Standardised Metrics and Methods for Synthetic Tabular Data Evaluation

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

13-09-2021 / 21-09-2021

CITATION

Hernandez, Mikel; Epelde, Gorka; Alberdi, Ane; Cilla, Rodrigo; Rankin, Debbie (2021): Standardised Metrics and Methods for Synthetic Tabular Data Evaluation. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.16610896.v1>

DOI

[10.36227/techrxiv.16610896.v1](https://doi.org/10.36227/techrxiv.16610896.v1)

Standardised Metrics and Methods for Synthetic Tabular Data Evaluation

Mikel Hernandez*, Gorka Epelde, Ane Alberdi, Rodrigo Cilla and Debbie Rankin

Abstract—Synthetic Tabular Data Generation (STDG) is a potentially valuable technology with great promise to augment real data and preserve privacy. However, prior to adoption, an empirical assessment of synthetic tabular data (STD) is required across the three dimensions of resemblance, utility, and privacy, trying to find a trade-off between them. A lack of standardised and objective metrics and methods has been found targeting this assessment in the literature and neither an organised pipeline or process for coordinating this evaluation has been identified. Therefore, in this work we propose a collection of metrics and methods to evaluate STD in the previously defined dimensions, presenting a meaningful orchestration of them and a pipeline unifying all of them. Additionally, we present a methodology to categorise STDG approaches performance for each dimension. Finally, we conducted an extensive analysis and evaluation to verify the usability of the proposed pipeline across six healthcare-related datasets, using four STDG approaches. The results of these analyses showed that the proposed pipeline can effectively be used to evaluate and benchmark the STD generated with one or more different STDG approaches, helping the scientific community to select the most suitable approaches for their data and application of interest.

Index Terms—synthetic tabular data generation, synthetic tabular data evaluation, resemblance evaluation, utility evaluation, privacy evaluation.

1 INTRODUCTION

WE are in an era where the amount of data being generated are ever-increasing, motivating a paradigm shift from traditional and manual processes towards artificial intelligence (AI) applications in different contexts. However, many AI developments are being slowed down by data protection and regulation laws or imbalanced data. Therefore, research on synthetic data (SD), which is 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) [1], has gained importance in recent years. Since SD does not contain data from the original real data (RD) set and less information is lost compared to other anonymisation or data balancing techniques, synthetic data generation (SDG) has shown great promise to (1) augment RD, balancing datasets or supplementing the available data to train predictive models [2] and (2) preserve privacy to enable secure and private data sharing [3].

Within healthcare, SDG has been researched for different modalities including biomedical signals [4], medical images [5], electronic health records (EHR) free-text content [6], time-series smart-home activity data [7] and EHR tabular data [8]. In this study we have focused on synthetic tabular data generation (STDG), as it is the largest exponent of structured data for developing machine learning (ML) models

for healthcare decision making. Therefore, tabular health data potentially offers the most valuable opportunities in the development of AI-based health care systems.

Although STDG is a potentially valuable technology, prior to adoption, an empirical assessment of synthetic tabular data (STD) generated with various approaches is required across the three dimensions of resemblance, utility, and privacy. The resemblance dimension evaluates how well SD represents RD. The utility dimension evaluates the usability of the 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. A lack of standardised and objective metrics for the evaluation of STD dimensions has been noticed in the literature. Thus, an organised pipeline or process is required for the assessment of these three dimensions to enable the selection of the best approach or approaches to generate the desired STD.

In this paper, we propose a collection of standardised metrics and methods for the evaluation of resemblance, utility, and privacy dimensions of healthcare-related STD. The proposed metrics and methods were selected from literature and configured in a standardised evaluation pipeline, to enable a guided and comparable evaluation. The proposed approach can be applied to any kind of STD. Our contributions can be summarised as follows:

- M. Hernandez, G. Epelde and R. Cilla are with the Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009, Donostia-San Sebastián, Spain. E-mail: {mhernandez, gepelde, rcilla}@vicomtech.org
- A. Alberdi is with Mondragon Unibertsitatea, Loramendi Kalea 4, 20500, Arrasate-Mondragón, Spain. E-mail: aalberdi@mondragon.edu
- D. Rankin is with School of Computing, Engineering and Intelligent Systems, Ulster University, Derry-Londonderry, United Kingdom. E-mail: d.rankin1@ulster.ac.uk

- (1) We propose a set of metrics and methods to evaluate STD in the previously defined dimensions (resemblance, utility, and privacy) presenting a meaningful orchestration of them. For each dimension, different universal metrics and methods are suggested for independent use.
- (2) To the best of our knowledge, this work is the first attempt to propose a complete and universal

STD evaluation pipeline covering resemblance, utility, and privacy dimensions. Additionally, we present a methodology to categorise the performance of each STDG approach in each dimension as “Excellent”, “Good” and “Poor”.

- (3) We conduct an extensive analysis and evaluation of the suggested pipeline using six different healthcare-related open-source datasets and four STDG approaches. The proposed evaluation metrics and methods are used to evaluate the STD generated for each dataset and STDG approach combination.
- (4) Based on the obtained 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 to select the most suitable approaches for their data and application of interest.

The remainder of this article is organised as follows: Section 2 presents related work in STD generation and evaluation for the healthcare context. Section 3 describes the proposed pipeline of metrics and methods for STD evaluation. Section 4 presents the experimental results, and they are discussed in Section 5. Section 6 states the conclusions of this work. Additionally, in the Supplemental Material a more extensive description of the results can be found.

2 RELATED WORK

This section presents some of the existing STDG approaches for tabular healthcare data (Section 2.1), and the most commonly used metrics and methods for evaluating STD dimensions (Section 2.2).

2.1 STDG Approaches

To generate STD in the healthcare context many approaches can be found in the literature. The simplest STDG approaches include Gaussian Multivariate (GM) [8], [9], Bayesian Networks (BN) [3], [10], [11], Categorical maximum entropy model (CMEM) [12] and Movement-based kernel density estimation (MKDE) [13] can be found. These approaches employ a statistical model to learn the multivariate distributions of the RD to sample a set of STD. These approaches are typically used for small amounts of data and are not very scalable.

Due to the efficiency and popularity of the Generative Adversarial Networks (GAN) in other areas and applications of healthcare, there is an interest in finding out if they have promising potential to synthesise tabular healthcare data. GANs principally consists of two neural networks (a generator and a discriminator) that learn to generate high quality STD through an adversarial training process. This approach has been used by several different authors, presenting improvements, tuning some hyper-parameters, or adding new features. While ehrGAN [2], [14] and medGAN [2], [8], [15], [16], [17] were proposed to synthesise mainly numerical and binary data, Wasserstein GAN (WGAN) with Gradient Penalty [2], [8], [16], [17], [18], [19], healthGAN

[8] and Conditional Tabular GAN (CTGAN) [20] synthesise numerical, binary and categorical data efficiently.

Furthermore, the Synthetic Data Vault (STDV) consists of an ensemble approach which combines several probabilistic graphical modelling and Deep Learning (DL) based techniques [21].

2.2 STD evaluation metrics and methods

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

2.2.1 Resemblance evaluation

The first step in resemblance evaluation is to analyse whether the distribution of STD attributes is equivalent to the distribution of the RD. Che et al. [14] and Chin-Cheong et al. [19] compared the distributions of the attributes of RD against STD. Yang et al. [2] compared the frequency of the attributes. Additionally, Choi et al. [15], Wang et al. [22], Abay et al. [11], Baowaly et al. [16] and Yale et al. [8] compared the dimensional probability or probability distributions of RD and STD.

For distributions comparison Yang et al. [2] and Rashidian et al. [18] analysed the mean absolute error (MAE) between mean and standard deviation values of RD and STD. Some authors also use statistical tests to analyse univariate resemblance of STD. Baowaly et al. [16] used Kolmogorov-Smirnov (KS) tests to compare distributions, Dash et al. [23] applied Welsch t-tests to compare mean values of the attributes, and Yoon et al. [17] performed Student t-tests to compare mean values of the attributes and Chi-squared tests to compare the independence of categorical attributes. In these studies, they analysed the p -values obtained from the statistical tests to determine whether the STD attributes preserved the properties of the RD attributes.

In the evaluation of multivariate relationships, Rankin et al. [3], Yale et al. [8], Wang et al. [22] and Rashidian et al. [18] visually compared the Pairwise Pearson Correlation (PPC) matrices to see whether correlations between attributes of RD are maintained in STD. Additionally, Principal Component Analysis (PCA) transformation has been used by Yale et al. [8] to compare the dimensional properties of STD and RD.

To analyse if the semantics or significance from RD is maintained in STD, Choi et al. [15], Wang et al. [22], Beaulieu-Jones et al. [24] and Lee et al. [25] asked some clinical experts to evaluate STD qualitatively giving a score between 1 and 10. This score indicated how real the STD records appear to them where a score of 10 is most realistic. Another method that can be used if access to clinical experts is not available is to train some ML classifiers to label records as real or synthetic as Lee et al. [25] proposed in their study.

2.2.2 Utility evaluation

The evaluation of the utility dimension has mainly been performed using STD in ML models, by training and analysing 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. [26], Wang et al. [22], Beaulieu-Jones et al. [24], Chin-Cheong et al. [19], Baowaly et al. [16] and Rashidian et al. [18]. These authors trained ML models with RD and with STD separately and then tested them with held-out RD not used for training. They use different classification metrics (Accuracy, F1-score, ROC, AUC-ROC, etc) to evaluate and analyse the differences in the models' performance when training the models with RD and with STD.

On the other hand, Che et al. [14], Wang et al. [22] and Yang et al. [2] augmented the training set of RD with STD. The authors analysed whether the trained ML models have a slight difference when training only with RD or when training with mixed RD and STD.

2.2.3 Privacy evaluation

The few metrics and methods that 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. [26] used distance to the closest record (DCR) 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., [27] the maximum real to synthetic (RTS) similarity value, computed by the cosine similarity, indicates whether the model has memorised and stored RD and is really generating data and not copying it. Other distance-based metrics, used by Yoon et al. [17], are the Jensen-Shannon divergence (JSTD) and the Wasserstein distance. The authors used them to compute the balance between the identifiability and quality of STD.

To assess the re-identification of RD disclosure risk through STD, several disclosure simulation attacks have been proposed in the literature. On the one hand, Choi et al. [15], Park et al. [26], Yale et al. [28] and Mendelevitch et al. [29] simulated membership inference attacks to analyse 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. On the other hand, Choi et al. [15] and Mendelevitch et al. [29] additionally simulated attribute inference attacks to quantify the disclosure risk of some attributes of the dataset. Defining quasi-identifiers (QID) attributes and training some ML models with STD to predict the rest of the attributes, they analyse how accurately an attacker could predict some RD attributes if they obtained access to STD.

3 STD EVALUATION METRICS AND METHODS

The proposed metrics and methods for the evaluation of STD can be clustered into three different dimensions: resemblance, utility, and privacy. Within each dimension, different metrics and methods from the literature have been selected and configured in an organised way. Figure 1 shows a 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. This pipeline was developed due to the lack of a complete STD evaluation method that covers resemblance, utility, and privacy dimensions. Although the included metrics and methods are not new, the orchestration of them in an organised way and the calculation of overall scores for each dimension are the novelty of this study. Furthermore, to the best of our knowledge this work is the first attempt to propose a complete and universal STD evaluation process that is generalisable to any kind of STD, due to the fact that the metrics and methods have been selected according to the most used and standardised STD evaluation metrics and methods reported in the literature.

3.1 Resemblance evaluation

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

3.1.1 Univariate Resemblance Analysis (URA)

This analysis proposes to analyse the attributes of RD and STD independently to see if 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 (h_0) is accepted. Otherwise, the alternative hypothesis (h_1) is accepted. The properties analysed in each test are preserved in STD if h_0 is accepted. For numerical attributes, the following tests are proposed:

- Student T-test for the comparison of means.
 - h_0 : Means of RD feature and STD attribute are equal.
 - h_1 : Means of RD feature and STD attribute are different.
- Mann Whitney U-Test for population comparison.
 - h_0 : RD feature and STD attribute come from the same population.
 - h_1 : RD feature and STD attribute do not come from the same population.
- Kolmogorov-Smirnov test for distributions comparison.
 - h_0 : RD feature distribution and STD attribute distribution are equal.
 - h_1 : RD feature distribution and STD attribute distribution are not equal.

For categorical features, the Chi-Squared test (χ^2 test) is proposed to analyse the feature independence between real and synthetic categorical attributes. In this case, if the h_0 is

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

accepted, statistical properties are not preserved. h_0 and h_1 are defined as:

- h_0 : There is not a statistical relationship between the real categorical variable and the synthetic categorical variable.
- h_1 : There is a statistical relationship between the real categorical variable and the synthetic categorical variable.

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. In total, 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)-(5), r is the attribute of RD and s is the attribute of STD.

Cosine distance is defined using the cosine similarity, which is the cosine of the angle between two n-dimensional vectors in n-dimensional space; the dot product of the two vectors is divided by the product of the two vectors' lengths (Equation 1). Using this distance, a threshold of 0.3 has been proposed to indicate that the STD attribute resembles RD attribute.

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

Jensen-Shannon distance is the square root of the Jensen-Shannon divergence, 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). To compute these distances the probability distributions of the features have been used: p is the probability distribution of the RD attribute and q is the probability distribution of the STD attribute. Thus, a value lower than 0.1 represents perfect resemblance.

$$js_dist(p, q) = \sqrt{\frac{D(p \parallel m) + D(p \parallel m)}{2}} \quad (2)$$

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (3)$$

Wasserstein distance can be seen as the minimum amount of work required to transform a vector (r) into another vector (s), where the work 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.

$$was_dist(r, s) = \int_{-\infty}^{+\infty} |R - S| \quad (4)$$

For both statistical tests and distance calculation, the number of STD attributes that fulfil the requisite to resemble RD attributes has been considered to categorise 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 attributes fulfilled it, as "Poor".

Finally, in the last step for URA, it 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, respectively. The STDG approaches have been categorised depending on the number of attributes that maintained resemblance of RD attributes.

To get an overall score of URA, the performances for all previously presented metrics can be calculated according to the results obtained from them. For that, first the categorisation is translated to a numerical value ("*Excellent*" = 3, "*Good*" = 2 and "*Poor*" = 1), and then, the same weight (0.25) is given for the four methods presented (statistical tests of numerical attributes, statistical tests of categorical attributes, distance calculations and visual comparisons). The resulting score, rounded to the nearest integer, gives a value between 1 to 3 that indicates the URA score of STD.

3.1.2 Multivariate Relationship Analysis (MRA)

This analysis suggests to analyse 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: Pairwise Pearson Correlation (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 normalised contingency tables for categorical variables. The correlation matrices of RD can be visually compared with the matrices of STD using heatmaps.

$$PPC(x, y) = \frac{\sum_{i=1}^n (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^n (x_i - \hat{x})^2} \sqrt{\sum_{i=1}^n (y_i - \hat{y})^2}} \quad (5)$$

Additionally, for each matrix, the differences between the correlations of RD and STD are calculated, to then compute the percentage of relationships maintained in STD (those that have a difference value lower than 0.1). This way, two values between 0 and 1 are obtained, one expressing the percentage of numerical attributes relationships maintained, and the other the percentage of categorical attributes relationships. If the values are higher than 0.6, the STDG approach is categorised 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 categorised 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 the half of the relationships. After doing this categorisation for each matrix, a total MRA performance is obtained by calculating the mean performance of both analyses.

3.1.3 Dimensional Resemblance Analysis (DRA)

To analyse if the dimensional properties of RD are preserved in STD, it is proposed to analyse the performance of a linear (PCA) and a non-linear (Isomap) dimensionality reduction method for RD and STD. For the two methods, the transformation should be independently computed for RD and STD, after scaling the numerical attributes and one-hot encoding the categorical attributes. After computing the transformations, the results can be visually analysed with scatter plots for each method, the more similar the shapes of RD and STD plots the more resemblance is maintained. Figure 2 shows some examples that indicate "*Excellent*", "*Good*" and "*Poor*" resemblance respectively.

Fig. 2. DRA Categorisation Example. (a) Result for Excellent resemblance. (b) Result for Good resemblance. (c) Result for Poor Resemblance.

Additionally, to enable the calculation of a numerical value from the visual results, a distance metric between RD and STD dimensionality reduction plots is proposed. This distance metric is the joint distance of the baricenters distance and spread distance of both plots (Equation 6). The baricenters 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 regularisation parameter that gives different weights to each of the two distances. In the developed experiments a α value of 0.05 has been experimentally chosen as the baricenters distance using PCA and Isomap is very low since in both dimensionality reduction methods the baricenters are calculated in a similar way. As this distance metric cannot be normalised, 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 are the dimensionality reduction plots for RD and STD. For this reason, this analysis has not been considered for total resemblance calculation, but the results of both the plots and the distance metric are provided and discussed in the supplemental material.

$$joint_dist = \alpha \cdot bar_dist + (1 - \alpha) \cdot spread_dist \quad (6)$$

3.1.4 Data Labelling Analysis (DLA)

The final step proposed for resemblance evaluation can be used to evaluate the semantics of STD. For this method, it is proposed to analyse the performance of some classifiers when labelling records as real or synthetic. Firstly, real and synthetic datasets should be mixed and labelled (0 for RD and 1 for STD) in a single dataset. Secondly, this dataset must be split into train and test sets, i.e., 80:20 split. Thirdly, the train and test data should be pre-processed, i.e., standardize numerical attributes and one-hot encode categorical attributes. Finally, some ML classifiers can be trained (with training data) and evaluated (with test data) to analyse the performance of a classifier to label records as real or synthetic. The following commonly used and diverse ML classifiers are proposed to be used in these analyses with the specified parameter values:

- RF ($n_estimators = 100$, $random_state = 9$, $verbose = True$, $n_jobs = 3$)
- KNN ($n_neighbors = 10$, $n_jobs = 3$)
- DT ($random_state = 9$)

- SVM ($C = 100$, $max_iter = 300$, $kernel = "linear"$, $probability = True$, $random_state = 9$, $verbose = 1$)
- MLP ($hidden_layer_sizes = (128, 64, 32)$, $max_iter = 300$, $random_state = 9$, $verbose = 1$)

After training and testing the models, classification performance metrics (accuracy, precision, recall and F1-score) can be analysed 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 indicates "Good" resemblance, while values equal to or greater than 0.8 indicates "Poor" resemblance.

3.1.5 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 to give more or less importance to each result based on the researchers' STDG goals. Based on our experimental work, DRA has been excluded from the total resemblance score due to the difficulty of objectively comparing and establishing thresholds for classifying proposed quantitative metric results into defined quality levels. 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 prioritise 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, total score is categorised as "Excellent" if the resulting value is 3, "Good" if the resulting values is 2 or "Poor" if the resulting value is 1.

3.2 Utility evaluation

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

Before training and testing the models, the data must be pre-processed, i.e., standardize numerical attributes and one-hot encode 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 analyse the classification results accuracy, precision, recall and F1-score classification metrics are proposed, as well as their absolute differences when TRTR and TSTR. To assure that STD utility is "Excellent", the metrics differences should not exceed a proposed threshold of 0.4. If differences between 0.4 and 0.8 are obtained the performance in the utility dimension should be categorised as "Good" and if they are higher than 0.8 the performance is "Poor".

3.3 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.

3.3.1 Similarity Evaluation Analysis (SEA)

In the SEA it is proposed to evaluate how private STD is when compared to RD. 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 defined as the square root of the sum of squares of differences between RD features and STD features, as defined in Equation 7. 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 analysed. 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 or equal 0.3 indicates that privacy is preserved.

$$euc_dist(r, s) = \sqrt{\sum_{i=1}^n (r_i - s_i)^2} \quad (7)$$

To compute the STR similarity, the cosine similarity metric is proposed, which computes similarity as the normalised dot product of two datasets (Equation 8; 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 analysed. If the mean value is higher than 0.5 the STD is very close to the RD, so the privacy is not preserved. In all other cases, it can be said that privacy is preserved.

$$cosine_similarity(S, R) = \frac{\sum_{i=1}^n s_i \cdot r_i}{\sqrt{\sum_{i=1}^n s_i^2} \cdot \sqrt{\sum_{i=1}^n r_i^2}} \quad (8)$$

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 9; R is the real dataset; S is the synthetic dataset and $d(s_i, r_i)$ is a metric between points s_i and r_i , in this case, the Euclidean distance). 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 STD is far from RD. Since this metric is not bounded between 0 and 1, a value higher than 1 has been considered to assure that the privacy is preserved.

$$haus_dist(S, R) = \max\{h(S, R), h(R, S)\} \quad (9)$$

For these three distances-based metrics, if all three metrics fulfil the condition to preserve privacy the categorisation for privacy preservation is "Excellent". Otherwise, if two or 1 of the metrics fulfil the condition it is "Good" and in other cases it is "Poor".

3.3.2 Re-Identification Risk Analysis (RIRA)

In this analysis, it is proposed 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) membership inference attack (MIA) when an attacker tries to identify if real patient records have been used to train the STDG approach (Figure 3), 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 (Figure 4) [29].

3.3.2.1 Membership Inference Attack: (MIA)

In a 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. Figure 3 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 the RD in the STD, meaning that r has been used to generate STD. The success rate of the attacker is proposed to be evaluated by simulating this kind of attack by calculating the Hamming distance (proportion of non-equal attributes between two records) between each row of the RD subset and STD rows. Reasonable thresholds to assure that r is close enough to a record in STD are 0.4, 0.3, 0.2 and 0.1. 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 that was identified as such by the attacker) values have been calculated. Accuracy and precision values should be 0.5 or lower for all thresholds to obtain an “Excellent” privacy preservation categorisation. Any value above 0.5, indicates increasing levels of disclosure risk, obtaining a “Good” or “Poor” privacy preservation categorisation depending on the number of thresholds where the values are higher than 0.5. To analyse and interpret these results, the accuracy and precision values have been plotted as a function of the proportion of records in the STD that are present in the RD subset known by the attacker for each threshold.

Fig. 3. Simulated MIA. Inspired from [29].

3.3.2.2 Attribute Inference Attack: (AIA)

In an AIA an attacker has access to the STD and a subset of attributes for some RD records (generally QID such as

age, gender, height, weight, etc). As shown in Figure 4, the attacker will use ML models trained with STD to predict the values of the rest of the attributes of the RD records. The success of this attack can be measured defining the QID of each dataset and then using them to train ML models, i.e., Decision Tree (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, whilst lower RMSE values (close to 0) reflect higher disclosure risk. The metric values obtained for each QID combination, and all attributes are evaluated visually with a box-plot for each risk attribute. Additionally, the percentage of correctly predicted attributes has been calculated to categorise the re-identification risk. The mode of the analysed metrics has been considered to determine if an attribute has been predicted; for categorical attributes accuracy mode equal to 1 and for numerical attributes RMSE mode equal to 0. If a high percentage (higher than 0.6) is obtained, as more than half of the attributes has been re-identified the AIA results is categorised as “Poor”. For a percentage between 0.4 and 0.6, the result is categorised as “Good” since approximately the 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.

Fig. 4. Simulated AIA. Inspired from [29].

3.3.3 Total privacy performance

After computing 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 (section 4), 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). Higher weight was assigned to RIRA, to prioritise the metrics analysing if RD could be inferred from STD, over the metric investigating how similar is STD to RD. Finally, once weighting is applied, and after the result's rounding to the nearest integer is performed, the privacy of STD is categorised as “Excellent” (3), “Good” (2) or “Poor”

TABLE 1
Brief description of the selected datasets

ID	Dataset name	Year	Data type	Num. attrib.	Num. records
A	Diabetes 130-US hospitals for years 1999-2008 dataset [30]	2014	Num. Cat.	55	101766
B	Cardiovascular disease dataset [31]	2019	Num. Cat.	13	70000
C	Estimation of obesity levels based on eating habits and physical condition dataset [32]	2019	Num. Cat.	17	2111
D	Contraceptive Method Choice dataset [33]	1997	Num. Cat.	9	1473
E	Pima Indians Diabetes Database [34]	2016	Num. Cat.	9	769
F	Indian Liver Patient dataset [35]	2012	Num. Cat.	11	583

(1) based on the total score obtained when applying these weights.

4 STD EVALUATION

To prove and trust the efficiency and usability of the proposed pipeline for STD evaluation, a few different datasets have been selected and then synthesised 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 repository¹.

4.1 Selected data

Six open-source healthcare-related datasets have been selected for synthesis. A brief description of these datasets, together with the data types and the number of attributes and records is presented in Table 1. Each dataset has firstly been pre-processed, deleting missing values and performing a data split into two subsets. 80% of the records have been used for training the STDG approaches and 20% of the records for utility dimension evaluation and RIRA simulations.

4.2 STDG approaches

To generate STD using the previously presented datasets four STDG approaches have been used, two of which are GANs and the other two are classical approaches. These approaches are as follows:

4.2.1 GM

A classical STDG approach based on statistical modelling that implements a multivariate distribution by using a Gaussian Copula to combine marginal probabilities estimated using univariate distributions² [36].

4.2.2 SDV

This approach is a STDG ecosystem of libraries that uses several probabilistic graphical modelling and DL-based techniques. To enable a variety of data storage structures, they employ unique hierarchical generative modelling and recursive sampling techniques³ [37].

4.2.3 CTGAN

A STDG approach proposed by Xu et al. in 2019 [20] defined as a collection of DL models based on GAN models for a single data table. It can learn from real data and generate synthetic clones with high fidelity⁴.

4.2.4 WGANGP

This approach is a GAN proposed by Yale et al. in 2020 [8] 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 optimisation function⁵.

4.3 Results

Using the datasets and STDG approaches described in Sections 4.1 and 4.2, the metrics and methods proposed in Section 3 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, a comparison and benchmarking of the STDG approaches has been performed based on the proposed strategy for STD evaluation.

Table 2 shows the results obtained from applying the proposed STD evaluation pipeline to the STD synthesised with each STDG approach and for each dataset. The detailed and complete description of all results are available in the Supplementary Material. Based on the results of each evaluation analysis presented in the pipeline, the performance of each synthesised dataset has been categorised as “Excellent”, “Good” or “Poor” for each dimension (resemblance, utility, and privacy).

The resemblance dimension has been perfectly maintained with GM for the six datasets. SDV has performed excellently on the resemblance dimension for the half of the datasets (C, E, F) and well for the 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 to resemble RD, as it has resembled RD good for four datasets (A, B, C and D) and poorly for the other two (E and F). Despite the poor resemblance for the last two datasets (E and F), by applying the proposed metrics and methods for resemblance evaluation it can be assured that resemblance has been maintained for most of the datasets using the four STDG approaches.

The utility dimension has been perfectly maintained with all approaches for five datasets (A, B, D, E and F). For dataset C, the utility of the STD has been poorly maintained with SDV and CTGAN, whilst with GM and WGANGP it

1. <https://github.com/Vicomtech/STDG-evaluation-metrics>

2. <https://github.com/SDV-dev/Copulas>

3. <https://github.com/SDV-dev/SDV>

4. <https://github.com/SDV-dev/CTGAN>

5. <https://github.com/ydataai/ydata-synthetic>

TABLE 2
STD Evaluation results

Dataset	Num. records	Num. attrib.	STDG Approaches	Resemblance	Utility	Privacy
A - Diabetes	101766 (81412 for training and 20354 for testing)	20 (10 numerical and 10 categorical)	GM	(3) Excellent	(3) Excellent	(2) Good
			SDV	(2) Good	(3) Excellent	(2) Good
			CTGAN	(2) Good	(3) Excellent	(3) Excellent
			WGANGP	(2) Good	(3) Excellent	(3) Excellent
B - Cardiovascular Disease	70000 (56000 for training and 14000 for testing)	13 (6 numerical and 7 categorical)	GM	(3) Excellent	(3) Excellent	(2) Good
			SDV	(2) Good	(3) Excellent	(2) Good
			CTGAN	(3) Excellent	(3) Excellent	(2) Good
			WGANGP	(2) Good	(3) Excellent	(2) Good
C - Obesity level estimation	2111 (1688 for training and 423 for testing)	17 (8 numerical and 9 categorical)	GM	(3) Excellent	(2) Good	(2) Good
			SDV	(3) Excellent	(1) Poor	(3) Excellent
			CTGAN	(2) Good	(1) Poor	(3) Excellent
			WGANGP	(2) Good	(2) Good	(2) Good
D - Contraceptive Method Choice	1473 (1178 for training and 295 for testing)	10 (2 numerical and 8 categorical)	GM	(3) Excellent	(3) Excellent	(2) Good
			SDV	(2) Good	(3) Excellent	(3) Excellent
			CTGAN	(2) Good	(3) Excellent	(3) Excellent
			WGANGP	(2) Good	(3) Excellent	(3) Excellent
E - Pima Indians Diabetes	769 (616 for training and 153 for testing)	9 (8 numerical and 1 categorical)	GM	(3) Excellent	(3) Excellent	(2) Good
			SDV	(3) Excellent	(3) Excellent	(2) Good
			CTGAN	(1) Poor	(3) Excellent	(2) Good
			WGANGP	(1) Poor	(3) Excellent	(2) Good
F - Indian Liver Patient	583 (467 for training and 116 for testing)	11 (9 numerical and 2 categorical)	GM	(3) Excellent	(3) Excellent	(2) Good
			SDV	(3) Excellent	(3) Excellent	(3) Excellent
			CTGAN	(2) Good	(3) Excellent	(3) Excellent
			WGANGP	(1) Poor	(3) Excellent	(2) Good

has been maintained but not perfectly. From these results, it can be assured that using the proposed methods for utility evaluation, the utility of STD has been perfectly maintained for all cases, except for one dataset (C).

The privacy of STD has been quite well maintained with GM for all the datasets. With SDV privacy has been perfectly maintained with three datasets (C, D and F) and quite well for the other three datasets (A, B and E). With CTGAN privacy has also been perfectly maintained for three datasets (A, C and F) and quite well maintained for the other three datasets (B, D and E). With WGANGP, privacy has only been perfectly maintained for two datasets (A and D) and quite well for the other four datasets (B, C, E and F). Therefore, by applying the proposed privacy dimension evaluation metrics and methods the STD generated with GM have yielded to a better privacy preservation than for the other approaches. However, with the other approaches privacy has been quite well preserved, offering better preservation for certain dataset and STDG approach combinations.

5 DISCUSSION

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. In the experiments carried out, none of the STDG approaches have been better than any other across all STD dimensions considered. Therefore, it can be said that it is difficult to get a trade-off between resemblance, privacy, and utility scores. However, the categorisation system provided and overall score calculation for each dimension can help to select the most appropriate STDG approaches, by looking at individual dimension scores and configuring overall scores by computing weights, according to priorities defined for each specific application.

A collection of 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, as statistical tests and distance calculations provide quantitative results, these 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 to analyse whether the multivariate relationships between attributes and how many of them are maintained in STD. Although DRA has not been included in the total resemblance calculation due to the difficulty to interpret the obtained results it can be useful to see how well the dimensional properties of RD are preserved. DLA has been less effective due to the lack of medical specialists to qualitatively evaluate generated data's significance. However, the analysis developed composed of different ML models trained to label records as real or synthetic has simplified this process giving an approximation of how a medical expert would label the records.

In terms of the utility dimension TRTR and TSTR methods have been proposed, where a few ML classification models have been trained with RD, and separately with STD. This methodology has been very useful to evaluate whether STD could be used instead of RD for data modelling. Further work in this area may include a ML model spot-check to select the best ML models for the specific application and the use of mixed data for models training in contrast to separate use of RD or STD. Furthermore, some other data modelling tasks can be proposed to assess the utility of STD in the same way, e.g., regression or clustering.

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 in a more reliable way. Thus, future work might include the definition of a strategy that helps identify which attributes are more prone to re-identification and considers the real consequences of potential personal data disclosure.

Although, the proposed STD evaluation pipeline has been used to evaluate some STD generated with different approaches and in different 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 anonymised or synthesised 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 with RD that comes directly from hospitals or laboratories without any anonymisation, or other modification processes applied to them. Furthermore, the proposed methodology for STD and the obtained results must be compared with the methodologies that other authors followed to evaluate STD quality.

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

Apart from that, several metrics and methods should also be proposed and developed to evaluate the performance of the STDG approaches in terms of training time and footprint. This way, STDG approaches could be compared analysing the computational resources needed. Furthermore, a web-based application can be developed unifying the proposed metrics and methods, to help researchers who work on STDG to evaluate the generated STD. This tool could help them to be more focused on improving or proposing STDG approaches without having to invest time in defining and developing a STD evaluation process.

6 CONCLUSION AND FUTURE WORK

In this work, we proposed a comprehensive and universal STD evaluation pipeline covering resemblance, utility, and privacy dimensions, with a methodology to categorise 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 healthcare-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 effectively be used to evaluate and benchmark different approaches for STDG, helping the scientific community to select the most suitable

approaches for their data and application of interest. Although other authors have proposed metrics or methods to evaluate STD, none of them have used a complete pipeline that covers all dimensions of STD evaluation.

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 trade-off between STD evaluation dimensions of resemblance, utility, and privacy, (3) previously synthesised or anonymised 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) to judge and benchmark the metrics and methods in the proposed pipeline with more datasets, in other contexts and with real data from health-care authorities, (2) to improve the proposed RIRA in the privacy dimension, (3) to propose new metrics and methods to evaluate the performance of STDG approaches in terms of time and footprint, (4) to improve the STDG approaches to improve the trade-off between the dimensions, and (5) to unify all the proposed metrics and methods into an easy-to-use web-based application.

REFERENCES

- [1] K. El Emam and R. Hoptroff, "The Synthetic Data Paradigm for Using and Sharing Data," *DATA ANALYTICS & DIGITAL TECHNOLOGIES*, vol. 19, no. 6, p. 12, 2019.
- [2] F. Yang, Z. Yu, Y. Liang, X. Gan, K. Lin, Q. Zou, and Y. Zeng, "Grouped Correlational Generative Adversarial Networks for Discrete Electronic Health Records," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 906–913.
- [3] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing," *JMIR Medical Informatics*, vol. 8, no. 7, p. e18910, 2020.
- [4] A. Hernandez-Matamoros, H. Fujita, and H. Perez-Meana, "A novel approach to create synthetic biomedical signals using BiRNN," *Information Sciences*, vol. 541, pp. 218–241, 2020.
- [5] M. Stephens, R. S. J. Estepar, J. Ruiz-Cabello, I. Arganda-Carreras, I. Macía, and K. López-Linares, "MRI to CTA Translation for Pulmonary Artery Evaluation Using CycleGANs Trained with Unpaired Data," in *Thoracic Image Analysis*, ser. Lecture Notes in Computer Science, J. Petersen, R. San José Estepar, A. Schmidt-Richberg, S. Gerard, B. Lassen-Schmidt, C. Jacobs, R. Beichel, and K. Mori, Eds. Cham: Springer International Publishing, 2020, pp. 118–129.
- [6] J. Guan, R. Li, S. Yu, and X. Zhang, "Generation of synthetic electronic medical record text," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 374–380.
- [7] J. Dahmen and D. Cook, "SynSys: A synthetic data generation system for healthcare applications," *Sensors*, vol. 19, no. 5, p. 1181, 2019, number: 5 Publisher: Multidisciplinary Digital Publishing Institute.
- [8] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," *Neurocomputing*, vol. 416, pp. 244–255, 2020.
- [9] Gonçalves A., Ray P., Soper B., Stevens J., Coyle L., and Paula Sales A., "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, vol. 20, no. 1, p. 108, 2020.
- [10] Z. Wang, P. Myles, and A. Tucker, "Generating and evaluating synthetic UK primary care data: Preserving data utility patient privacy," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 2019, pp. 126–131, ISSN: 2372-9198.
- [11] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds. Springer International Publishing, 2019, pp. 510–526.

- [12] H. Wu, Y. Ning, P. Chakraborty, J. Vreeken, N. Tatti, and N. Ramakrishnan, "Generating realistic synthetic population datasets," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 4, pp. 45:1–45:22, 2018.
- [13] E. E. Fowler, A. Berglund, M. J. Schell, T. A. Sellers, S. Eschrich, and J. Heine, "Empirically-derived synthetic populations to mitigate small sample sizes," *Journal of Biomedical Informatics*, vol. 105, p. 103408, 2020.
- [14] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 787–792, ISSN: 2374-8486.
- [15] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," *Proceedings of Machine Learning for Healthcare (MLHC)*, 2018.
- [16] M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019.
- [17] J. Yoon, L. N. Drumright, and M. v. d. Schaar, "Anonymization through data synthesis using generative adversarial networks (ADS-GAN)," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378–2388, 2020.
- [18] S. Rashidian, F. Wang, R. Moffitt, V. Garcia, A. Dutt, W. Chang, V. Pandya, J. Hajagos, M. Saltz, and J. Saltz, "SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation," in *Artificial Intelligence in Medicine*, ser. Lecture Notes in Computer Science, M. Michalowski and R. Moskovitch, Eds. Springer International Publishing, 2020, pp. 37–48.
- [19] K. Chin-Cheong, T. Sutter, and J. E. Vogt, "Generation of heterogeneous synthetic electronic health records using GANs," in *Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019, pp. 1–6, accepted: 2020-06-16T05:38:56Z.
- [20] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019.
- [21] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 10 2016, pp. 399–410.
- [22] L. Wang, W. Zhang, and X. He, "Continuous patient-centric sequence generation via sequentially coupled adversarial learning," in *Database Systems for Advanced Applications*, ser. Lecture Notes in Computer Science, G. Li, J. Yang, J. Gama, J. Natwichai, and Y. Tong, Eds. Springer International Publishing, 2019, pp. 36–52.
- [23] S. Dash, A. Yale, I. Guyon, and K. P. Bennett, "Medical time-series data generation using generative adversarial networks," in *Artificial Intelligence in Medicine*, ser. Lecture Notes in Computer Science, M. Michalowski and R. Moskovitch, Eds. Springer International Publishing, 2020, pp. 382–391.
- [24] Beaulieu-Jones Brett K., Wu Zhiwei Steven, Williams Chris, Lee Ran, Bhavnani Sanjeev P., Byrd James Brian, and Greene Casey S., "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, p. e005122, 2019.
- [25] D. Lee, H. Yu, X. Jiang, D. Rogith, M. Gudala, M. Tejani, Q. Zhang, and L. Xiong, "Generating sequential electronic health records using dual adversarial autoencoder," *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1411–1419, 2020, publisher: Oxford Academic.
- [26] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, 2018.
- [27] S. Norgaard, R. Saeedi, K. Sasani, and A. H. Gebremedhin, "Synthetic sensor data generation for health applications: A supervised deep learning approach," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1164–1167, ISSN: 1558-4615.
- [28] A. Yale, S. Dash, K. Bhanot, I. Guyon, J. S. Erickson, and K. P. Bennett, "Synthesizing quality open data assets from private health research studies," in *Business Information Systems Workshops*, ser. Lecture Notes in Business Information Processing, W. Abramowicz and G. Klein, Eds. Springer International Publishing, 2020, pp. 324–335.
- [29] O. Mendelevitch and M. D. Lesh, "Fidelity and privacy of synthetic medical data," *arXiv:2101.08658 [cs]*, 2021. [Online]. Available: <http://arxiv.org/abs/2101.08658>
- [30] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, vol. 2014, pp. 1–11, 2014.
- [31] S. Ulianova. (2018) Cardiovascular disease dataset. [Online]. Available: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>
- [32] F. M. Palechor and A. d. I. H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico," *Data in Brief*, vol. 25, p. 104344, 2019-08-01.
- [33] T.-S. Lim. (1997) Contraceptive method choice data set. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>
- [34] U. M. Learning. (2016) Pima indians diabetes data. [Online]. Available: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- [35] B. V. Ramana, S. P. Babu, and N. Venkateswarlu. (2012) ILPD (indian liver patient dataset) data set. [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)](https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset))
- [36] (2021) Multivariate distributions — copulas 0.5.0 documentation. [Online]. Available: https://sdv.dev/Copulas/tutorials/03_Multivariate_Distributions.html#Gaussian-Multivariate
- [37] (2018-05-11) sdv-dev/SDV. [Online]. Available: <https://github.com/sdv-dev/SDV>

and industry contexts.

Mikel Hernandez holds a degree in Telecommunication Systems Engineering (2019) and a master's degree in Biomedical Technologies (2021) from the Faculty of Engineering of Mondragon University. Currently, is working as a Research Assistant in the Precision Medicine and Big Data line of the Digital Health and Biomedical Technologies department at the Vicomtech foundation, on which he developed his final master's degree thesis. His work is related with synthetic data generation and evaluation in the healthcare

Gorka Epelde is a Project Leader and Senior Researcher in Vicomtech. In 2014, Gorka obtained his Computer Science PhD from the University of the Basque Country. From 2000 until 2007 Gorka held the position of Assistant Researcher at Ikerlan. From 2007 onwards, Gorka has been a Staff Researcher at Vicomtech's eHealth and Biomedical Applications department. Since 2009, he is part of the eHealth group under the Bioengineering Area of the BioDonostia Health Research Institute. His fields of interest include interoperability architectures, data engineering, as well as the human computer interaction and the advanced visualisation of data.

Ane Alberdi. Lecturer and researcher at MGEP since 2017. Double graduate in Electronics Engineering from INP-ENSEEIH and MGEP, she also holds a Master's degree in Embedded Systems. She received her Ph.D. in 2017, where she worked on the early detection of diseases using smart environments (SE) and AI techniques. As a visiting researcher at the CASAS Lab. at Washington State University (WA, USA) during her Ph.D., she studied the use of SE to support the independent living of the cognitively impaired elderly. Ane has published articles in several scientific journals of impact. Currently, her research is focused in health-related applications of AI.

Rodrigo Cilla holds a degree in Computer Engineering (2007), a Master in Computer Science and Technology (2008) and a PhD (2012) from Universidad Carlos III de Madrid. He worked as a postdoctoral researcher in Instituto de Telecomunicações (Lisbon, Portugal, 2013), as a Research Fellow at the Victor Hatini Lab of Tufts University (2013 - 2016) and as a postdoctoral researcher at Fetal I+D Research Center of Fundació Sant Joan de Dèu (Espluges de Llobregat, Catalonia, 2016-2018). Since 2018 he is a researcher at Fundació Vicomtech (Donostia, Basque Country) where he is involved in different projects related to biomedical computer vision and machine learning.

Debbie Rankin received her B.Sc. and Ph.D. degrees in Computer Science from Ulster University, Northern Ireland (UK) in 2008 and 2012, respectively. She is currently a Lecturer in Computer Science at Ulster University in the School of Computing, Engineering and Intelligent Systems. Her research covers Health Informatics, Synthetic Data, Machine Learning, Data Mining, Big Data and Computer Vision.