IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 14, 2021, accepted January 8, 2022, date of publication January 19, 2022, date of current version January 31, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144765

A Multi-Dimensional Evaluation of Synthetic Data Generators

FIDA K. DANKAR[1], MAHMOUD K. IBRAHIM[2], AND LEILA ISMAIL[3,4], (Member, IEEE)
[1]Department of Information Systems and Security, United Arab Emirates University, Al Ain, United Arab Emirates
[2]Faculty of Engineering Science, Katholieke Universiteit (KU) Leuven, 3000 Leuven, Belgium
[3]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
[4]Intelligent Distributed Computing and Systems (INDUCE) Research Laboratory, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding author: Fida K. Dankar (fida.dankar@uaeu.ac.ae)

This work was supported in part by United Arab Emirates (UAE) University under Research Grant UPAR 31T121, and in part by the UAE University Big Data Analytics Center.

ABSTRACT Synthetic datasets are gradually emerging as solutions for data sharing. Multiple synthetic data generators have been introduced in the last decade fueled by advancement in machine learning and by the increased demand for fast and inclusive data sharing, yet their utility is not well understood. Prior research tried to compare the utility of synthetic data generators using different evaluation metrics. These metrics have been found to generate conflicting conclusions making direct comparison of synthetic data generators very difficult. This paper identifies four criteria (or dimensions) for masked data evaluation by classifying available utility metrics into different categories based on the measure they attempt to preserve: attribute fidelity, bivariate fidelity, population fidelity, and application fidelity. A representative metric from each category is chosen based on popularity and consistency, and the four metrics are used to compare the overall utility of four recent data synthesizers across 19 datasets of different sizes and feature counts. The paper also examines correlations between the selected metrics in an attempt to streamline synthetic data utility.

INDEX TERMS Data utility, privacy enhancing technologies, synthetic data generators.

I. INTRODUCTION

The technological advances of recent years led to the collection and storage of huge amounts of data. A recent report by IBM titled "10 key marketing trends for 2017" stated that more than 2.5 quintillion bytes of data is being created daily [1]. These large volumes of data have the potential to solve real world problems across multiple domains, and to enrich our lives with new technologies [2]–[5]. However, the data is generally not accessible to the broader research community due to privacy concerns. One particular domain where data sharing is challenging is healthcare. The availability of healthcare data for widespread research is limited by privacy laws, fears of data breaches, as well as administrative strategies seeking to benefit from the assumed value of the data [6]–[8]. A recent report from the US Government Accountability Office identifies data availability as a main barrier to the application of artificial intelligence (AI) in healthcare [9]. The report reveals that a lot of effort goes into accessing and curating data to make it usable for machine learning applications, thus severely delaying the pace of research and creating a barrier to data profit and progress in health care.

An increasingly popular way to overcome issues of data availability is the use of fully synthetic data. Synthetic data (SD) is artificial data that is simulated from real data to mimic its statistical properties. It is considered a safe approach for the wider release of sensitive data as it contains no identifiable information about the dataset it was generated from [10]–[12]. The exploitation of SD is at an early stage yet moving very fast. In a recent WSJ article, Gartner predicts that, 60% of the data used for AI and analytics will be synthetically generated by 2024 [13].

Various synthetic data generators (SDGs) were developed in the last decade, fueled by advances in machine learning and by the increasing demand for fast and inclusive data-sharing. However, empirical evidence of their utility has not been fully explored. Synthetic data is still in the experimental stage and is currently used to carry out exploratory analyses and to generate preliminary models, with the final analysis almost always obtained from the original dataset [14]. Utility of the synthetic data generators will determine whether synthetic data will be used outside the exploratory phase. Few research papers tried to investigate the utility of synthetic

The associate editor coordinating the review of this manuscript and approving it for publication was Junggab Son.

data generators [15]–[20]. They do so either by measuring a chosen statistical distance between the original and synthesized datasets [16], [21], or, more commonly, by measuring the differences in specific models generated from the original versus synthetic data [15], [17], [19], [20]. The choice of the measures/models is guided by the application of interest and the provided conclusions apply to that specific context. None offered any guidelines or criteria that synthetic data should satisfy in general when released for public use.

In this paper, we identify multiple dimensions of utility (referred to as *quality dimensions*) and use them to conduct a systematic multi-dimensional comparison between four recent synthetic data generators. We assess whether the identified dimensions are correlated, and whether any subset can be used to predict the overall masked data utility. The specific contributions of the paper are as follows:

1. We identify four dimensions of masked data utility by classifying available utility metrics based on the measure they attempt to preserve: attribute fidelity, bivariate fidelity, population fidelity, and application fidelity. Then we choose a representative metric from each of the identified dimensions based on popularity and consistency, the metrics are referred to as *quality metrics*, and their values for a specific synthetic dataset define the dataset *quality*.
2. We analyze the performance of four recent (open source) SDGs, investigate whether one generator consistently produces better synthetic datasets across all quality metrics, and how often the various metrics agree on this conclusion.
3. Considering supervised machine learning as one performance indicator (i.e. as a measure of the success of synthetic datasets when employed in real scenarios), we assess whether:
   a. The SDs with higher quality produce machine learning models of higher accuracy, and whether
   b. The SD quality affects the choice of the machine learning classifier. It is necessary that models selected when training on SD match the model selected when training on real data since these models are eventually applied on the real data for the final analysis.
4. Finally, to reduce the number of required metrics, we ask whether the identified metrics are correlated, and whether one or more metric can be used to assess the overall utility of masked data.

II. METHODS

A. SYNTHETIC DATA GENERATORS

Synthetic datasets are generated from a model that is fit to a real dataset. The model captures the statistical properties and patterns of the original dataset. SDGs employ one of two mechanisms for model generation: (i) statistical methods or (ii) machine learning methods [22]. Statistical methods generate the model by estimating the distribution of the population from which the data was drawn, while machine learning methods generate the model by training on the real dataset. Once a model is generated it is used to produce the synthetic dataset. Such production is nondeterministic, implying that the model generates a different synthetic dataset each time. It is generally accepted that synthetic data can be shared widely as it is considered non-identifiable and falls (to date) outside the scope of privacy regulations [23].

We consider four publicly available methods for synthetic data generation that are among the most influential work in this area [15]. Two of the methods are statistical, these are a Bayesian network based data synthesis technique [24], and a copula-based data synthesis technique [25]. The other two are based in machine learning, a parametric data synthesis technique [26], and a non-parametric tree-based data synthesis technique [27]. While other SDGs exist in the literature, they are often developed for specific applications [15], [28], [29].

The Bayesian network method, referred to as Datasynthesizer or DS, generates a Bayesian network model that captures the correlations between the different attributes in the real data and produces synthetic data samples from the constructed model. The copula-based method, known as synthetic data vault or SDV, generates the model by estimating the joint distribution of the attributes in the dataset. The joint distribution is estimated from individual (marginal) attribute distributions and a Gaussian Copula reflecting the dependency structure among the attributes, data samples are then produced from the generated model. The parametric and non-parametric methods, referred to as Synthpop parametric and Synthpop nonparametric or SP-p and SP-np respectively, generate the model by synthesizing the different attributes sequentially. The first attribute is synthesized after estimating its marginal distribution from the raw data, and following attributes are synthesized after estimating their conditional distribution using all prior attributes as predictors. SP-p uses linear regression for estimating conditional distributions, while SP-np uses classification and regression trees, CART [30].

B. UTILITY METRICS: OVERVIEW AND CLASSIFICATION

Data utility attempts to measure whether the data is appropriate for processing and analysis. In the general broad sense, this translates to how beneficial and reliable the data is to society. Such benefit is impossible to quantify as it is data and knowledge dependent. The practice is rather to capture utility of a masked dataset through comparison with the original dataset, to check whether any function or statistical measure is preserved between original and masked datasets.

There is a wealth of synthetic data utility measures each focusing on a specific statistical measure or aiming to preserve a specific function. In 2019, Drechler and Reiter [31] classified available utility measures into two categories, narrow and broad. Narrow measures assess the ability of the synthetic data to replicate a specific analysis performed on the original data (such as data summaries or coefficients of a trained model). Broad measures capture general features of the entire dataset (such as differences in marginal distributions or overall distributional similarity between the two datasets). Prior to that, in 2017, Snoke et al [14] had the same categorization of the utility measures referring to them as specific and general measures respectively.

Narrow or specific measures are widely used for assessing synthetic data [15], [19], [20], [27], [32], [33]. They are useful when the analysis to be performed on the synthetic data is known ahead of time. Broad or general measures are more helpful in allowing an overall assessment of synthetic data. They measure the extent of agreement between inferences obtained from synthetic data and real data. They are used when the exact analysis to be done on the data is not known at the time of data release [14], [26], [34], [35].

Each of the categories incorporates many utility metrics (refer to Table 1 for examples). Up to now, there are no general guidelines on which of these metrics to use when comparing the overall utility of SDGs, making this task extremely difficult. Therefore, we reviewed available utility metrics and further categorized the *broad measures* category into three sub-categories (or dimensions) depending on the statistics they attempt to compute: attribute fidelity, bivariate fidelity and population fidelity. These three dimensions along with the application fidelity form our *quality dimensions*. A representative metric is chosen from each dimension, and the four metrics will be used together to assess the four recent SDGs introduced earlier. An illustration of the four dimensions is provided in Figure 1.

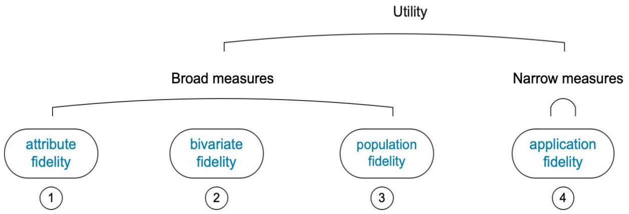

FIGURE 1. The four quality dimensions.

C. QUALITY METRICS

The four quality dimensions are described in detail in this section. A representative metric is chosen from each dimension to perform the comparison. The choice is based on the popularity and consistency of the metric. The chosen metrics are together denoted as quality metrics and are defined below as well.

1) ATTRIBUTE FIDELITY

Attribute fidelity (or univariate fidelity) covers metrics that measure the basic structural similarity between the datasets. Each attribute in the masked data should have similar structure and similar fundamental aggregated statistics (variable types, formats, names means, and ranges) or similar univariate distributions for continuous and discrete variables. Univariate fidelity measures are commonly used in the synthesis literature and are necessary to determine whether the same code can be applied to both datasets without producing syntactical errors [34], [36]. Measures commonly used are Hellinger distance [37] and Kullback-Leibler divergence [38]. Univariate fidelity is the minimum requirement for synthetic data to be useful, meaning that all marginal distributions of original and synthetic datasets should be matching [34]. Hellinger (H) is a popular univariate utility measure. It was shown to be consistent and easy to interpret (as it produces values between 0 and 1). For each variable v (data column), Hellinger distance is calculated as follows:

$$H\left(v_o, v_s\right) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

where v_o is the original column and v_s is the corresponding synthesized one, p_i and q_i correspond to the probability for every distinct value of variable v in the original and synthesized datasets respectively. To calculate the overall Hellinger distance, we compute the mean Hellinger distance across all variables. The smaller the H value, the closer the synthetic dataset is to the real dataset in terms of univariate distributions across all variables [34].

2) BIVARIATE FIDELITY

Bivariate fidelity refers to metrics measuring correlations among pairs of variables in the dataset so as to capture the statistical dependency structure of the original and released datasets. Retaining such structure in the synthetic dataset ensures that the underlying relationships between the attributes are preserved (for example symptoms are attributed to the right diagnosis and employment status is attributed to an appropriate age) such association is crucial to ensure truthful representation of the original dataset. Pairwise Correlations are sometimes measured using pairwise correlation plots such as heat maps [27], [33], but more often using statistical measures such as pairwise correlation difference (PCD) [18]. We assess the correlations between attribute pairs using the latter. PCD is defined as:

$$PCD\left(R, S\right) = ||Corr\left(R\right) - Corr\left(S\right)||_F$$

where R, S represent the real and synthetic data matrices, and $Corr$ stands for correlation coefficient. Smaller values for PCD imply that the synthetic and real data are close in terms of linear correlations across variables. PCD measures the difference in terms of Frobenious norm and is defined at the dataset level.

3) POPULATION FIDELITY

Population fidelity metrics reflect the similarity on the entire distribution of the masked data in comparison with the original data. They attempt to reflect large-scale features of the entire distributions. Many metrics fall under this category as it is the most commonly used approach for the evaluation of masked data. They are helpful in allowing a global assessment of how well the final inference might agree with what would have been obtained had the user had access to the original data [39]. Many metrics from this category have been proposed in the literature:

TABLE 1. Utility measures classification and examples of representative metrics.

Utility	Measurements	Example Metrics
Global measures-Univariate fidelity	structural similarity	Range, averages, variable names and formats
	Univariate distribution	Hellinger, graphical models [34], Normalized Kullback-Leibler divergence [38]
Global measures-bivariate fidelity	Bi-variable correlations	pairwise correlation difference [18], mutual and heat maps [33]
Global measures-population fidelity	Difference in statistical dependency structure [multivariate correlation)	Cross classification [18]
	Cluster-based distributional difference	Log-cluster [40]
	Difference in empirical distributions	Kolmogorov-Smirnov type statistics [40]
	Likelihood metrics	Bayesian network based or Gaussian mixture models based [41]
	Distinguishability	Propensity [14], [36]
Application fidelity-analysis specific measures	Prediction Accuracy	Ability to replicate studies performed on real data

a. *The cross-classification metric* measures how well a synthetic data captures the statistical dependency structure of the original dataset. It measures the dependence via prediction accuracy. Each variable in the real dataset is predicted from all other variables (via a chosen classifier), the resulting classifiers are tested for accuracy on both the synthetic dataset and some hold-out real data. The comparison between the obtained accuracies is useful for evaluating if the statistical properties of both datasets are similar. For more information, the reader is referred to [18].

b. *The log-cluster metric* measures the similarity of the underlying dependency structure in terms of clustering [18], [40]. The real and synthetic datasets are merged into one dataset, a cluster analysis with a fixed number of clusters is performed on the merged dataset, placing records into clusters of similar values. Finally, a metric is calculated to reflect the distribution of the synthetic dataset across the different clusters. If the allocation to the different clusters is similar for synthetic and real, then this suggests similar distributions.

c. *Likelihood metrics* compare the datasets by fitting the real data to a probabilistic model (such as Bayesian network or Gaussian mixture models) then computing the likelihood that the synthetic datasets follows the same distribution [41].

d. *Difference in Empirical distributions type metrics* measure the differences between the empirical cumulative distribution functions calculated for real and synthetic datasets. Kolmogorov-Smirnov type statistics [40] estimate the empirical distributions difference. They first calculate the discrete empirical distributions of both datasets (from the supplied sample), then they calculate the average square differences between the two.

e. *Distinguishability type metrics* characterize the extent to which it is possible to distinguish the original dataset from the synthesized one. The most prominent distinguishability measure is the propensity score [14], [36]. It involves building a classification model to distinguish between the real and released records. A high utility implies the inability of the model to perform the distinction.

The most popular population level fidelity metric is propensity score (from the distinguishability subset) [14]. It is advocated as the best measure for synthetic data evaluation and was cited as the most promising measure for comparing synthetic data [14], [27], [34], [40], [42].

To calculate propensity, the original and synthetic datasets are joined in one group with a binary indicator assigned to each record depending on whether the record is real or synthesized (1 for synthetic rows and zero for original rows). A binary classification model is constructed to discriminate between real and synthetic records. The model is then used to compute the propensity score $\hat{p}_i$ for each record i (predicted value for the indicator) [40]. The propensity score is then calculated from the predicted value as follows:

$$pMSE = \frac{1}{N}\sum_{i}(\hat{p}_i - 0.5)^2$$

where N is the size of the joint dataset. Propensity score varies between 0 and 0.25, with 0 indicating no distinguishability between the two datasets. This can happen if the generator overfits the original dataset and creates a synthetic that is indistinguishable from the original one (leading to a score of $\hat{p}_i = 0.5$ for every record). On the other extreme, if the two datasets are completely distinguishable, the propensity score for each record will be 1 or 0 (1 for synthetic rows and zero for original rows), leading to an overall score of 0.25. Propensity score is cited as the most practical measure for predicting the overall utility of a dataset [34], it is also valuable for comparing different synthesis approaches [42].

4) APPLICATION FIDELITY

Application fidelity evaluates the performance of masked data in a specific application or analysis. Multiple analysis specific measure are used for assessing synthetic data, the most common of which is prediction accuracy [15], [20].

TABLE 2. **Datasets description.**

Dataset name	Short name	Number of observations	Number of attributes (predictors)	Number of labels	Origin
BankNote	D_1	1,372	4	2	UCI
Titanic	D_2	891	7	2	Kaggle
Ecoli	D_3	336	7	8	UCI
Diabetes	D_4	768	9	2	UCI
Cleveland heart	D_5	297	13	2	UCI
Adult	D_6	48,843	14	2	UCI
Breast cancer	D_7	570	30	2	UCI
Dermatology	D_8	366	34	6	UCI
SPECTF Heart	D_9	267	44	2	UCI
Z-Alizadeh Sani	D_{10}	303	55	2	UCI
Diabetic Data	D_{11}	101766	50	3	Cerner clinical database
Colposcopies	D_{12}	287	68	2	UCI
ANALCATDATA	D_{13}	841	71	2	OpenML
Mice Protein	D_{14}	1,080	80	8	UCI
Diabetic Mellitus	D_{15}	281	97	2	OpenML
Tecator	D_{16}	240	124	2	OpenML
Colorectal	D_{17}	690	176	2	Datasphere NCT00384176
Arrhythmia	D_{18}	452	279	2	UCI
Scene	D_{19}	2407	293	2	OpenML

In prediction models, if inferences agree between synthetic and real data, then the synthetic data is said to have high utility. We use 4 classification algorithms to assess application fidelity: Logistic regression (LR), support vector machines (SVM), Random Forest (RF) and decision trees (DT). For each dataset, the 4 different classification models (CM) are trained. Models are trained on the real training dataset as well as the synthetic datasets and tested on the real data. Testing on real data allows us to determine how well a model trained on synthetic data will perform in real-life. We use the prediction accuracy measure (PA) as well as and F1-scores to test the accuracy of the generated models. The F-score is essential in case of imbalanced dataset, as it reveals how much a model is correctly classifying the minority class, which may not be detected by accuracy [43], [44].

It is important to note, that application-level fidelity is hard to capture from one application/analysis, or even multiple applications. We chose classification as it is a popular tool for synthetic data evaluation.

On the other hand, one of the objectives of this investigation is to evaluate whether the other three dimensions of quality are good predictors of application-level fidelity through prediction accuracy [15], [19], [20], [45].

D. SYNTHETIC DATA GENERATION PROCESS

To assess the synthetic data generators, several datasets with variable sizes and feature counts were used (Table 2). As the generation of synthetic data is stochastic, there will be utility variation owing to the generation process itself. Thus, for each generator SDG_i and dataset D_j, the synthesis is repeated 20 times to generate 20 synthetic datasets.

Specifically, the process for preparing, synthesizing and testing the synthesizers' utility is described below:

1. We used the raw unprocessed real datasets as our synthesizers' input
2. We performed a repeated holdout method, where we randomly generate 4 splits of each real dataset into 70% training and 30 % testing.
3. For each split, we repeatedly apply the four data synthesis methods 5 times with the real training data as the input. The generated synthetic datasets are of equal length as the real training data.
4. All utility metrics (apart from prediction accuracy) are calculated for each of the synthetic datasets generated. The *average utility* for a [dataset, generator] pair represents the average across the 20 corresponding datasets.
5. Prediction accuracy and F1 scores are calculated for each synthetic dataset, and each CM using the corresponding real testing dataset. In other words, the different machine learning models are trained on the synthetic data and tested on the real data. The *average accuracy/F1-score* for each [dataset, generator, CM] corresponds to the average across the 20 corresponding datasets.

We use 19 datasets in our experiments contained within the University of California Irvine repository [46], OpenML platform [47], Datasphere [48], Cerner clinical database [49] and

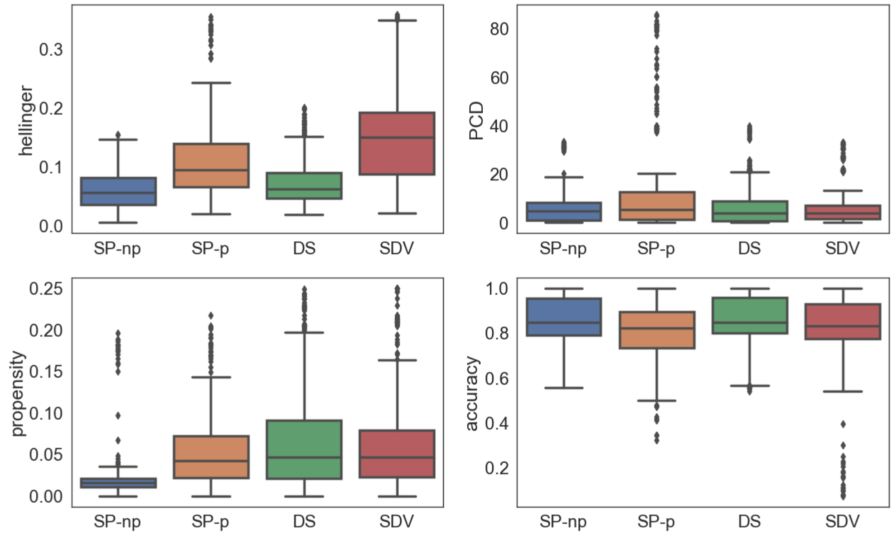

FIGURE 2. Performance of the different synthetic data generators on each quality metric across all datasets. Lower metric values are more desirable across all metrics except accuracy.

Kaggle community platform [50]. Details about the different datasets are given in Table 2.

E. SET-UP

Data generation was performed on an AWS virtual machine, instance type: r5a.8xlarge, having 32 vCPUs with 256 GiB memory [51]. Synthetic data are generated from raw unprocessed real data (as recommended by recent experiments[35]). When generating synthetic data, default generation settings were used for all synthetic data generators except for SDV (defaults are the settings suggested by the authors of each SDG, readers are invited to check [24]–[26] for more information). For SDV, the creators of the system recommended changing the default distribution for categorical attributes to Gaussian KDE [52].

We used the mice function from the R mice package for imputation, and the ps function from the R twang package for calculating the propensity scores (using GBM model). For machine learning models, the python scikit-learn library was used, and for clustering mixed numerical and categorical data the Kprototype function from the Python Kmodes library was used [53].

III. RESULTS

The results of the multi-dimensional comparison between the four synthetic data generators (Datasynthesizer or DS, Synthpop parametric or SP-p, Synthpop nonparametric SP-np, and Synthetic Data Vault or SDV) are presented consecutively in four subsections: (i) first the results related to the data generator with the highest quality are presented, then (ii) results related to the investigation into a correlation between the different metrics is presented, (iii) the third subsection presents results related to the accuracy of the winning SDG, and lastly (iv) results on SD agreement with the real data on the winning classifiers are presented.

TABLE 3. Count of winning SDGs in terms of different metrics.

SDG	Hellinger	PA	PCD	Propensity
SP-np	14	8	7	15
SP-p	1	0	1	0
SDV	0	3	5	3
DS	4	8	6	1
Total	19	19	19	19

A. PERFORMANCE COMPARISON

1) COMPARISON ON AVERAGE RESULTS

We use all (four) quality metrics to compare the performance of the synthetic data generators. Figure 2 depicts the performance of the generators on each metric using boxplots. It shows better average performance and standard deviation for SP-np across all metrics. SDV exhibits similar performance to SP-np for PCD, and DS for accuracy.

The results of Figure 2 are further analyzed in Tables 3, 4, 5 and 6. Table 3 counts, for each quality metric, the number of times each of the SDGs produced the best result across the 19 datasets. The previous conclusions are echoed in the table, SP-np exhibits better overall results. The table suggests that DS follows SP-np on all dimensions except population fidelity.

To better understand the average performance of the different SDGs, Table 4 displays the average results for each metric. The table shows that SP-np achieves the best average results across all metrics.

Table 5 reports on the mean and stability for different SDGs across the different datasets. The results indicate that SP-np provides the best stability for application fidelity, population fidelity and bivariate fidelity. It also displays a close-to-best

TABLE 4. Average of the metrics across all the datasets.

SDG	Hellinger	PA	PCD	Propensity
REAL		0.8746276		
SP-np	0.0617296	0.8409588	6.2960989	0.0233651
SP-p	0.1171702	0.7956391	14.130144	0.0557168
SDV	0.1539435	0.7957867	7.4813360	0.0647602
DS	0.0829068	0.8355981	10.193042	0.0724779
winning	SP-np	SP-np	SP-np	SP-np

TABLE 5. Stability measures across all datasets (RA stands for relative accuracy, it is the difference between the accuracy of the SD from its corresponding real).

SDG	RA	Hellinger	PCD	Propensity
DS	(0.027,0.027)	(0.02,0.027)	(1.277,1.864)	(0.019,0.024)
SDV	(0.035,0.026)	(0.023,0.013)	(0.721,0.821)	(0.022,0.023)
SP-np	(0.03,0.02)	(0.014,0.012)	(0.546,0.586)	(0.008,0.018)
SP-p	(0.059,0.057)	(0.012,0.011)	(1.319,3.022)	(0.012,0.019)

TABLE 6. Average bias and stability for PA across all datasets (the arrow symbols on the right side of metric's name indicate that lower values are desired).

SDG	Range PA↓	Range Hellinger↓	Range PCD↓	Range Propensity↓
SP-np	6.301647695	0.040265719	1.36195216	0.02601857
SP-p	11.39524974	0.034346952	2.990892833	0.03638797
SDV	10.35558468	0.056885875	1.541746138	0.06904959
DS	6.814676059	0.046063943	2.668823648	0.06587738

stability for attribute fidelity. SP-p displays the best stability for attribute fidelity.

In other work, stability was defined as the average range of metric values across all datasets [17]. These alternative values are shown in Table 6 and they offer the same conclusion.

2) METRIC AGREEMENT ON BEST SDG

It is evident from the previous sub-section that, overall, SP-np shows a clear lead over other SDGs. However, the level of agreements on a leading SDG (on a case by case basis) among the different metrics is not evident. It is important to note that the four quality metrics are measuring different aspects of utility and will thus differ in their evaluation of the different synthetic datasets. Our goal is to check the extend of this divergence/agreement between the metrics when choosing the best synthetic generator.

TABLE 7. Kappa score measuring the agreement of different metrics on the winning SDG.

	Hellinger	PCD	Propensity	PA
Hellinger		0.368421	0.508772	0.298246
PCD	0.368421		0.017544	0.578947
Propensity	0.508772	0.017544		0.087719
PA	0.298246	0.578947	0.087719	

TABLE 8. Kappa score measuring the agreement of different metrics on one of the best two SDGs.

	Hellinger	PCD	propensity	PA
Hellinger		1	1	0.761569
PCD	1		1	1
propensity	1	1		0.88511
PA	0.761569	1	0.88511	

Figure 3 below shows the winning SDG for every dataset based on each metric. It shows that the four metrics rarely agree (unanimously) on a winning SDG- D8, D11, and D19 are the only exceptions- however, 3 out of 4 agree on a winner 58% of the time. Looking at majority votes, the majority agrees on SP-np as the winner 63% of the time, and on DS 10.5% of the time. The remaining cases reveal a tie between SP-np and DS (16%) or SP-np and SDV (10.5%).

Next, we consider another measure of agreements between the different metrics- the kappa score. The kappa score measures the degree of agreement between two evaluators, also known as inter-rater reliability [54], it takes into consideration the probability of agreeing by chance. Perfect agreement is achieved when Cohen's kappa equals 1; a value of Cohen's kappa equal to zero suggests that the agreement is no better than that which would be obtained by chance alone. Although there is no formal scale, the following categories are often considered appropriate for judging the level of agreement: agreement is Slight if $0.00 \leq \kappa \leq 0.20$, Fair if $0.21 \leq \kappa \leq 0.40$, Moderate if $0.41 \leq \kappa \leq 0.60$, Substantial if $0.61 \leq \kappa \leq 0.80$ and Almost perfect if $\kappa > 0.80$.

Table 7 below reports the agreement among the different metrics on the winning SDG. The degree differs between metrics, ranging from slight to moderate. The pairs (Hellinger, propensity) and (PA, PCD) are in highest agreement on the winner, followed by (PCD, Hellinger) and (PA, Hellinger).

However, requiring all measures to agree on a winner may be too restrictive, Once we relax the requirement to an agreement on one of the first two winning SDGs (result in Table 8), all the metrics become in *substantial* or *almost perfect* agreement with each other.

To sum up, SP-np exhibited the best overall performance on all quality metrics and displayed the best stability. All metrics support this conclusion although they do not do so on a case-by case basis. Looking at individual datasets, most of the metrics agree on SP-np as the winner 63% of the time, and on DS 10.5% of the time.

B. OVERALL UTILITY MEASURE

In this sub-section, we aim to investigate whether one of the metrics can be considered as a good overall measure of

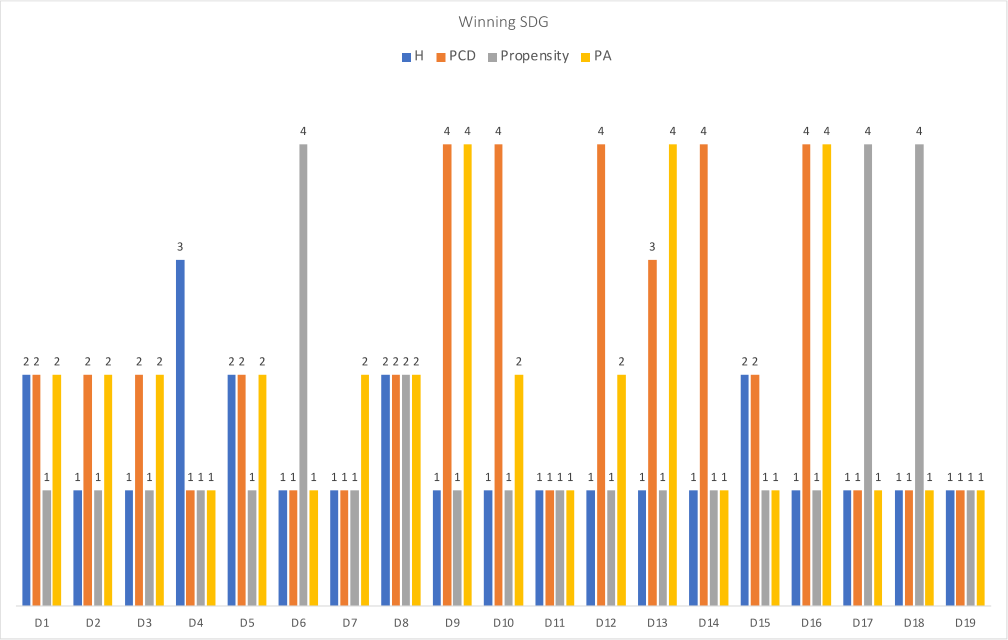

FIGURE 3. Winning SDG for every dataset and metric. 1 refers to SP-np, 2 for DS, 3 for SP-p and 4 for SDV.

TABLE 9. Correlation matrix between metrics.

	Hellinger	PCD	Propensity	PA
Hellinger	1	0.535184	0.268217	-0.2636
PCD	0.535184	1	0.257282	-0.2684
Propensity	0.268217	0.257282	1	-0.33437
PA	-0.2636	-0.2684	-0.33437	1

quality across all synthetic datasets (irrespective of how the data was generated). In other words, we ask whether any one metric can be used as an indicator/predictor for all quality dimensions. The answer to this question is very useful as such metric can be used to optimize on when generating multi-purpose synthetic datasets.

To answer this question, we looked at the affinity between pairs of metrics by calculating pairwise correlation. Table 9 reports on the correlations between the different metrics. A high correlation between 2 metrics indicates that the measures maybe correlated. The results suggest that, apart from (PCD, Hellinger) pair, correlations are of low degree (<0.4).

In an attempt to further understand the type, reliability and direction of these correlations, we try to express the relation in terms of an equation using regression. Figure 4 represents the results in 12 different graphs, one for each metric pair. Individual graphs represent the values of the metrics in relation with each other across all generated synthetic datasets (metric values for different datasets are represented on the x and y axis). The red line on the graphs represents an attempt at finding the best linear relation between the two metrics, or the regression line with values on the x-axis being the predictors.

Regression attempts to establish how/whether any change in any of the metrics causes the other to change as well. The direction and strength of the relationship between the two metrics is indicated by slope of the regression, and the error in the estimation (or goodness of fit for the regression line) is measured through $r2$ (correlation coefficient squared). The values for $r2$ are depicted inside the individual graphs.

The graphs suggest no strong relation between any of the metrics, and that no single metric can be used as a predictor of (multi-dimensional) quality. However, there exists some relation (bi-directional) between (PCD, Hellinger) which, along with the results of Table 9, suggest that scientists need not compute both metrics (Hellinger, PCD) for a given dataset. Another weak relation exists between Accuracy and Propensity.

C. GENERATOR FOR BEST ACCURACY

If prediction accuracy is considered a performance indicator, or a tangible measure of the success of synthetic datasets in preserving complex patterns in machine learning applications, then it becomes important to investigate the SDG that provides the best accuracies and to check the stability of such conclusion.

Table 10 presents the average change in accuracy for each SDG and each classifier across the datasets. SP-np shows lowest accuracy loss across 3 of the classifiers (DT, LR, RF), while DS dominates on SVM with SP-np trailing by less than 1% point. Thus, if prediction accuracy is taken as a performance indicator, then it becomes evident that SP-np has the highest performance, with an average reduction in accuracy of less than 3.5%.

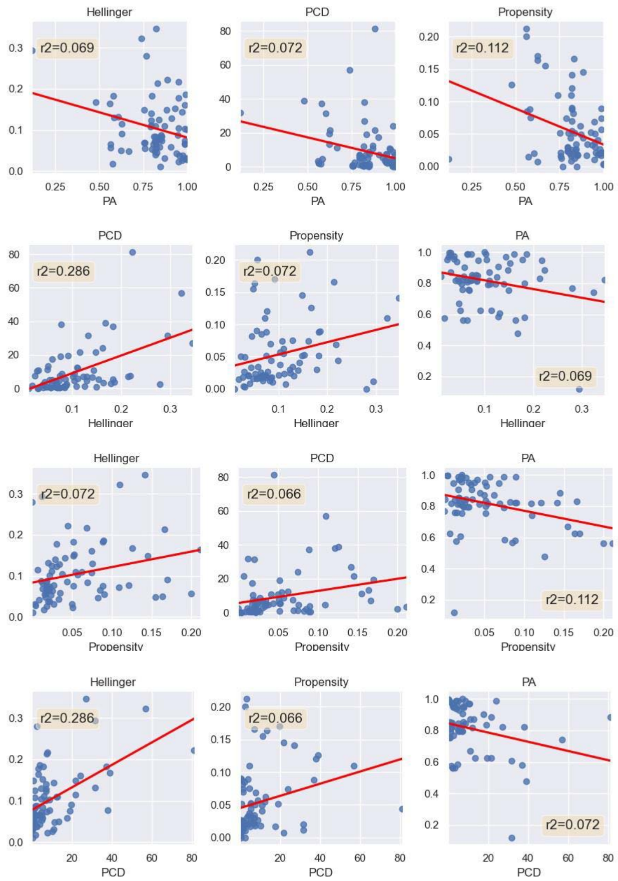

FIGURE 4. Twelve graphs representing different metric pairs in relation with each other (on the x and y axis respectively) across all generated synthetic datasets.

TABLE 10. Mean absolute difference in accuracy between the real and synthetic datasets for each machine learning model and synthetic data type.

SDG	DT	LR	RF	SVM	Mean	Range Accuracy
SP-np	2.43%	3.81%	3.53%	4.17%	3.49%	6.301648
SP-p	11.75%	8.07%	9.59%	8.62%	9.51%	11.39525
SDV	14.67%	7.57%	9.90%	7.54%	9.92%	10.35558
DS	5.97%	4.33%	4.18%	3.49%	4.49%	6.814676

In addition to accuracy scores, we also consider changes in F1 scores. Table 11 presents the mean absolute values for the F1 scores and provides the same conclusion about SP-np. The stability of both measures (Accuracy and F1 score) is shown in the last columns of Tables 10 and 11 respectively. The results indicates a higher stability for SP-np which provide good support for our conclusion.

Figure 5 presents the results in a more granular form. The absolute difference between real and synthetic datasets on accuracy and F1 score are shown per classification algorithm across all datasets. The results confirm the conclusion of higher performance for SP-np followed (closely) by DS.

TABLE 11. The first 5 columns report on the mean absolute difference in F1 between real and synthetic data, the last column reports on the range for the F1 value for each SDG.

SDG	DT	LR	RF	SVM	Mean	Range F1
SP-np	2.49%	4.00%	4.26%	5.17%	3.98%	7.13867
SP-p	11.97%	8.58%	10.91%	9.81%	10.32%	11.5511
SDV	14.31%	7.91%	10.66%	8.45%	10.33%	11.5349
DS	5.80%	4.48%	5.24%	4.77%	5.07%	7.64472

TABLE 12. Count of winning classifiers according to SDG.

SDG	LR	SVM	RF	DT	Total
REAL	6	4	8	1	19
SP-np	6	5	5	3	19
SP-p	12	4	3	0	19
SDV	11	5	3	0	19
DS	8	5	5	1	19

TABLE 13. Number of matched of winning classifier trained on real data versus when trained on synthetic data.

SDG	match
SP-np	15/19 (0.79)
SP-p	9/19 (0.47)
SDV	10/19 (0.53)
DS	8/19 (0.42)

We conclude that SP-np, in addition to displaying the best results on all quality dimensions, exhibits accuracy loss of less than 3.5% points on average, with a range of less than 6.5% across all datasets.

D. MATCHES ON WINNING CLASSIFIER

Synthetic datasets are currently used for exploratory analysis, and it is often the case that final analysis is almost always performed on real data. For synthetic data to be useful under this scenario, it is important for the models generated from synthetic data to be applicable on real data. As such, we compare the winning classifiers when trained on real data with the winning classifiers when trained on synthetic data. Table 12 shows the count of winning classifiers across the 4 classification algorithms considered. RF has the highest number of wins on real datasets followed by LR. Whereas, for synthetic datasets LR has the highest wins followed by (RF and SVM) for SP-np and DS and SVM for SP-p and SDV.

Table 13 shows the number of times the winning classifier trained on synthetic data matched the winning classifier trained on real data across all 19 datasets. SP-np shows the

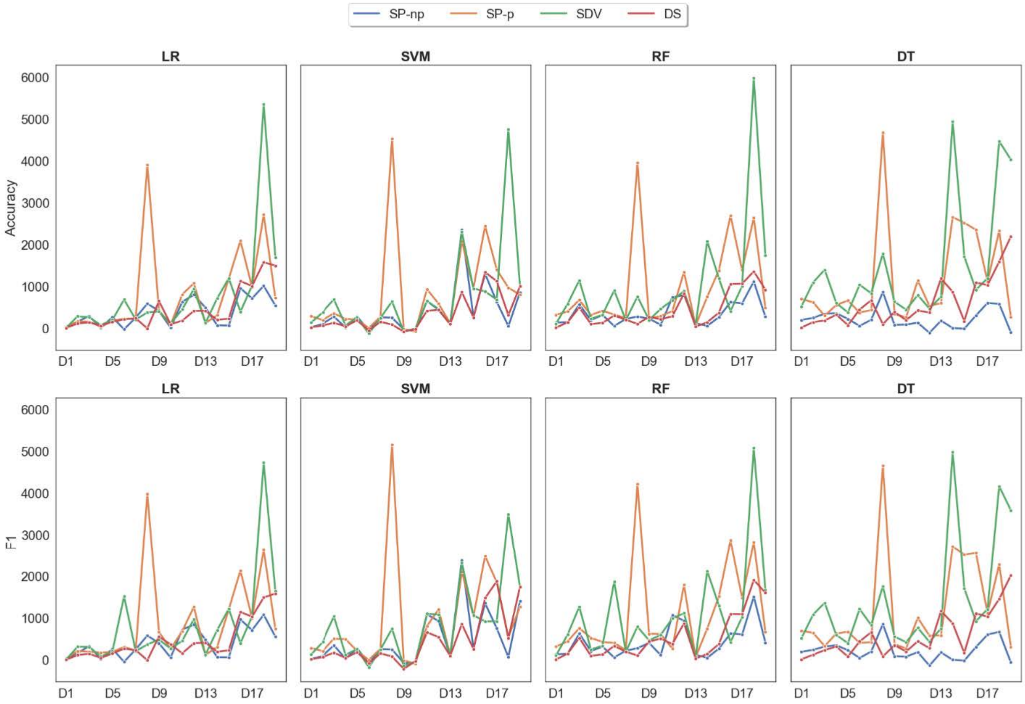

FIGURE 5. Average absolute difference between real and synthetic data for accuracy and F1 score values. Lower values are desired.

highest match of 79% followed by (not closely) by SDV at 53%.

IV. DISCUSSION

Healthcare data science is one area where privacy protection is particularly essential, yet where machine learning is critical for improving health and decision making. Synthetic data can solve the issues of data availability as well as delays in acquiring data if proven to have high utility. This paper identified four quality metrics for assessing the overall utility of masked data and used these metrics to compare four recent SDGs across 19 datasets of different sizes and feature counts. It also investigated correlations between the different metrics in an attempt to streamline synthetic data utility.

Our results indicate that SP-np is the overall winning SDG. SP-np provides the best average values across all metrics as well as the best (overall) stability and consistency. SP-np displays the best overall accuracy with a mean of 3.49% accuracy loss and accuracy range of 6.5% across all experiments. SP-np also resulted in the highest number of matches with real data on winning classifiers (79%) followed by SDV (53%). This constitutes compelling evidence for the use of SP-np for different analysis purposes. Given a private dataset to be shared publicly, an ensemble of multiple synthetic datasets generated using SP-np may provide the highest quality (when the purpose of the analysis is undefined or diverse). Moreover, it offers the highest chance of providing analysts with the best classifier to use on the real data once available.

On another front, our results suggest no strong correlation between the different quality metrics, which implies that all metrics are required when evaluating the overall utility of synthetic data. A moderate correlation is exhibited between PCD and Hellinger. More experiments are needed to better define the strength and type of this correlation.

V. CONCLUSION

The abundance of utility metrics, and the absence of general guidelines for utility measurement makes overall utility comparison between SDGs extremely difficult. Here, to enable direct comparison of the generators, we investigated commonly used metrics for masked data assessment and classified them into four dimensions depending on the function they attempt to preserve -attribute fidelity, bivariate fidelity, population fidelity, and application fidelity. The categorization was then used to compare four eminent SDGs and provided conclusions on the best overall SDG.

Further investigations with more datasets and machine learning algorithms are needed to validate the results. As an extension to this work, research into a single quality measure that captures all quality metrics would greatly help data holders in optimizing the quality of the released dataset.

Every privacy enhancing technology (PET) leads to reduction in data utility. Although PETs have been used for data sharing since a long time, it is still not clear, how much utility can be retained in the shared data without compromising privacy, nor what is an acceptable decrease in performance in different applications such as healthcare. Thus, our results can inform on the SDG with best multi-dimensional utility, but not on whether the utility is at an acceptable level. Comparisons between synthetic data and other PETs may give some insights in this direction.

ABBREVIATION

AI:	Artificial intelligence.
SDG:	Synthetic data generator.
SD:	Synthetic dataset.
DS:	Data synthesizer.
SDV:	Synthetic data vault.
SP-np:	Nonparametric Synthpop.
SP-p:	Parametric Synthpop.
H:	Hellinger.
PCD:	Pairwise correlation difference.
pMSE:	Propensity.
PA:	Performance accuracy.
RA:	Relative accuracy.
LR:	Linear regression.
SVM:	Support vector machine.
RF:	Random forest.
DT:	Decision trees.
CM:	Classification models.
PET:	Privacy enhancing technology.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments on earlier versions of this paper.

REFERENCES

[1] (Dec. 15, 2016). *10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations*. Bizibl.Com. Accessed: Sep. 1, 2020. https://bizibl.com/marketing/download/10-key-marketing-trends-2017-and-ideas-exceeding-customer-expectations

[2] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare J.*, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: 10.7861/futurehosp.6-2-94.

[3] E. A. McGlynn, T. A. Lieu, M. L. Durham, A. Bauck, R. Laws, A. S. Go, J. Chen, H. S. Feigelson, D. A. Corley, D. R. Young, A. F. Nelson, A. J. Davidson, L. S. Morales, and M. G. Kahn, "Developing a data infrastructure for a learning health system: The PORTAL network," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 4, pp. 596–601, Jul. 2014, doi: 10.1136/amiajnl-2014-002746.

[4] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-assisted decision-making in healthcare," *Asian Bioethics Rev.*, vol. 11, no. 3, pp. 299–314, Sep. 2019, doi: 10.1007/s41649-019-00096-0.

[5] *17 Big Data Examples & Applications*. Built in. Accessed: Sep. 1, 2020. [Online]. Available: https://builtin.com/big-data/big-data-examples-applications

[6] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J. P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the genomic era," *ACM Comput. Surv. CSUR*, vol. 48, no. 1, p. 6, 2015.

[7] F. K. Dankar and R. Badji, "A risk-based framework for biomedical data sharing," *J. Biomed. Informat.*, vol. 66, pp. 231–240, Feb. 2017.

[8] F. K. Dankar, M. Gergely, B. Malin, R. Badji, S. K. Dankar, and K. Shuaib, "Dynamic-informed consent: A potential solution for ethical dilemmas in population sequencing initiatives," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 913–921, Jan. 2020, doi: 10.1016/j.csbj.2020.03.027.

[9] *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development (STAA) Policy Briefs & Reports EPTA Network*. Accessed: Sep. 1, 2020. [Online]. Available: https://eptanetwork.org/database/policy-briefs-reports/1898-artificial-intelligence-in-health-care-benefits-and-challenges-of-machine-learning-in-drug-development-staa

[10] J. Hu, "Bayesian estimation of attribute and identification disclosure risks in synthetic data," 2018, *arXiv:1804.02784*.

[11] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proc. VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018, doi: 10.14778/3231751.3231757.

[12] N. Ruiz, K. Muralidhar, and J. Domingo-Ferrer, "On the privacy guarantees of synthetic data: A reassessment from the maximum-knowledge attacker perspective," in *Privacy in Statistical Databases*. Cham, Switzerland: Springer, 2018, pp. 59–74, doi: 10.1007/978-3-319-99771-1_5.

[13] S. Castellanos. (2021). *Fake it to Make it: Companies Beef Up AI Models With Synthetic Data*. Accessed: Sep. 22, 2021. [Online]. Available: https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601

[14] J. Snoke, G. Raab, B. Nowok, C. Dibben, and A. Slavkovic, "General and specific utility measures for synthetic data," 2016, *arXiv:1604.06651*.

[15] 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 Med. Informat.*, vol. 8, no. 7, Jul. 2020, Art. no. e18910.

[16] B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee, "Synthetic data for social good," 2017, *arXiv:1710.08874*.

[17] A. R. Benaim, R. Almog, Y. Gorelik, I. Hochberg, L. Nassar, T. Mashiach, M. Khamaisi, Y. Lurie, Z. S. Azzam, J. Khoury, D. Kurnik, and R. Beyar, "Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies," *JMIR Med. Informat.*, vol. 8, no. 2, Feb. 2020, Art. no. e16492, doi: 10.2196/16492.

[18] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–40, Dec. 2020.

[19] R. Heyburn, R. R. Bond, M. Black, M. Mulvenna, J. Wallace, D. Rankin, and B. Cleland, "Machine learning using synthetic and real data: Similarity of evaluation metrics for different healthcare datasets and for different algorithms," in *Proc. Data Sci. Knowl. Eng. Sens. Decis. Support*, Sep. 2018, pp. 1281–1291.

[20] M. Hittmeir, A. Ekelhart, and R. Mayer, "On the utility of synthetic data: An empirical evaluation on machine learning tasks," in *Proc. 14th Int. Conf. Availability, Rel. Secur.*, Aug. 2019, pp. 1–6.

[21] K. E. Emam, L. Mosquera, and C. Zheng, "Optimizing the synthesis of clinical trial data using sequential trees," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 1, pp. 3–13, Jan. 2021, doi: 10.1093/jamia/ocaa249.

[22] L. Xu, "Synthesizing tabular data using conditional GAN," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2020. [Online]. Available: https://hdl.handle.net/1721.1/128349

[23] K. E. Emam. (2021). *Could Synthetic Data Be the Future of Data Sharing*. CPO Magazine. Accessed: Sep. 28, 2021. [Online]. Available: https://www.cpomagazine.com/data-privacy/could-synthetic-data-be-the-future-of-data-sharing/

[24] H. Ping, J. Stoyanovich, and B. Howe, "Datasynthesizer: Privacy-preserving synthetic datasets," in *Proc. 29th Int. Conf. Sci. Stat. Database Manage.*, 2017, pp. 1–5.

[25] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2016, pp. 399–410, doi: 10.1109/DSAA.2016.49.

[26] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke creation of synthetic data in R," *J. Stat. Softw.*, vol. 74, no. 11, pp. 1–26, Oct. 2016.

[27] B. Nowok, "Utility of synthetic microdata generated using tree-based methods," Administ. Data Res. Centre, Univ. Edinburgh, Edinburgh, Scotland, 2015.

[28] M. Templ, B. Meindl, A. Kowarik, and O. Dupriez, "Simulation of synthetic complex data: The R package simPop," *J. Stat. Softw.*, vol. 79, no. 10, pp. 1–38, 2017, doi: 10.18637/jss.v079.i10.

[29] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 3, pp. 230–238, Mar. 2018, doi: 10.1093/jamia/ocx079.

[30] R. J. Lewis, "An introduction to classification and regression tree (CART) analysis," in *Proc. Annu. Meeting Soc. Academic Emergency Med.*, vol. 14. San Francisco, CA, USA, 2000, pp. 1–14.

[31] J. Drechsler and J. Reiter, "Synthetic data: Balancing data confidentiality & quality in public use files," 2019.

[32] J. Miranda and L. Vilhuber, "Using partially synthetic microdata to protect sensitive cells in business statistics," *Stat. J. IAOS*, vol. 32, no. 1, pp. 69–80, Feb. 2016, doi: 10.3233/SJI-160963.

[33] M. Hittmeir, A. Ekelhart, and R. Mayer, "Utility and privacy assessments of synthetic data for regression tasks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 5763–5772.

[34] G. M. Raab, B. Nowok, and C. Dibben, "Guidelines for producing useful synthetic data," 2017, *arXiv:1712.04078*.

[35] F. K. Dankar and M. Ibrahim, "Fake it till you make it: Guidelines for effective synthetic data generation," *Appl. Sci.*, vol. 11, no. 5, p. 2158, Feb. 2021, doi: 10.3390/app11052158.

[36] *Practical Synthetic Data Generation [Book]*. Accessed: Sep. 6, 2020. [Online]. Available: https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/

[37] L. L. Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. New York, NY, USA: Springer, 2000, doi: 10.1007/978-1-4612-1166-2.

[38] A. Dandekar, R. A. M. Zen, and S. Bressan, "Comparative evaluation of synthetic data generation methods," in *Proc. ACM Conf. (Deep Learn. Secur. Workshop)*, 2017.

[39] *General and Specific Utility Measures for Synthetic Data Snoke 2018 Journal of the Royal Statistical Society: Series A (Statistics in Society) Wiley Online Library*. Accessed: Nov. 19, 2020. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12358

[40] M.-J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr, "Global measures of data utility for microdata masked for disclosure limitation," *J. Privacy Confidentiality*, vol. 1, no. 1, pp. 1–14, Apr. 2009, doi: 10.29012/jpc.v1i1.568.

[41] *Synthetic Data Evaluation*. SDV. Accessed: Oct. 14, 2021. [Online]. Available: https://sdv.dev/SDV/user_guides/evaluation/

[42] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. New York, NY, USA: Springer, 2011, doi: 10.1007/978-1-4614-0326-5.

[43] B. J. Erickson and F. Kitamura, "Magician's corner: 9. Performance metrics for machine learning models," *Radiol., Artif. Intell.*, vol. 3, no. 3, May 2021, Art. no. e200126, doi: 10.1148/ryai.2021200126.

[44] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation," *Arch. Comput. Methods Eng.*, vol. 29, pp. 313–333, Apr. 2021.

[45] K. El Emam, "Seven ways to evaluate the utility of synthetic data," *IEEE Secur. Privacy*, vol. 18, no. 4, pp. 56–59, Jul. 2020.

[46] *UCI Machine Learning Repository*. Accessed: Oct. 14, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[47] J. Vanschoren. *OpenML*. OpenML: Exploring Machine Learning Better, Together. Accessed: Oct. 14, 2021. [Online]. Available: https://www.openml.org

[48] *DATASPHERE*. Datasphere. Accessed: Oct. 14, 2021. [Online]. Available: https://www.datasphere.online/en/

[49] *Real-World Data Solution | Cerner*. Accessed: Oct. 14, 2021. [Online]. Available: https://www.cerner.com/solutions/real-world-data

[50] *Kaggle: Your Machine Learning and Data Science Community*. Accessed: Oct. 14, 2021. [Online]. Available: https://www.kaggle.com/

[51] *Cloud Services Amazon Web Services (AWS)*. Accessed: Oct. 14, 2021. [Online]. Available: https://aws.amazon.com/

[52] (May 4, 2021). *Meet the Synthetic Data Vault*. SDV Blog. Accessed: Oct. 17, 2021. [Online]. Available: https://sdv.dev/blog/intro-to-sdv/

[53] N. J. de Vos, *Kmodes: Python Implementations of the k-Modes and k-Prototypes Clustering Algorithms for Clustering Categorical Data*. Accessed: Oct. 14, 2021. [Online]. Available: https://github.com/nicodv/kmodes

[54] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.

FIDA K. DANKAR received the Ph.D. degree in computer science from the University of Ottawa, Canada. She is currently an Associate Professor at United Arab Emirates (UAE) University. Prior to joining UAE University, she was a Research Scientist at the IBM Canada Research and Development Centre, where she worked on the secure and private mining of distributed health data. Prior to joining IBM, she worked for eight years at the Children's Hospital of Eastern Ontario Research Institute, her work centered around facilitating the sharing of health information for secondary purposes while protecting the privacy of patients and the identity of providers. Her research interests include the development of multidisciplinary approaches for the private and secure mining/sharing of health data that uses methods from cryptography, biomedical knowledge modeling, and policy analysis.

MAHMOUD K. IBRAHIM received the Bachelor of Electrical Engineering degree from the American University of Beirut, in 2020. He is currently pursuing the Advanced Masters of Artificial Intelligence degree with Katholieke Universiteit (KU) Leuven, Belgium. Prior to starting his master's studies, he was a Research Assistant at the United Arab Emirates University, working on data privacy and anonymity problems. He is an AI Researcher with focus on data privacy, anonymity, uncertainty, and deep learning.

LEILA ISMAIL (Member, IEEE) is the Founder and the Director of the Intelligent Clouds and Distributed Computing Systems (INDUCE) Research Laboratory, College of Information Technology, United Arab Emirates University (UAEU), and an Associate Professor at the Department of Computer Science and Software Engineering. She has vast industrial and academic experience at the Sun Microsystems Research and Development Center, France, working the design and implementation of highly available distributed systems, and participated in the deposit of a U.S. patent. She served in teaching at Grenoble I, France, and has been serving as an Adjunct Professor at the Digital Ecosystems and Business Intelligence Institute, Curtin University, Australia. She has been very active in creating smart and efficient digital ecosystems responding to now-a-days emergency needs for better living in our dynamic global habitat, introducing blockchain, the Internet of Things (IoT), machine learning, deep learning, and artificial intelligence approaches for different applications domains, such as healthcare, energy, and smart transportation. She was the recipient of several awards and appreciation certificates, including the IBM Shared University Research Award (SURA) and the IBM Faculty Award, very competitive worldwide, and the UAE University Award for high achievements publishing in top ranked journals. She won funding for major projects as a principal investigator on grid and cloud computing, intelligent systems, and smart applications, and awards of top achievements. She has been very active in international research collaborations within Australia and USA, and has been invited as a keynote speaker in several conferences, including the Women in Data Science International Conference (WiDS 2021), organized by Stanford University. She has been participating in the success of many IEEE and ACM international conferences in several roles, such as the general chair, the organizing committee chair, and the technical program chair. She served as an Associate Editor for the *International Journal of Parallel, Emergent and Distributed Systems* for several years. She is the author of many scientific publications in journals and conferences, and the Editor of the book *Information Innovation Technology in Smart Cities* (Springer Nature).