDV scored "Excellent" for three datasets, CTGAN for only two and WGANGP for only one dataset. This suggests that GM and SDV have given the best overall results from the chosen STDG approaches and datasets when applying equal weights to the three dimensions.

Weighting Scenario 2: External Software Company

In this weighting scenario that might be used by an external software company for software and STDG approaches validation, more importance should be given to resemblance and privacy than to utility. Specifically, due to privacy guarantees, the dimension that should have more importance in this scenario is privacy. Therefore, the applied weights for this scenario are the following: 40% to resemblance, 10% to utility, and 50% to privacy.

►**Table 6** shows the results obtained when applying this weighting scenario. Using the evaluation metrics defined and the applied weights the results are quite similar to the previous weighting scenario except for a few differences. SDV scored "Excellent" for four datasets, GM and CTGAN for three datasets, and WGANGP for only two datasets. This suggests that SDV has given the best overall results from the chosen STDG approaches and datasets when giving more importance to privacy and resemblance and less importance to utility.

Weighting Scenario 3: Internal Organization with Security Measures

This weighting scenario might be used by an internal organization that implements security measures. Therefore, in this scenario, privacy is the dimension that has least importance and utility is the one with the most importance. However, a high level of significance should be given to the resemblance evaluation. The applied weights for this scenario are as follows: 30% to resemblance, 60% to utility, and 10% to privacy.

►**Table 7** shows the results obtained when applying this weighting scenario. Using the evaluation metrics defined and the applied weights the results have changed a lot from the previous scenarios. GM, CTGAN, and WGANGP scored "Excellent" for three datasets while SDV only for one dataset. Apart from that, there are two datasets in which none of the four STDG approaches have yielded an "Excellent" score. This suggests that GM, CTGAN, and WGANGP have given the best overall results from the chosen STDG approaches and datasets when giving more importance to utility and less importance to privacy. Additionally, in this specific case, for certain datasets an optimal score has not been obtained.

Discussion

Principal Results

Overall, the results have shown that the proposed pipeline for STD evaluation in the three dimensions defined (resemblance, utility, and privacy) can be used to assess and benchmark STD generated with different approaches and applying different weighting scenarios according to which dimension the researchers want to prioritize. Contrary to the evaluation frameworks proposed by Dankar et al,[41] Hittmeir

Table 6 Overall results of the STD evaluation for all datasets and STDG approaches combination when applying Weighting Scenario 2 (privacy the most important dimension and utility the least)

ID	Data shape	STDG appr.	Resemblance (40%)	Utility (10%)	Privacy (50%)	Final score
A	Records: 101,766 Attributes: 20	GM* SDV CTGAN* WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (3) Excellent (3) Excellent (3) Excellent	(2) Good (2) Good (3) Excellent (3) Excellent	(3) Excellent (2) Good (3) Excellent (3) Excellent
B	Records: 70,000 Attributes: 13	GM* SDV CTGAN* WGANGP	(3) Excellent (2) Good (3) Excellent (2) Good	(3) Excellent (2) Good (3) Excellent (3) Excellent	(2) Good (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (2) Good
C	Records: 2,111 Attributes: 17	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (2) Good (2) Good	(2) Good (2) Good (1) Poor (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (2) Good
D	Records: 1,473 Attributes: 10	GM* SDV* CTGAN WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (3) Excellent	(3) Excellent (3) Excellent (2) Good (3) Excellent
E	Records: 769 Attributes: 9	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (1) Poor (1) Poor	(2) Good (3) Excellent (2) Good (2) Good	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (2) Good (2) Good
F	Records: 583 Attributes: 11	GM SDV* CTGAN* WGANGP	(3) Excellent (3) Excellent (2) Good (1) Poor	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good

Note: The STDG approaches that yielded an "Excellent" score are marked with *.

Table 7 Overall results of the STD evaluation for all datasets and STDG approaches combination when applying Weighting Scenario 3 (preference on utility, resemblance also important, but privacy not important)

ID	Data shape	STDG appr.	Resemblance (30%)	Utility (60%)	Privacy (10%)	Final score
A	Records: 101,766 Attributes: 20	GM* SDV* CTGAN* WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (3) Excellent (3) Excellent (3) Excellent	(2) Good (2) Good (3) Excellent (3) Excellent	(3) Excellent (3) Excellent (3) Excellent (3) Excellent
B	Records: 70,000 Attributes: 13	GM* SDV CTGAN* WGANGP*	(3) Excellent (2) Good (3) Excellent (2) Good	(3) Excellent (2) Good (3) Excellent (3) Excellent	(2) Good (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (3) Excellent
C	Records: 2,111 Attributes: 17	GM SDV CTGAN WGANGP	(3) Excellent (3) Excellent (2) Good (2) Good	(2) Good (2) Good (1) Poor (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (2) Good (2) Good (2) Good
D	Records: 1,473 Attributes: 10	GM* SDV CTGAN* WGANGP*	(3) Excellent (2) Good (2) Good (2) Good	(3) Excellent (2) Good (3) Excellent (2) Good	(2) Good (3) Excellent (2) Good (3) Excellent	(3) Excellent (2) Good (3) Excellent (2) Good
E	Records: 769 Attributes: 9	GM SDV* CTGAN WGANGP	(3) Excellent (3) Excellent (1) Poor (1) Poor	(2) Good (3) Excellent (2) Good (2) Good	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (2) Good (2) Good
F	Records: 583 Attributes: 11	GM SDV CTGAN WGANGP	(3) Excellent (3) Excellent (2) Good (1) Poor	(2) Good (2) Good (2) Good (2) Good	(2) Good (3) Excellent (3) Excellent (2) Good	(2) Good (2) Good (2) Good (2) Good

Note: The STDG approaches that yielded an "Excellent" score are marked with *.

et al,[39] and Platzer and Reutterer,[43] our pipeline has covered and evaluated all three dimensions identified as relevant for ensuring the delivery of high-quality STD in the health domain as well as giving the option to tune the weighting methodology according to the desired output. Additionally, the proposed categorization strategy is useful for comparing different STDG approaches to select the best ones and evaluating the quality of STD generated by one approach.

None of the STDG approaches have been better across all STD dimensions considered in the experiments described for the three proposed weighting scenarios. Therefore, it can be said that it is difficult to get a trade-off between resemblance, privacy, and utility scores. However, the categorization system provided and the per dimension score calculation and overall score calculation can help select the most appropriate STDG approaches by looking at individual scores and configuring overall scores by computing weights according to priorities defined for each specific application.

Resemblance Evaluation

Different metrics and methods have been proposed to evaluate the resemblance of STD at different levels: univariate, multivariate, dimensional, and semantics. Among the metrics used in URA, statistical tests and distance calculations provide quantitative results, which have been more trustworthy in assessing how well STD attributes resemble RD than visual comparisons of the attribute distributions, which provide a more qualitative view. MRA has appeared to be useful in analyzing the multivariate relationships between attributes and how well these are maintained in STD. Although DRA has not been included in the total resemblance calculation due to the difficulty in interpreting the results obtained, it can be useful in illustrating how well the dimensional properties of RD are preserved. DLA has been less effective due to the lack of medical specialists to qualitatively evaluate the generated data's significance. However, the analysis composed of different ML models trained to label records as real or synthetic has simplified this process, approximating how a medical expert would label the records.

Utility Evaluation

TRTR and TSTR methods have been proposed in terms of the utility dimension, where a few ML classification models have been trained with RD and separately with STD. This methodology has been very useful in evaluating whether STD could be used instead of RD for data modeling. Further work in this area may include trying and validating different ML models to select the best ones for the specific application and using mixed data for model training in contrast to the separate use of RD or STD. Furthermore, other data modeling tasks can be proposed to assess the utility of STD in the same way, e.g., regression or clustering. Additionally, statistical tests can be applied to analyze if the classification metrics differences are statistically significant so the categorization into "Excellent"; "Good" and "Poor" performance can be made based on the results of these statistical tests.

Privacy Evaluation

Regarding the privacy of STD, the similarity between STD and RD has first been evaluated in the SEA, and then a pair of data inference attacks (MIA and AIA) have been simulated in the RIRA analysis. Although these simulations have not been quite significant, they could be useful to estimate the quantification of the re-identification risk of STD. However, these metrics and methods must be improved to quantify the re-identification risk of STD more reliably. Thus, future work might include defining a strategy that helps identify which attributes are more prone to re-identification and considers the real consequences of potential personal data disclosure.

Weighting Scenarios for Overall Score

Regarding the three proposed weighting scenarios to obtain an overall score for the four STDG approaches per dataset, similar results have been obtained for the first two scenarios (giving equal weights and prioritizing privacy over resemblance and utility). For the proposed third weighting scenario (prioritizing utility over resemblance and privacy), the results have changed a lot, yielding no "Excellent" STDG approach for two datasets. Apart from that, in the first scenario GM and SDV were the best approaches while in the second scenario it was SDV. In the third scenario the best approaches were GM, CTGAN and WGANGP. From these results it can be concluded that SDV is a good approach for privacy preservation, while the other three (GM, CTGAN and WGANGP) have performed better for utility preservation.

Limitations and Future Work

Although the proposed STD evaluation pipeline has been used to evaluate STD generated with different approaches and contexts, the datasets used for the evaluation have been limited due to the lack of quality health-related open-source datasets. From the six datasets selected, only two (A and B) comprised an appropriate number of records to be considered representative of real health-related data, with the remaining four containing a limited number of entries. Moreover, these open-source datasets might have been anonymized or synthesized before, inducing new bias in STD and analysis. Therefore, further work is required to judge and benchmark the proposed pipeline with more datasets in other contexts and RD that comes directly from hospitals or laboratories without any anonymization or other modification processes applied to the data. Apart from that, different weighting scenarios should be tested when computing the final score that considers the three defined dimensions and also weighting of each of the metrics and methods categorized in each dimension. Furthermore, the proposed methodology for STD and the results obtained must be compared with the methodologies other authors followed to evaluate STD quality.

Another important finding from this work is the lack of a trade-off between the resemblance, utility, and privacy dimensions in STD generated with different approaches in the evaluation section. Thus, further work on improving the STDG approaches to generate more quality STD that maintains an appropriate trade-off between these dimensions is required.

In addition, the proposed evaluation pipeline has been centered on the resemblance, utility, and privacy dimensions, which are the most commonly used dimensions for evaluating the quality of STD for targeted health domain applications found in the literature. Several metrics and methods should also be proposed and developed as part of future work to evaluate the performance of STDG approaches in terms of other dimensions, such as cost (computational resources, training time, and footprint) or diversity (How diverse is the generated STD? Is it biased to any minority class?).

Furthermore, an online toolkit or library could be developed to unify the proposed metrics and methods to help researchers who work on STDG evaluate the generated STD. This tool could help them focus more on improving or proposing STDG approaches without investing time in defining and developing an STD evaluation process.

Conclusion

In this work, we proposed a comprehensive and universal STD evaluation pipeline covering resemblance, utility, and privacy dimensions, with a methodology to categorize the performance of STDG approaches across each dimension. Additionally, we conducted an extensive analysis and evaluation of the proposed STD evaluation pipeline using six different health care-related open-source datasets and four STDG approaches to prove the efficiency and veracity of the proposed STD evaluation pipeline. This analysis has shown that the proposed pipeline can be used effectively to evaluate and benchmark different approaches for STDG, helping the scientific community select the most suitable approaches for their data and application of interest. Although other authors have proposed metrics or methods to evaluate STD, none have defined or used a complete pipeline covering the resemblance, utility, and privacy dimensions in providing different weighting scenarios to get a final score that indicates the quality of generated STD.

Regarding the limitations of this work, we have found that (1) some metrics and methods are not as trustworthy as initially considered, (2) it is difficult to find a perfect trade-off between STD evaluation dimensions of resemblance, utility, and privacy, (3) previously synthesized or anonymized data has been used and (4) the pipeline has not been compared with other methods used in the literature for STD evaluation (as no similar proposal has been identified in the literature).

Future work includes (1) judging and benchmarking the metrics and methods in the proposed pipeline with more datasets, in other contexts, more weighting scenarios and with RD from health care authorities, (2) improving the proposed RIRA in the privacy dimension, (3) proposing new metrics and methods to evaluate the performance of STDG approaches in terms of time and footprint, (4) enhancing the STDG approaches to improve the trade-off between the dimensions, and (5) unifying all the proposed metrics and methods into an easy-to-use online toolkit or library.

Funding

This research was partially funded by the Department of Economic Development and Infrastructure of the Basque Government through Emaitek Plus Action Plan Programme.

Ane Alberdi is part of the Intelligent Systems for Industrial Systems research group of Mondragon Unibertsitatea (IT1676-22), supported by the Department of Education, Universities and Research of the Basque Country.

Conflict of Interest

None declared.

References

1 Rubin DB. Discussion statistical disclosure limitation. J Off Stat 1993;9(02):461–468
2 Little RJA. Statistical Analysis of Masked Data. J Off Stat 1993;9 (02):407–426
3 El Emam K, Hoptroff R. The synthetic data paradigm for using and sharing data. DATA Anal Digit Technol 2019;19(06):12
4 Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. Neurocomputing 2022;493:28–45
5 Hu J. Bayesian estimation of attribute and identification disclosure risks in synthetic data. arXiv preprint arXiv:1804.02784, 2018
6 Reiter JP. New approaches to data dissemination: a glimpse into the future. Chance 2004;17(03):11–15
7 Taub J, Elliot M, Pampaka M, Smith D. Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: Domingo-Ferrer J, Montes F, eds. Privacy in Statistical Databases. Cham: Springer International Publishing; 2018:122–137
8 Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Generation and evaluation of privacy preserving synthetic health data. Neurocomputing 2020;416:244–255
9 Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. Machine learning for healthcare conference 2017: 286–305
10 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002; 16:321–357
11 He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Paper presented at: 2008 IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence; 2008:1322–1328
12 Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 2014;28(01): 92–122
13 Yang F, Yu Z, Liang Y, et al. Grouped Correlational Generative Adversarial Networks for Discrete Electronic Health Records. Paper presented at: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019:906–913
14 Hernandez-Matamoros A, Fujita H, Perez-Meana H. A novel approach to create synthetic biomedical signals using BiRNN. Inf Sci 2020;541:218–241
15 Andreini P, Ciano G, Bonechi S, et al. A Two-Stage GAN for High-Resolution Retinal Image Generation and Segmentation. Electronics (Basel) 2022;11(01):60
16 Porcu S, Floris A, Atzori L. Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems. Electronics (Basel) 2020;9(11):1892
17 Han C, Hayashi H, Rundo L, et al. GAN-based synthetic brain MR image generation. Paper presented at: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018:734–738

18 Stephens M, Estepar RSJ, Ruiz-Cabello J, Arganda-Carreras I, Macía I, López-Linares K. MRI to CTA Translation for Pulmonary Artery Evaluation Using CycleGANs Trained with Unpaired Data. In: Petersen J, San José Estépar R, Schmidt-Richberg A, et al., eds. Thoracic Image Analysis. Cham: Springer International Publishing; 2020:118–129
19 Dahmen J, Cook D. SynSys: a synthetic data generation system for healthcare applications. Sensors (Basel) 2019;19(05):1181
20 Li Z, Ma C, Shi X, Zhang D, Li W, Wu L. TSA-GAN: A Robust Generative Adversarial Networks for Time Series Augmentation. 2021. Paper presented at: International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE; 2021:1–8
21 Che Z, Cheng Y, Zhai S, Sun Z, Liu Y. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. Paper presented at: 2017 IEEE International Conference on Data Mining (ICDM). 2017:787–792
22 Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inform 2020;8(07):e18910
23 Hernandez M, Epelde G, Beristain A, et al. Incorporation of synthetic data generation techniques within a controlled data processing workflow in the health and wellbeing domain. Electronics (Basel) 2022;11(05):812
24 Kotal A, Piplai A, Chukkapalli SSL, Joshi A. PriveTAB: Secure and Privacy-Preserving sharing of Tabular Data. ACM Int Workshop Secur Priv Anal; 2022
25 Bourou S, El Saer A, Velivassaki T-H, Voulkidis A, Zahariadis T. A review of tabular data synthesis using GANs on an IDS dataset. Information (Basel) 2021;12(09):375
26 Piacentino E, Guarner A, Angulo C. Generating Synthetic ECGs Using GANs for Anonymizing Healthcare Data. Electronics (Basel) 2021;10(04):389
27 Hazra D, Byun Y-C. SynSigGAN: generative adversarial networks for synthetic biomedical signal generation. Biology (Basel) 2020;9 (12):441
28 Norgaard S, Saeedi R, Sasani K, Gebremedhin AH. Synthetic Sensor Data Generation for Health Applications: A Supervised Deep Learning Approach. Paper presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018:1164–1167
29 Wang Z, Myles P, Tucker A. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility Patient Privacy. Paper presented at: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). 2019:126–131
30 Beaulieu-Jones BK, Wu ZS, Williams C, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes 2019;12(07):e005122
31 Wang L, Zhang W, He X. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In: Li G, Yang J, Gama J, Natwichai J, Tong Y, eds. Database Systems for Advanced Applications. Cham: Springer International Publishing; 2019:36–52
32 Rashidian S, Wang F, Moffitt R, et al. SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation. In: Michalowski M, Moskovitch R, eds. Artificial Intelligence in Medicine. Cham: Springer International Publishing; 2020:37–48
33 Yoon J, Drumright LN, van der Schaar M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE J Biomed Health Inform 2020;24(08):2378–2388
34 Baowaly MK, Lin C-C, Liu C-L, Chen K-T. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc 2019;26(03):228–241
35 Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020;20(01):108
36 Guan J, Li R, Yu S, Zhang X. Generation of Synthetic Electronic Medical Record Text. Paper presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018:374–380
37 Dash S, Yale A, Guyon I, Bennett KP. Medical Time-Series Data Generation Using Generative Adversarial Networks. In: Michalowski M, Moskovitch R, eds. Artificial Intelligence in Medicine. Cham: Springer International Publishing; 2020: 382–391
38 Chin-Cheong K, Sutter T, Vogt JE. Generation of Heterogeneous Synthetic Electronic Health Records using GANs. ETH Zurich, Institute for Machine Learning; 2019
39 Hittmeir M, Ekelhart A, Mayer R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. Paper presented at: Proceedings of the 14th International Conference on Availability, Reliability and Security. Canterbury CA United Kingdom: ACM; 2019:1–6
40 Giles O, Hosseini K, Mingas G. Faking feature importance: A cautionary tale on the use of differentially-private synthetic data. arXiv preprint arXiv:2203.01363, 2022
41 Dankar FK, Ibrahim MK, Ismail L. A multi-dimensional evaluation of synthetic data generators. IEEE Access 2022;10:11147–11158
42 Hittmeir M, Ekelhart A, Mayer R. Utility and Privacy Assessments of Synthetic Data for Regression Tasks. Paper presented at: 2019 IEEE International Conference on Big Data (Big Data). 2019:5763–5772
43 Platzer M, Reutterer T. Holdout-based empirical assessment of mixed-type synthetic data. Front Big Data 2021;4: 679939
44 Alaa AM, van Breugel B, Saveliev E, van der Schaar M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. International Conference on Machine Learning 2022:290–306
45 Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G, eds. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing; 2019:510–526
46 Wu H, Ning Y, Chakraborty P, Vreeken J, Tatti N, Ramakrishnan N. Generating realistic synthetic population datasets. ACM Trans Knowl Discov Data 2018;12(04):45:1–45:22
47 Fowler EE, Berglund A, Schell MJ, Sellers TA, Eschrich S, Heine J. Empirically-derived synthetic populations to mitigate small sample sizes. J Biomed Inform 2020;105:103408
48 Alqahtani H, Kavakli-Thorne M, Kumar G. Applications of generative adversarial networks (GANs): an updated review. Arch Comput Methods Eng 2021;28(02):525–552
49 Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems; 2019:32
50 Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. Paper presented at: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016:399–410
51 The Synthetic Data Vault. Put synthetic data to work! 2022. Accessed January 24, 2022, at: https://sdv.dev/
52 SYNTHO. 2022. Accessed January 13, 2022, at: https://www.syntho.ai/
53 The Medkit-Learn(ing) Environment. 2022. Accessed January 24, 2022, https://github.com/vanderschaarlab/medkit-learn
54 Build better datasets for AI with synthetic data. 2022. Accessed January 24, 2022, at: https://ydata.ai
55 Lee D, Yu H, Jiang X, et al. Generating sequential electronic health records using dual adversarial autoencoder. J Am Med Inform Assoc 2020;27(09):1411–1419
56 Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proc VLDB Endow 2018;11(10):1071–1083
57 Mendelevitch O, Lesh MD. Fidelity and privacy of synthetic medical data. arXiv preprint arXiv:2101.08658, 2021
58 Ghosheh G, Li J, Zhu T. A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation

measures and data sources. arXiv preprint arXiv:2203.07018, 2022
59 Hernandez M, Epelde G. Synthetic Tabular Data Evaluation Metrics. 2022. Accessed June 1, 2022, at: https://github.com/Vicomtech/STDG-evaluation-metrics
60 Multivariate Distributions – Copulas 0.5.0 documentation. 2022. Accessed March 3, 2021, at: https://sdv.dev/Copulas/tutorials/03_Multivariate_Distributions.html#Gaussian-Multivariate
61 Patki N, Wedge R, Veeramachaneni K. "The Synthetic Data Vault." Paper presented at: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016:399–410. Doi: 10.1109/DSAA.2016.49
62 Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems; 2019, 32
63 Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. Adv Neural Inf Process Syst 2017;30:5767–5777
64 El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Ilustrated. O'Reilly Media, Incorporated; 2020
65 Strack B, DeShazo JP, Gennings C, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int 2014;2014; 781670
66 Ulianova S. Cardiovascular Disease dataset. Kaggle 2018. Accessed January 26, 2021, at: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
67 Palechor FM, Manotas AH. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data Brief 2019;25; 104344
68 Machine Learning Repository UCI. Contraceptive Method Choice Data Set. 2022. Accessed March 14, 2022, at: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice
69 Pima Indians Diabetes Database. 2022. Accessed March 14, 2022, at: https://kaggle.com/uciml/pima-indians-diabetes-database
70 Machine Learning Repository UCI. ILPD (Indian Liver Patient Dataset) Data Set. 2022. Accessed March 14, 2022, at: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

Appendix A: Abbreviations list

Abbreviation	Definition
AI	Artificial Intelligence
AIA	Attribute Inference Attack
AUC	Area Under the Curve
BN	Bayesian Network
CMC	Contraceptive Method Choice
CMEM	Categorical Maximum Entropy Model
CTGAN	Conditional Tabular Generative Adversarial Network
CVD	CardioVascular Disease
DCR	Distance to the Closest Record
DL	Deep Learning
DLA	Data Labelling Analysis
DRA	Dimensionality Reduction Analysis
DT	Decision Tree
EHR	Electronic Health Records
GAN	Generative Adversarial Network
GM	Gaussian Multivariate
ILP	Indian Liver Patient
JSD	Jensen-Shannon Divergence
KNN	K-Nearest Neighbors
KS	Kolmogorov-Smirnov
MAE	Mean Absolute Error
MIA	Membership Inference Attack
MKDE	Movement-based Kernel Density Estimation
ML	Machine Learning
MLP	MultiLayer Perceptron
MRA	Multivariate Relationship Analysis
PCA	Principal Component Analysis
PID	Pima Indians Diabetes
PPC	Pairwise Pearson Correlation
QID	Quasi-IDentifiers
RD	Real Data
RF	Random Forest
RIRA	Re-Identification Risk Analysis
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RTS	Real To Synthetic
SD	Synthetic Data
SDG	Synthetic Data Generation
SDV	Synthetic Data Vault
SEA	Similarity Evaluation Analysis
STD	Synthetic Tabular Data
STDG	Synthetic Tabular Data Generation

(*Continued*)

(*Continued*)

Abbreviation	Definition
STR	Synthetic To Real
SVM	Support Vector Machines
TRTR	Train on Real Test on Real
TSTR	Train on Synthetic Test on Real
URA	Univariate Resemblance Analysis
WGAN	Wasserstein Generative Adversarial Network
WGANGP	Wasserstein Generative Adversarial Network with Gradient Penalty

Appendix B: Evaluation metrics and methods not included in the Final Methodology

Appendix B.1. Visual analysis of individual attributes

An extra method for the URA, is proposed to visually compare the values of each attribute (RD vs. STD). For numerical and categorical attributes, distribution plots and histograms can be used. The STDG approaches have been categorized depending on the number of attributes that maintained resemblance of the RD attributes. Since this visualization is subjective and cannot be given an objective evaluation value it has been deleted from the proposed evaluation methodology. However, the results of applying this method can be seen in the **Supplementary Material 1**, available in the online version only, available in the online version only).

Appendix B.2. Dimensionality resemblance analysis (DRA)

To analyze if the dimensional properties of RD are preserved in STD, it is proposed to analyze the performance of a linear, principal component analysis (PCA), and a non-linear, Isomap, dimensionality reduction method for RD and STD. The transformation should be independently computed for RD and STD for the two methods after scaling the numerical attributes and one-hot encoding the categorical attributes. After computing the transformations, the results can be visually analyzed with scatter plots for each method. The more similar the shapes of RD and STD plots are, the more resemblance is maintained.

Additionally, a distance metric between RD and STD dimensionality reduction plots is proposed to calculate a numerical value from the visual results. This distance metric is the joint distance of the barycenter distance and spread distance of both plots (Equation 10). The barycenter distance is the distance between the mean values of the RD and STD dimensionality reduction matrices, while the spread distance is the distance between the standard deviation values of the same matrices. α is a regularization parameter that gives different weights to each distance. In the equation μ is the mean value and σ is the standard deviation or RD (r) and STD (s).

$$joint\_dist_{(r,s)} = \propto \cdot \sqrt{\mu_{(r)}{}^2 + \mu_{(s)}{}^2} + (1-\alpha) \cdot \sqrt{\sigma_{(r)}{}^2 + \sigma_{(s)}{}^2}$$

Equation 10: Joint Distance of dimensionality reduction plots

In these experiments, $\alpha = 0.05$ has been chosen experimentally, as the barycenter distance using PCA and Isomap is very low since the barycenters are calculated similarly in both dimensionality reduction methods. As this distance metric cannot be normalized, it is not possible to define a methodology to quantitatively classify the resemblance of STD generated with one or more STDG approaches into "*Excellent*," "*Good*," and "*Poor*," although the lower this distance value is, the more similar the dimensionality reduction plots for RD and STD are. For this reason, this analysis has not been considered for the total resemblance calculation, but the results of both the plots and the distance metric are provided and discussed in the **Supplementary Material 1**, available in the online version only).