BMC Medical Informatics and Decision Making

SOFTWARE

Open Access

How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data

Gabriele Santangelo[1*], Giovanna Nicora[1], Riccardo Bellazzi[1] and Arianna Dagliati[1]

Abstract

Background The exponential growth in patient data collection by healthcare providers, governments, and private industries is yielding large and diverse datasets that offer new insights into critical medical questions. Leveraging extensive computational resources, Machine Learning and Artificial Intelligence are increasingly utilized to address health-related issues, such as predicting outcomes from Electronic Health Records and detecting patterns in multi-omics data. Despite the proliferation of medical devices based on Artificial Intelligence, data accessibility for research is limited due to privacy concerns. Efforts to de-identify data have met challenges in maintaining effectiveness, particularly with large datasets. As an alternative, synthetic data, that replicate main statistical properties of real patient data, are proposed. However, the lack of standardized evaluation metrics complicates the selection of appropriate synthetic data generation methods. Effective evaluation of synthetic data must consider resemblance, utility and privacy, tailored to specific applications. Despite available metrics, benchmarking efforts remain limited, necessitating further research in this area.

Results We present SynthRO (Synthetic data Rank and Order), a user-friendly tool for benchmarking health synthetic tabular data across various contexts. SynthRO offers accessible quality evaluation metrics and automated benchmarking, helping users determine the most suitable synthetic data models for specific use cases by prioritizing metrics and providing consistent quantitative scores. Our dashboard is divided into three main sections: (1) Loading Data section, where users can locally upload real and synthetic datasets; (2) Evaluation section, in which several quality assessments are performed by computing different metrics and measures; (3) Benchmarking section, where users can globally compare synthetic datasets based on quality evaluation.

Conclusions Synthetic data mitigate concerns about privacy and data accessibility, yet lacks standardized evaluation metrics. SynthRO provides an accessible dashboard helping users select suitable synthetic data models, and it also supports various use cases in healthcare, enhancing prognostic scores and enabling federated learning. SynthRO's accessible GUI and modular structure facilitate effective data evaluation, promoting reliability and fairness. Future developments will include temporal data evaluation, further broadening its applicability.

Keywords Synthetic data, Synthetic Tabular Data evaluation, Benchmarking Tools, Utility and privacy evaluation, Electronic Health Records (EHRs)

*Correspondence:
Gabriele Santangelo
gabriele.santangelo01@universitadipavia.it

[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

Background

Synthetic-data-driven applications

The gathering of patients' data by healthcare providers, governments, and private industry is growing at an exponential pace, encompassing vast volumes and diverse types of abundant information that can give novel insights into essential medical questions [1]. Thanks to the increasing availability of large computer infrastructure and computational resources, Machine Learning (ML) and Artificial Intelligence (AI) are progressively being exploited to solve health-related problems, such as prognosis prediction from Electronic Health Records (EHRs) or pattern detection in multi-omics data. ML and AI approaches are increasingly being translated from bench to bedside, with 171 enabled AI-based medical devices from the Food and Drug Administration (FDA) as of October 2023 [2].

Despite their potential value, such datasets are mainly inaccessible to wider research communities, mainly due to concerns about patient's privacy. Even when access is granted, for instance within federated research networks, the process of ensuring proper data usage and protection poses significant delays to research progress and slows the translation of solutions based on these data from bench to bedside. A possible solution to address these challenges is to de-identifying patient data through various means, such as removing identifiable features, adding noise, or grouping variables. However, the effectiveness of such methods, especially with large datasets, remains uncertain, leaving open the possibility of patient re-identification when combined with other datasets. To this end, there has been a growing proposition to replace original data, derived from real patients, with the use of Synthetic Data (SD) that mimics the main statistical characteristics of their real counterparts.

Initially proposed by Rubin [3] and Little [4], SD refers to artificially generated datasets replicating real data distributions and structures. SD can also be defined as microdata records generated by a statistical model of the original data: the model allows the sampling of new data values that replicate the statistical properties of the original data [5]. Finally, according to Alan Turing Institute, "synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)". This definition underscores the broader aim of SD, which is to address various data science tasks [6]. A related aim of SD is to ensure privacy and preserve information from the original dataset while enhancing model training, especially in healthcare applications.

The usage of SD has multiple applications in healthcare from screening clinical intervention, to enhance ML pipelines for predictive analytics, to finetune models for specific populations [7]. In ML, they can be exploited to train classifiers and/or to evaluate their performance [8, 9]. SD can also drive a change in distributed research networks, allowing a shift from meta-learning models to direct data sharing and application of patient-level modeling techniques. Researchers have extensively explored the use of SD within the federated learning domain [10]. For instance, Azizi et al. investigated its role in assessing country-level disparities in cardiovascular diseases when compared to federated analysis based on real data [11]. Initiatives like the Open-CESP project by the French Centre de recherche en Epidémiologie et Santé des Populations have aimed to provide public access to synthetic datasets derived from research efforts [12]. These endeavors often focus on refining frameworks for generating and evaluating SD to maintain fidelity to ground truth data while ensuring privacy [13]. Notable approaches include generating SD modeled on United Kingdom primary care data [14], centralizing local simulated instances for AI model training [15], and integrating generative methods with federated learning for privacy-preserving high-quality SDG [16].

Generation and evaluation of synthetic tabular data

While unstructured data (i.e. images and free text) plays a crucial role in current ML-based systems, tabular data is still the predominant type of data used to develop models to aid healthcare decision making, still offering the most valuable opportunities to develop AI-based healthcare systems [17].

Several approaches exist for generating tabular synthetic patient data. Some methods rely on domain-specific knowledge, restricting their adaptability to different contexts. In contrast, entirely data-driven methods learn directly from the data, rendering them more versatile and easily applicable to diverse scenarios. Hernandez and colleagues [18] provide a systematic review of the available approaches for synthetic tabular data generation. The approaches are categorized into three main groups: (i) classical approaches, which include baseline methods and statistical and supervised ML approaches; (ii) deep learning approaches, where the generative model is realized using deep learning; lastly, (iii) approaches that do not fall into the previous categories (e.g. SDG methods that simulate a series of procedures). In the deep learning realm, Generative Adversarial Networks (GANs) have emerged as key ML-based generative models in healthcare. While GAN algorithms have been adapted for clinical tabular data, they often overlook the unique characteristics of such data. Ensuring the preservation of logical relationships and achieving class-balanced tabular SD are recognized as crucial factors in enhancing model performance and stability [19].

The wide array of methods available for SDG, coupled with the absence of standardized evaluation metrics,

present challenges in determining the most suitable approach. Evaluating SD as "good" or "appropriate" cannot be done universally, as their suitability may vary depending on the context and application. For instance, criteria such as statistical fidelity at both individual and population levels, as well as privacy disclosure, may be prioritized differently based on specific tasks. Within the process of SDG through data-driven methods, it is crucial to provide researchers with guidance in selecting methodologies tailored to their specific needs.

Metrics and methods for evaluating SD can be categorized into different dimensions to assist developers in selecting the most appropriate approach for their specific application requirements. These metrics are typically grouped into three main categories [17, 20, 21]:

Resemblance metrics These metrics serve as an initial validation step for SD, acting as a preliminary quality control measure before more comprehensive evaluations. Generally, they employ statistical approaches to analyze whether the correlation structure, both univariate and multivariate, among the features of the original dataset is preserved in the synthetic dataset.

Utility metrics These metrics assess the usability of the outcomes derived from ML models trained on SD. The evaluation involves comparing the performance achieved using the original dataset with that of the synthetic dataset.

Privacy metrics SD that closely resemble real records pose a risk of inference, potentially exposing the original data used for training the generative model. Therefore, evaluating the privacy of SD is crucial for minimizing the potential for data disclosure. Typically, metrics used in literature aim to evaluate the security of SD regarding the disclosure risk of private or sensitive information, including simulating cyberattacks.

Benchmarking synthetic data

While evaluation metrics provide a robust framework for comparing different datasets, there is a notable gap in benchmarking efforts, raising several types of concern.

Evaluation metrics need to be selected according to the final application of the SD. For instance, within federated healthcare projects, SD can be shared among participants. In this context, privacy should be highly prioritized. In the development phase of a ML model, utility should be optimized, as the goal is to implement ML models having consistent performance in comparison with models trained on the original dataset. Resemblance is of extreme importance when developing Clinical Decision Support Systems. The interplay among these three aspects has been studied in [22], highlighting the positive correlation between resemblance and utility and the potential trade-off between resemblance/utility and privacy.

Yan and colleagues [22] highlight the necessity for benchmarking frameworks to determine the most suitable tabular SDG models for specific use cases within given datasets. Their motivations include a lack of consensus regarding suitable evaluation metrics for assessing SD, which is crucial for effectively comparing and contrasting synthesis models; the diverse range of use cases for SD, which introduces varying priorities regarding the preservation of different data aspects; the inherent instability in training trajectories of generative models, which can yield disparate models and inconsistencies in the quality of generated data.

The literature on benchmarking approaches for SD remains scarce, focusing more on evaluating individual applications and methods. However, some notable studies have emerged. In [23], authors evaluate utility metrics to rank SD generation methods based on their performance in analytical workloads. More in detail, multiple metrics were tested across 30 health datasets and three generation methods (Bayesian network, GANs, sequential tree synthesis) to compare SD performances against real data. Other frameworks [24] evaluate the quality of differentially private SD from applied researchers' perspectives or outline [25] criteria for evaluating masked data, categorizing utility metrics and their fidelity at different levels, ranging from attributes to population distributions and finally to applications.

To the best of our knowledge, the only existing work that provides robust insight into SD benchmarking is the paper by Yan and colleagues [22]. Their results highlight a utility-privacy tradeoff in sharing synthetic health data, indicating that no single method emerges as the optimal choice across all criteria in every use case. This highlights again the importance of assessing SDG methods within their specific contexts. This work introduces a rank-based scoring mechanism that aggregates individual metric scores into a final score model, facilitating the consideration of competing evaluation metrics.

Synthetic data assessment tools

As the usage of SD within various contexts is increasing, tools for SD assessment are needed. We advocate four relevant aspects that a tool for SD evaluation should fulfill:

1) It should allow for the evaluation of all the evaluation categories (resemblance, privacy, and utility);
2) It should have a user-friendly Graphical User Interface (GUI) to allow for SD evaluation across different types of users, both programmers and end-users, such as researchers and clinicians;

3) It should provide a report with the results of the evaluation;
4) It should allow for benchmarking that the user can calibrate according to specific use cases.

Several open-source tools have been developed to generate and evaluate SD, each offering distinct features (Table 1). In [26], we presented a dashboard that incorporates these metrics for SD evaluation. This user-friendly dashboard application is designed to facilitate a comprehensive assessment of SD quality. With an intuitive interface, users can easily access and utilize functionalities to generate detailed reports. The dashboard enables to conduct general and qualitative analyses of synthetic datasets obtained through statistical or generative methods.

While existing tools offer various methods for SD evaluation, almost none specifically address benchmarking for ranking SD datasets in terms of multiple metrics and choosing them based on the preferred use case, in which one wants to prioritize one aspect (e.g. utility or privacy) over another.

Therefore, we introduce SynthRO (Synthetic data Rank and Order), a practical tool designed to benchmark health synthetic tabular data for various contexts and use cases. SynthRO provides a user-friendly interface for evaluating synthetic datasets, addressing the need for accessible quality evaluation metrics, and it automatically allows for SD benchmarking. In addition to the interface implementation, we emphasize the importance of our benchmarking approach, particularly its interface-level integration, a feature currently lacking in existing synthetic data evaluation tools. SynthRO implements a streamlined mechanism to determine which SD models are most appropriate for which use case for a given dataset. Users can prioritize the different metrics and get quantitative scores that enable consistency in SD evaluations. This represents a valuable contribution with potential applications extending beyond academic contexts.

Additionally, to test and validate the usability of our dashboard, various users were asked to use the dashboard and fill in a System Usability Scale (SUS) [27].

Implementation

Dashboard architecture

The SynthRO platform has been developed using the Python programming language. Specifically, the Dash package [36] has been utilized. Dash is a Python framework designed for the development of interactive, web-based data visualization applications. It facilitates rapid prototyping and iteration while offering extensive customization options for user interface design and functionality, by providing predefined interactive components such as graphs, tables, and input controls that can be customized. It also allows for easy linking of various components and dynamic updating of the user interface in response to user actions or changes in data. Users can easily use our dashboard locally.

The dashboard is divided into three main sections (see Fig. 1):

1. **Loading data** section, where users have the ability to locally upload the necessary files for quality analysis;
2. **Evaluation** section, where various analyses of quality assessment are conducted by calculating various metrics and measures;
3. **Benchmarking** section, where users can globally assess the quality of SD.

The modular design of our dashboard ensures that its sections are easily extendable. This flexibility allows for the seamless addition of new metrics introduced in the literature, such as fairness or Carbon footprint [37], enabling users to incorporate these metrics into the benchmarking process as they become relevant. Consequently, our tool

Table 1 List of open-source tools that perform an evaluation of synthetic data utility, resemblance, and privacy

Tool	Description	Metrics	GUI	Report	Benchmarking
Synthetic Data Vault [28] (SDGym and SDGMetrics module)	Python package to generate and evaluate	resemblance, privacy	no	yes	no
SDNist [29]	Python package for evaluation	resemblance, utility, privacy	no	yes	no
Anonymeter [30]	Python package to evaluate privacy	privacy	no	no	no
SynthGauge [31]	Python package	utility, privacy	no	no	no
synthpop [32]	R library	utility	no	no	no
Gretel.ai [33]	Dashboard to generate and summary evaluate	resemblance, privacy	yes	yes	no
SynthEval [34]	Python package for evaluation	resemblance, utility, privacy	no	no	no
Synthcity [35]	Python package for evaluation	resemblance, utility, privacy	no	no	yes
SynthCheck [26]	Python package for evaluation	resemblance, utility, privacy	yes	yes	no
SynthRO (our proposed tool)	**Python package for evaluation and benchmarking**	**resemblance**, **utility**, **privacy**	**yes**	**yes**	**yes**

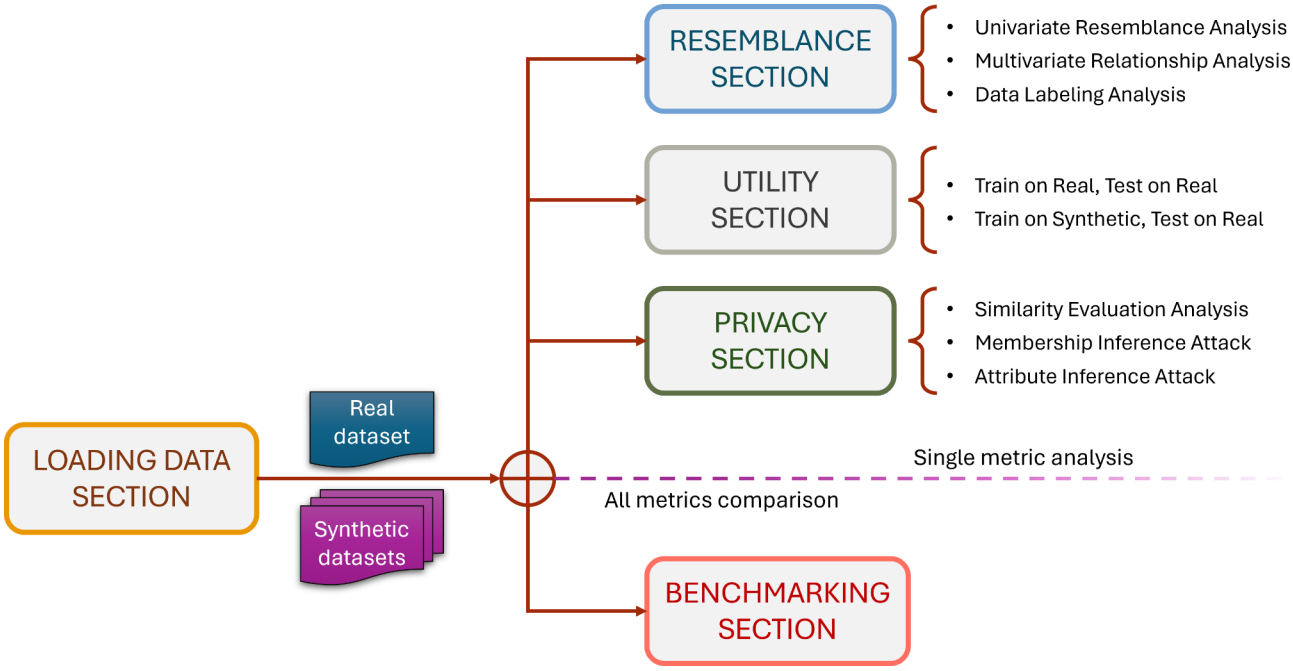

Fig. 1 Overall schematic diagram of the platform

can adapt to evolving evaluation criteria, maintaining its utility and comprehensiveness over time.

As shown in Fig. 1, the **Resemblance section** is categorized into three subsections: Univariate Resemblance Analysis (URA), Multivariate Relationships Analysis (MRA), and Data Labeling Analysis (DLA). URA assesses the preservation of univariate statistical properties from the original data within SD, using statistical tests and distance metrics. MRA determines whether SD replicates the original data's statistical properties in a multidimensional context. In DLA, several classifiers are trained to recognize whether the proposed record is original or synthetic.

In the **Utility section**, two approaches are performed:

a) "Train on Real Test on Real" (TRTR) approach, in which a classifier is trained on a portion of the original dataset and then evaluated using a test set from the same real dataset;
b) "Train on Synthetic Test on Real" (TSTR) approach, in which the same classifier is trained on SD but tested on real data.

In the end, performance metrics are compared between the two approaches.

Finally, privacy preservation is measured in the **Privacy section**, with two different analyses: the first analysis is called Similarity Evaluation Analysis (SEA), while the second includes simulating two different cyberattacks, i.e. Membership Inference Attack (MIA) and Attribute Inference Attack (AIA). SEA involves computing the distance between the original dataset and synthetic dataset records. MIA is a simulated attack in which an attacker has access to a portion of the original dataset and attempts to identify the records that are part of the training set used for SDG. In AIA simulation, the attacker aims to reconstruct one feature from a partial original dataset using ML models trained on a synthetic dataset. In general, if reconstruction is inaccurate, it suggests the preservation of privacy.

A preliminary description of the metrics implemented in the Evaluation section can be found in [26], while a detailed description of the procedure followed for comparison of synthetic datasets is provided in this article.

Benchmarking algorithm

In this paper, we have implemented the benchmarking framework developed by Yan et al. [22] in a specific section of the dashboard. All the performance metrics for evaluating SD are computed for several synthetic datasets to be compared. Then, for each metric, a ranking list based on the value assumed by that specific metric is generated: the lower the rank value assigned to the dataset, the better the performance on the given metric. To obtain a final score for the synthetic dataset, a weighted sum of all the obtained ranks considering the various metrics is performed, where the weights were tailored to a specific use case suggested by the authors.

In our platform, the user is allowed to select the relevant metrics to be calculated (see Benchmarking section). For each selected metric, a rank is assigned to each synthetic dataset (lower is better), and in the case of a

tie, the same rank is assigned (see Table 2 for the criteria used to assign ranks for each metric).

An upward arrow indicates that the lower the value considered, the lower the assigned rank (indicating better quality of the synthetic dataset).

Subsequently, as illustrated in Fig. 2, the obtained ranking lists are sorted into the three evaluation categories (resemblance, utility, and privacy), and within each category, the Cumulative Distribution Function (CDF) is calculated for each synthetic dataset, as the fraction of cases where the synthetic dataset was ranked in the first k positions. Furthermore, for each obtained curve, the area under the curve (AUC) has been calculated: the higher its value, the better the quality of the synthetic dataset. Finally, the calculation of the CDF and its related AUC is also performed considering all ranking lists, regardless of category.

The platform allows to allocate weights (between 0 and 1) as the final step of the analysis, to assign different degrees of importance to each evaluation category, which may vary according to the final goal of the SDG for the specific user problem. Weights are organized hierarchically, as shown in Fig. 3. A different weight can be assigned to each evaluation category, but only if at least one metric belonging to the considered category has been selected. Thus, at most, three different weights can be assigned in the first level of the hierarchy, and their sum must be equal to 1. The second level of the hierarchy includes the single selected metrics, divided by evaluation category. The values assumed by the weights are always between 0 and 1 but should be interpreted as a percentage of the value set in the main category: for example, if three resemblance metrics have been chosen and a weight of 0.6 has been assigned, while the three metrics have been assigned values of 0.5, 0.2, 0.3, they should be interpreted as 0.5*0.6 = 0.3, 0.2*0.6 = 0.12, and 0.3*0.6 = 0.18, respectively. At both levels, when the value of a weight is modified, the others will adjust accordingly to satisfy the sum constraint: using the previous example with values at the second level, if the first value is changed from 0.5 to 0.6, there will be a negative deviation in the sum of weights 1-(0.6 + 0.2 + 0.3)=-0.1, this difference will be equally distributed among the weights of the two unchanged metrics – 0.1/2=-0.05, so the adjusted value for the second and third metrics will be 0.2 – 0.05 = 0.15 and 0.3 – 0.05 = 0.25, respectively. Furthermore, it is possible to lock the value of a metric, preventing it from being considered in the automatic update of weights. Finally, once the weights have been selected, the weighted final score is calculated as follows: all the ranks in the ranking list of a metric will be multiplied by their respective weight, and then, for each synthetic dataset, all the weighted ranks assigned to it will be summed. Again, the lower the value, the higher the quality of the synthetic dataset.

Parallel computing

The various metrics and analyses implemented in the platform have to be calculated for all synthetic datasets uploaded to the platform whenever requested by the user. This can potentially slow down obtaining results if the datasets are processed sequentially. To address this, the Python package "multiprocessing", which provides support for parallel execution of tasks by spawning multiple processes and facilitates the management of multiple worker processes, is used to manage the calculation of various metrics for different datasets in parallel, thereby obtaining the desired results in less time.

Results

Use case

To validate the platform, we analyzed five synthetic datasets: three SD obtained using the HealthGAN SDG method implemented by Yale et al. [22], which consists of a modified version of a GAN; one SD was obtained using SDV method described in [38], in which SD are

Table 2 Implemented criteria for assigning ranks in a ranking list, divided by metric

Metrics Categorization			Low Rank Criteria
Resemblance	URA	Numerical tests	% accepted features ↑
		Categorical tests	% rejected features ↑
		Distance metrics	Average distance ↓
	MRA	Correlation matrices	Average matrices difference ↓
		Contingency tables	Average tables difference ↓
		Principal Component Analysis	RMSE real vs. synthetic ↓
	DLA		Average F1 score ↓
Utility		TRTR vs. TSTR	Average F1 scores difference ↓
Privacy	SEA	Cosine similarity	Average distance ↓
		Euclidean distance	Average distance ↑
		Hausdorff distance	Average distance ↑
	MIA		Accuracy ↓
	AIA		Accuracy ↓ RMSE ↑

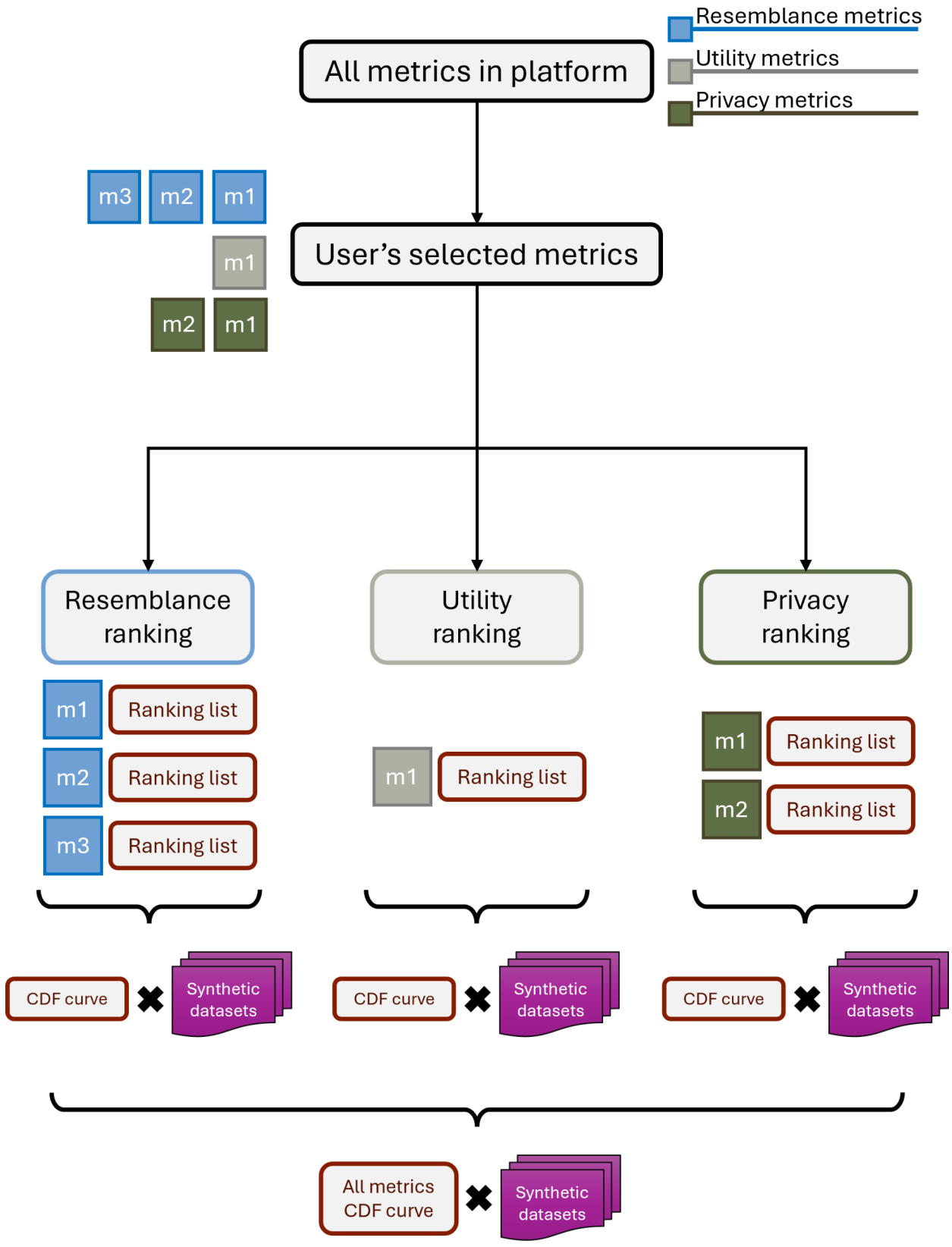

Fig. 2 Example of the workflow for generating ranking lists and CDF curves. In the figure, the user has selected three resemblance metrics (blue squares), one utility metric (gray square), and two privacy metrics (green squares). Therefore, the generated ranking lists total 3 + 1 + 2 = 6, while the number of CDF curves depends on the number of synthetic datasets uploaded

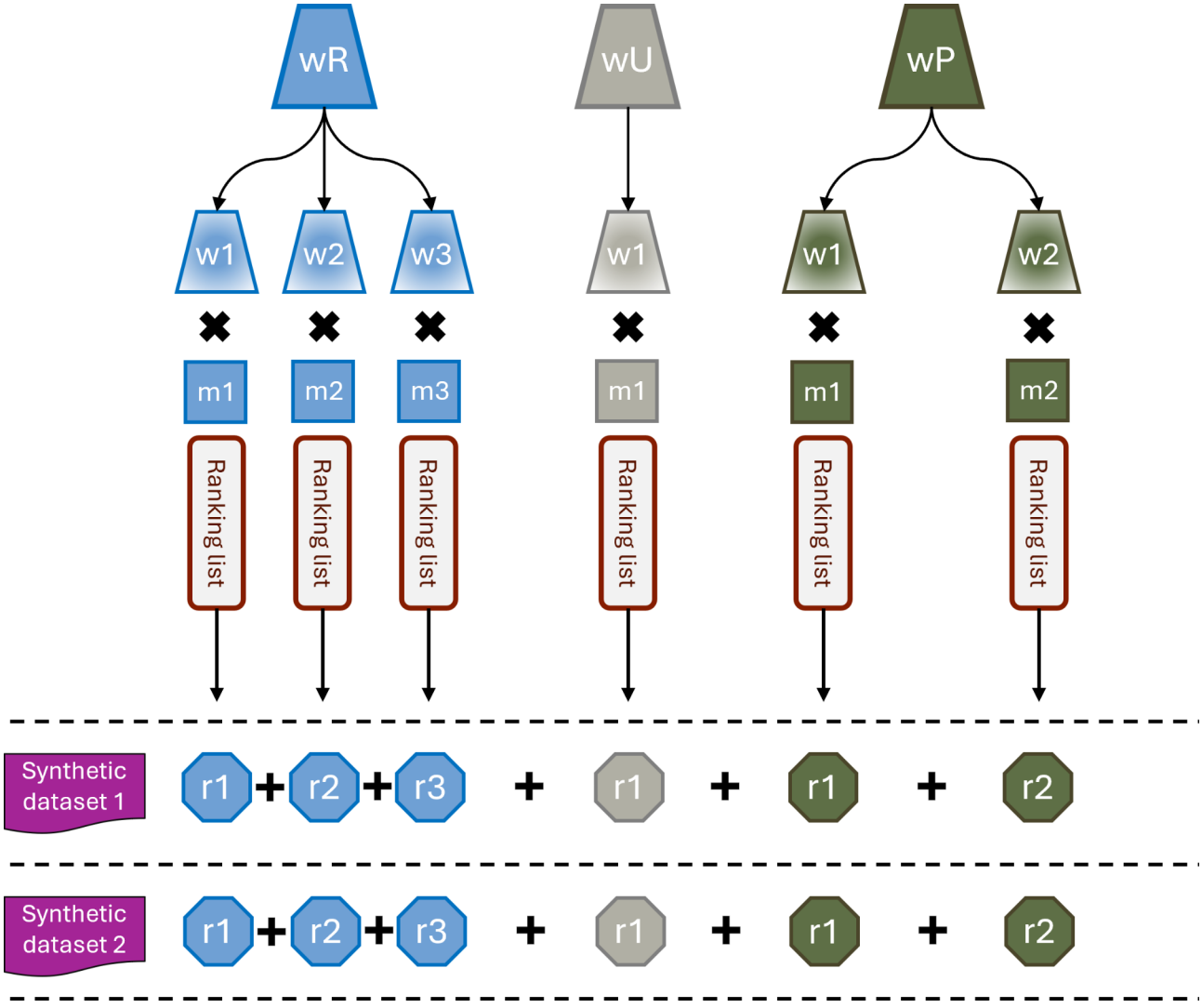

Fig. 3 Overview of the final score calculation with weighted metrics. For each synthetic dataset, the score is obtained by summing the ranks, contained in the various ranking lists, multiplied by the weights assigned to each metric

generated by applying statistical learning approach; the last SD is produced by sampling the univariate distribution of each feature (Baseline method).

The original dataset selected for training the generative model is extracted from the MIMIC-II dataset. This dataset [39] includes vital signs and diverse clinical data from 12,000 Intensive Care Unit (ICU) patients. For each patient, up to 42 variables were recorded at least once during the first 48 h after ICU admission. Aggregated features were obtained as detailed in [40], followed by the removal of features with at least 70% missing values. The resulting dataset comprises 109 features and 6,000 records, some of which contain missing values. Prior to generating SD, the dataset was split into a training set (80%) and a test set (20%). To address the missing data, the MICE (Multivariate Imputation by Chained Equations) method [41] was used.

As illustrated in Fig. 1, the SynthRO dashboard is composed of three main sections for (i) loading synthetic data, (ii) evaluating them across different metrics, and (iii) benchmarking the results. The latter section allows users to assess the overall quality of the synthetic data according to their planned use and deployment.

Loading data section

In this section, the user has to upload, through the interface, the original dataset used in the SDG phase. Unlike [26], the platform offers the possibility to upload multiple synthetic datasets simultaneously (derived from the same original dataset), in case the user would like to perform a comparative analysis of different generation methods or perform a stability analysis of a single method. The datasets to be uploaded must be in Comma Separated Values (CSV) format, and once uploaded to the platform,

an illustrative table of the uploaded data will be shown (see Fig. 4a-b). Note that, since the application is installed locally, the uploading of data onto the platform does not imply that data is shared over the network.

Furthermore, in this section, the user must specify the data type (numerical/categorical) of each feature in the uploaded datasets. As shown in Fig. 4c, this can be done by: (i) uploading a CSV file in which, for each feature name, a label is provided indicating whether it is "numerical" or "categorical"; (ii) directly from the user interface, by manually specifying the type of each feature using dropdown menus, whose values are initialized through an automatic detection process that relies on the data type identified by Python package "pandas" at the time of loading the original dataset and the count of unique values assumed by the features. This step is necessary, as some of the implemented metrics are calculated considering only one type of data or their output may be different depending on the type considered.

Evaluation section

After the users have uploaded the data, they are able to carry out the evaluation of the synthetic data's quality. This process comprises three panels, each performing a distinct quality analysis between resemblance, utility and privacy aspect. A preliminary description of the Evaluation components can be found in [26].

Since the platform allows for evaluation of multiple synthetic datasets, each section for visualizing the results obtained from a specific analysis is organized in a tabbed interface: for each synthetic dataset, there is a tab containing the output of a single metric (tables, figures, etc.). This applies to all panels related to the Evaluation section.

From the navigation bar at the bottom, the user can explore the various sections implemented: (i) from the Resemblance dropdown, three different subsections can be accessed (URA, MRA and DLA panels); (ii) in the Utility panel, the TRTR and the TSTR approaches are implemented; (iii) the Privacy dropdown includes all the analysis performed for privacy evaluation organized in three different subsections (SEA, MIA and AIA panels).

In addition, each panel allows the user to download a detailed report containing the graphs and/or tables displayed within that specific panel.

Benchmarking section

This section can be accessed by the user at any time, after data loading, without the need to perform analyses present in the single metric evaluation sections. The Benchmarking section consists of a series of panels that will appear to the user as they progress in the analysis.

Initially, the user must choose the relevant metrics for ranking by selecting them via various checkboxes available in the three tabbed interfaces, one for each evaluation category (resemblance, utility and privacy). Once the metrics are chosen, the analysis can proceed by pressing a specific button. This section will always be available, even as the analysis progresses, allowing the user to view the selected metrics and make any additions or deselections. In the example shown in Fig. 5, the following metrics were selected: numerical and categorical statistical tests (from the URA-Resemblance tab), correlation matrices and contingency tables (from the MRA-Resemblance tab), TRTR-TSTR approaches (from the Utility tab), and finally, the simulation of MIA and AIA cyberattacks (from the Privacy tab). Based on these metrics, the five uploaded synthetic datasets were compared.

If the user has selected metrics that require additional data and information for their computation, a second section will allow them to add the required information; the analysis will not proceed until the user has filled in all necessary fields. In case no further details need to be specified, the platform will directly display the last section with the obtained results. This last section comprises: (i) a section showing the ranking of synthetic datasets divided by evaluation category (see Fig. 5b-c-d); (ii) a section displaying the overall ranking, considering all the selected metrics to which the same "weight" is assigned (see Fig. 5e); (iii) a section where the user can adjust the weight assigned to each previously selected metric and see how the ranking of synthetic datasets changes.

Synthetic dataset 1: Baseline, *Synthetic dataset 2-3-4*: HealthGAN, *Synthetic dataset 5*: SDV.

Specifically, the latter section contains several sliders (see Fig. 6): if at least one metric in an evaluation category has been selected, a slider will be available to modify the weight assigned to that specific category (as shown in Fig. 6a, there are three sliders for resemblance, utility and privacy); in addition, each subcategory will have its slider to allow the user to adjust the weight of individual subcategories within the main category (see Fig. 6b-c-d). The values assumed by the sliders range from 0 (low "importance") to 1 (high "importance"), although if the constraint on the sum of the weights is not satisfied (see Benchmarking algorithm section), the value of the slider and/or the others will be automatically adjusted. Additionally, the user has the option to lock the value of one or more sliders (at least two must remain unlocked).

Synthetic dataset 1: Baseline, *Synthetic dataset 2-3-4*: HealthGAN, *Synthetic dataset 5*: SDV.

Once the user has assigned the desired weights, the result of the new ranking will be displayed after pressing a specific button. Here the user can decide whether to change the weights further or start a new analysis by changing the selected metrics.

In Fig. 7, four results obtained by modifying the weights assigned to the various selected metrics are presented.

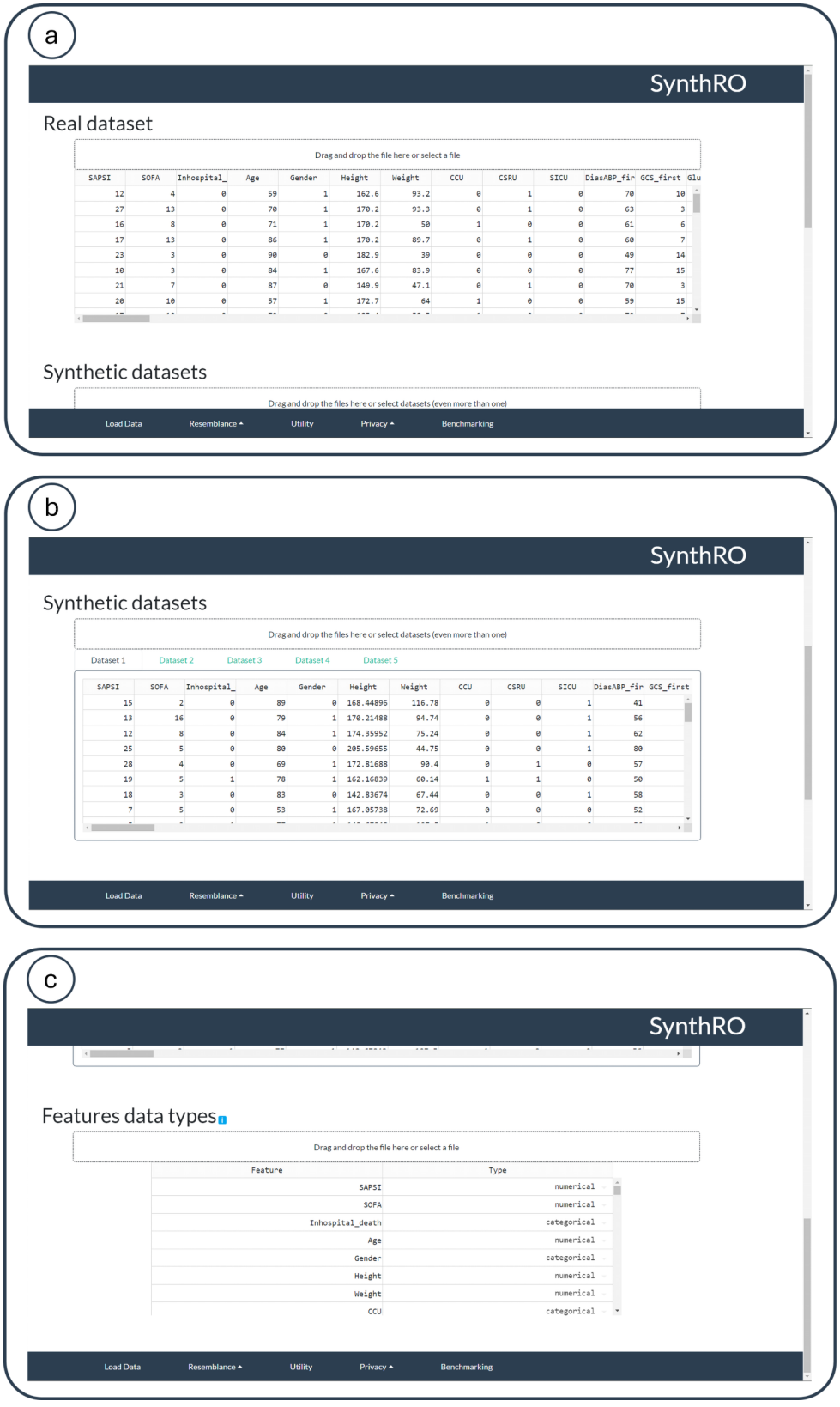

Fig. 4 **(a)** Section for uploading the real dataset and a summary table of the uploaded dataset. **(b)** Section for uploading the synthetic datasets with associated summary tables organized in a tabbed interface. **(c)** Section for specifying the types (numerical/categorical) of each feature, where a file containing this information can be uploaded or the type can be selected directly from the table

Fig. 5 Overview of the results section obtained through benchmarking analysis. **(a)** Tabbed interface for metrics selection. **(b)** Tables of the ranks assigned to the synthetic datasets for each resemblance metric and CDF plots obtained by considering only the resemblance metrics. **(c)** Table of the ranks assigned to the synthetic datasets for utility. **(d)** Tables of the ranks assigned to the synthetic datasets for each privacy metric and CDF plots obtained by considering only the privacy metrics. **(e)** CDF plots obtained by considering all metrics and a table reporting the AUC

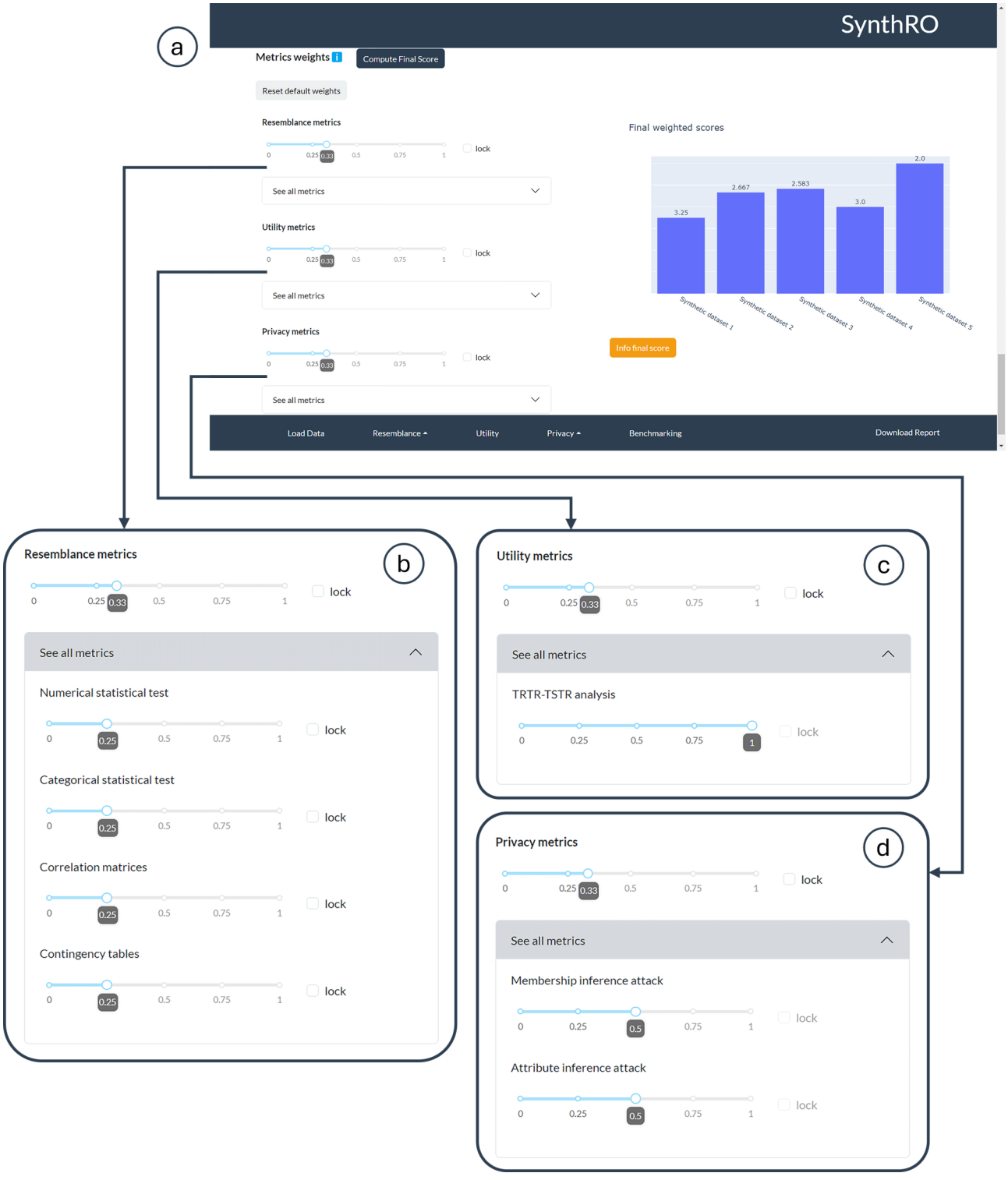

Fig. 6 Detail of the interface section where the user can modify the weights assigned to the metrics. **(a)** Main sliders for resemblance, utility, and privacy. **(b)** Secondary sliders for the resemblance metrics (in the example, four resemblance metrics were selected). **(c)** Secondary slider for utility (currently, only one utility metric is implemented on the platform). **(d)** Secondary sliders for the privacy metrics (in the example, the two cyberattacks MIA and AIA were selected)

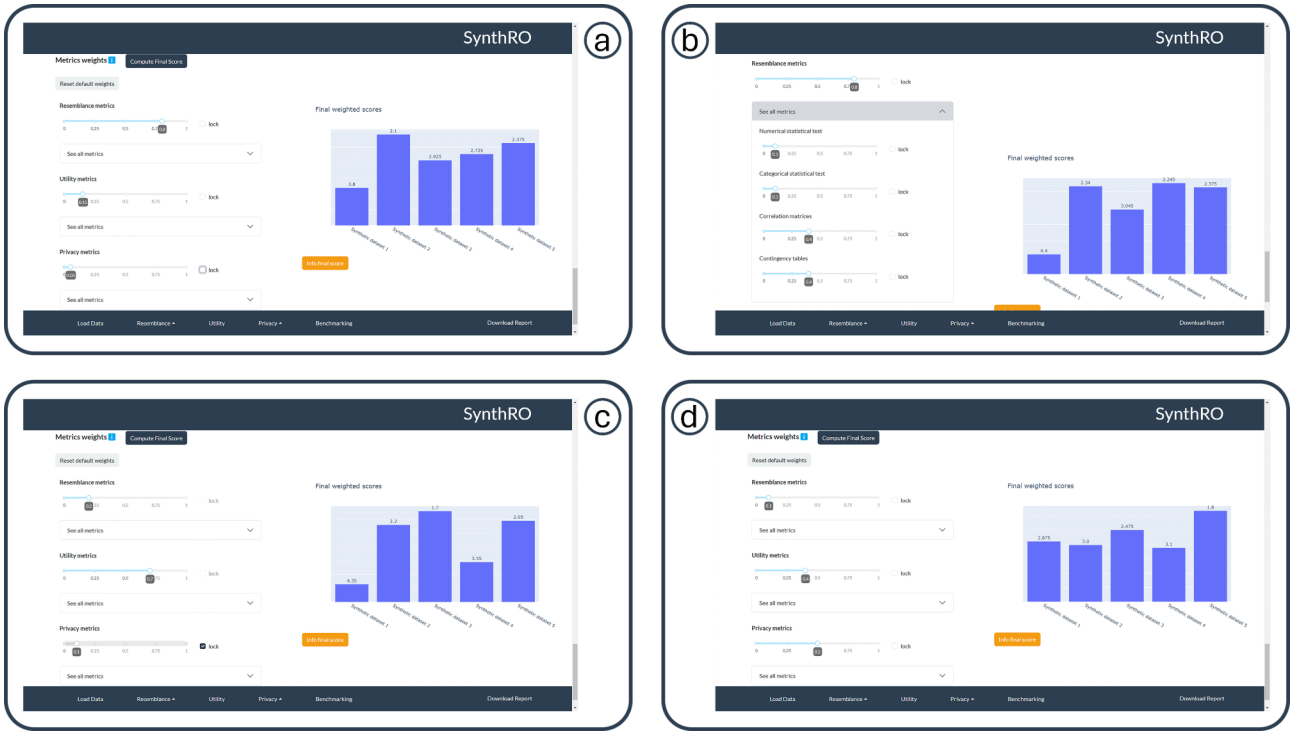

Fig. 7 Final weighted scores obtained by changing the weights assigned to the different metrics. **(a)** Case study where a greater weight was assigned to resemblance (0.8) compared to utility (0.15) and privacy (0.05). **(b)** Similar to the previous case study but with modified weights for single resemblance metrics, prioritizing multivariate ones. **(c)** Case study where more importance was given to utility (0.7) compared to resemblance (0.2) and privacy (0.1). **(d)** Case study in which privacy (0.5) and utility (0.4) were prioritized at the expense of resemblance (0.1)

In Fig. 7a, a greater weight was assigned to resemblance metrics (0.8) compared to utility (0.15) and privacy (0.05) metrics; in this case, the platform recommends the use of synthetic dataset 2 (one of those generated with HealthGAN method), as its final score is the lowest. Different results can be obtained by changing the weight of other metrics, as shown in Fig. 7c (higher weight for utility) and Fig. 7d (higher weight for privacy). As previously explained, the weights of individual metrics can be also adjusted, as in the case shown in Fig. 7b, where, among the four resemblance metrics selected for the analysis, those evaluating the preservation of similarity in a multivariate context were prioritized.

Synthetic dataset 1: Baseline, *Synthetic dataset 2-3-4*: HealthGAN, *Synthetic dataset 5*: SDV.

The computational time required for the various analyses depends not only on the size of the two datasets but also on the selected metrics. Generally, univariate metrics require less time compared to multivariate ones. To obtain the illustrated results, a computer with a 12th Gen Intel(R) Core(TM) i7-12650 H processor was used, and the time taken was 2 min and 30 s.

Report generation

A useful feature available in the dashboard is the ability to save the results of the analyses in Portable Document Format (PDF) files, which can be downloaded by the user. These reports are available in both the Evaluation section and the Benchmarking section by clicking a button located in the navigation bar. The reports contain tables and graphs obtained as results of a specific analysis. As an example, in the Supplementary materials, Fig.S1 illustrates a report available after performing a benchmarking analysis.

Usability study

We employed the SUS to assess users' experiences with our dashboard. The SUS is a standardized questionnaire comprising ten questions, developed by John Brooke [27], that allowed us to derive an overall score for perceived usability. Due to the multidisciplinary nature of this process, employing the SUS proved particularly advantageous. It facilitated an examination of non-expert users' autonomy and enabled the identification of challenges encountered by expert users. We recruited a diverse group of participants, categorized into experts (e.g., computer scientists, developers) and non-experts (e.g., medical professionals, legal experts). A total of seven participants were enrolled in the study, comprising four experts and three non-experts, with three males and four females. Further details can be found in the Supplementary materials.

The tool received an average SUS score of 83.93 (with a standard deviation of 8.76), which is a strong indicator of high usability. The results obtained by analyzing the two user categories separately show an average SUS score of 84.38 (standard deviation 8.26) for expert users, and 83.33 (standard deviation 11.27) for non-expert users.

Conclusions

The exponential growth in patient data collection offers unprecedented opportunities to address critical medical questions. However, the accessibility of this data for research purposes is significantly hindered by privacy concerns. SD, which replicates the main statistical properties of real patient data, has emerged as a promising alternative to mitigate these concerns.

Despite its potential, the lack of standardized evaluation metrics presents a significant challenge in selecting appropriate SDG methods. In response to these challenges, we developed SynthRO, a user-friendly software designed to benchmark health synthetic tabular data across various contexts.

SynthRO provides accessible quality evaluation metrics and automated benchmarking, enabling users to determine the most suitable SDG models for specific applications by prioritizing metrics and providing consistent quantitative scores. SynthRO represents a step forward in the benchmarking and quality assurance of synthetic health data. By providing a comprehensive, user-friendly tool that incorporates a wide range of evaluation metrics, SynthRO enhances the utility and reliability of SD in healthcare research. Nowadays, the literature offers numerous methods for generating synthetic datasets, including in clinical contexts. Therefore, it is crucial that users who wish to utilize synthetic data in their research and development projects have the possibility to easily and clearly compare the various generative methods available in the literature. By assigning different levels of importance to various quality evaluation criteria, users can investigate which generative method is best suited for a given synthetic data usage scenario.

We have demonstrated the software's capabilities in analyzing synthetic datasets using various methods, including baseline, HealthGAN and SDV. In particular, we have shown how adjusting the weights to prioritize one metric (utility) over another (privacy), and vice versa, can lead to the selection of different generative approaches. Even within the same method, these adjustments can result in prioritizing one dataset over another. These scenarios demonstrate how SynthRO can be applied to enhance data utility while addressing issues of privacy and data scarcity across various domains.

We also conducted a usability study to demonstrate the software's usability, involving users with varying levels of expertise, both experts and non-experts. The results indicate higher usability among expert users, suggesting that less experienced users may benefit from a brief preliminary training to fully utilize the tool's capabilities. This reflects the multidisciplinary nature of the process, where two distinct levels of expertise emerge: on the one hand, the technical expert, who understands the metrics but may lack domain-specific knowledge, and on the other, the domain expert, who understands the field of application but may not be as familiar with technical tools. Overall, the software achieved a high SUS score (83.93), which, according to the guidelines [42, 43], is an excellent result, exceeding the threshold of 68. This score confirms the robustness of the system from a general usability perspective, while also emphasizing the importance of seamless integration between technical and clinical expertise.

As with other state-of-the-art software tools [35], SynthRO can be applied to several use cases in the biomedical domain. It can serve to benchmark various generative models to rebalance and improve prognostic scores computed from EHR or biobanks, to enable data combination for federated learning approaches and applied ML models at patient level. The opportunity to leverage high-quality SD within a large research network paves theway to develop more accurate models at individual level and to better assess their reliability [44] via pointwise evaluations. This is of particular interest in order to address emerging conditions about fairness and underrepresented populations.

Robust tools that can be easily used to enhance SD quality and reliability facilitate more effective and ethical use of SD. As importantly noted in [37], future research should continue to expand on these efforts, ensuring that SD evaluations remain robust, fair, and environmentally responsible.

Our implementation of a benchmarking framework for the quality assurance of synthetic tabular data for healthcare can support effective communication and preparation for real-life model implementation. Differently from current tools, the easily accessible GUI provided by SynthRO allows users to visualize comparison results in a dashboard-style interface. This enables even non-expert users to evaluate the quality of SD across different contexts and in relation to its intended use, which is crucial for the deployment of final models.

Furthermore, SynthRO is structured in a modular way, which will allow it to easily add novel metrics to the benchmarking procedures and address key gaps in the current literature regarding the evaluation of SD. Future development will focus on expanding SynthRO's capabilities to include the evaluation of temporal data and images, further enhancing its applicability in healthcare research. Additionally, given the modularity of the dashboard, a section preceding the synthetic dataset

evaluation could be included, where various generative methods are implemented. This would allow users to generate and evaluate synthetic datasets using the same tool or when the evaluation metrics require repeated sampling from a SDG model [45]. Lastly, it may be beneficial to implement multi-user and session support in a scenario where the application is hosted on-premise within an institution to facilitate the evaluation of SD. SynthRO's framework promotes a standardized approach to these evaluations, ensuring more consistent, reliable, and reproducible results. By incorporating these elements, SynthRO not only advances the field of SD evaluation but also aligns with the broader goals of reliability and fairness in AI and ML applications in healthcare.

Abbreviations

ML	Machine Learning
AI	Artificial Intelligence
EHR	Electronic Health Records
FDA	Food and Drug Administration
SD	Synthetic Data
SDG	Synthetic Data Generation
GAN	Generative Adversarial Network
GUI	Graphical User Interface
SUS	System Usability Scale
SDV	Synthetic Data Vault
URA	Univariate Resemblance Analysis
MRA	Multivariate Relationships Analysis
DLA	Data Labeling Analysis
TRTR	Train on Real Test on Real
TSTR	Train on Synthetic Test on Real
SEA	Similarity Evaluation Analysis
MIA	Membership Inference Attack
AIA	Attribute Inference Attack
CDF	Cumulative Distribution Function
AUC	Area under the curve
ICU	Intensive Care Unit
MICE	Multivariate Imputation by Chained Equations
CSV	Comma Separated Values, PDF: Portable Document Format

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02731-9.

Supplementary Material 1

Acknowledgements

GS is a PhD student enrolled in the National PhD program in Artificial Intelligence, XXXIX cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. This work was supported by "Fit4MedRob - Fit for Medical Robotics" (Grant No. B53C22006950001).

Author contributions

GS wrote and edited the source code and the manuscript. GN and AD wrote and edited the manuscript. RB and AD supervised the study. All authors read and approved the final manuscript.

Funding

This work was supported by "Fit4MedRob - Fit for Medical Robotics" (Grant No. B53C22006950001). The project is aimed at revolutionizing current rehabilitation and assistive models by means of novel (bio)robotic and allied digital technologies.

Data availability

The MIMIC-II dataset is available at https://archive.physionet.org/physiobank/tutorials/using-mimic2/. Access to the MIMIC-II database was obtained after completing the necessary training and certification required by the data custodians.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Availability and requirements

Project name: SynthRO.
Project home page: https://github.com/bmi-labmedinfo/SynthRO.
Operating system(s): Platform independent.
Programming language: Python 3.12.
Other requirements: Refer to the GITHUB repository for package versions.
License: CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/?ref=chooser-v1).
Any restrictions to use by non-academics: Non commercial.

Received: 20 June 2024 / Accepted: 21 October 2024
Published online: 18 February 2025

References

1. Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research, *Nat. Rev. Cancer*, vol. 22, no. 11, pp. 625–639, Nov. 2022, https://doi.org/10.1038/s41568-022-00502-0
2. Health R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices, *FDA*, Oct. 2023, Accessed: Nov. 27, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
3. Rubin DB. Discussion statistical disclosure limitation. Discuss Stat Discl Limit. 1993;9(2):461–8.
4. Little RJA. Statistical Analysis of Masked Data. Stat Anal Masked Data. 1993;9(2):407–26.
5. Philpott D, editor. *A guide to Federal terms and acronyms*, Second edition. Lanham: Bernan Press, 2018.
6. Jordon J et al. Synthetic Data -- what, why and how? 2022, https://doi.org/10.48550/ARXIV.2205.03257
7. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. Npj Digit Med. Oct. 2023;6(1):186. https://doi.org/10.1038/s41746-023-00927-3.
8. Chen A, Chen DO. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. Sci Rep. Oct. 2022;12(1):17917. https://doi.org/10.1038/s41598-022-23011-4.
9. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. Npj Digit Med. Nov. 2020;3(1):147. https://doi.org/10.1038/s41746-020-00353-9.
10. Little C, Elliot M, Allmendinger R. Federated learning for generating synthetic data: a scoping review. Int J Popul Data Sci. Oct. 2023;8(1). https://doi.org/10.23889/ijpds.v8i1.2158.
11. Azizi Z, et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. Sci Rep. Jul. 2023;13(1):11540. https://doi.org/10.1038/s41598-023-38457-3.
12. Chapelle R, Falissard B. Statistical properties and privacy guarantees of an original distance-based fully synthetic data generation method, 2023, https://doi.org/10.48550/ARXIV.2310.06571
13. Haendel MA, The National COVID Cohort Collaborative (N3C). Mar,.: Rationale, design, infrastructure, and deployment, *J. Am. Med. Inform. Assoc.*, vol. 28, no. 3, pp. 427–443, 2021, https://doi.org/10.1093/jamia/ocaa196
14. Wang Z, Myles P, Tucker A, Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data UtilityPatient Privacy, in. 2019 *IEEE 32nd*

International Symposium on Computer-Based Medical Systems *(CBMS)*, Cordoba, Spain: IEEE, Jun. 2019, pp. 126–131. https://doi.org/10.1109/CBMS.2019.00036
15. Rollo C, Pancotti C, Birolo G, Rossi I, Sanavia T, Fariselli P. SYNDSURV: a simple framework for survival analysis with data distributed across multiple institutions. Comput Biol Med. Apr. 2024;172:108288. https://doi.org/10.1016/j.compbiomed.2024.108288.
16. Xin B, et al. Federated synthetic data generation with differential privacy. Neurocomputing. Jan. 2022;468:1–10. https://doi.org/10.1016/j.neucom.2021.10.027.
17. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions, *Methods Inf. Med.*, vol. 62, no. S 01, pp. e19–e38, Jun. 2023, https://doi.org/10.1055/s-0042-1760247
18. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. Neurocomputing. Jul. 2022;493:28–45. https://doi.org/10.1016/j.neucom.2022.04.053.
19. Kang HYJ, Batbaatar E, Choi D-W, Choi KS, Ko M, Ryu KS. Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy, *JMIR Med. Inform.*, vol. 11, p. e47859, Nov. 2023, https://doi.org/10.2196/47859
20. Kaabachi B, et al. Can we trust Synthetic Data in Medicine? A scoping review of privacy and Utility Metrics. Nov. 2023;28. https://doi.org/10.1101/2023.11.28.23299124.
21. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: state of the art in health care domain. Comput Sci Rev. May 2023;48:100546. https://doi.org/10.1016/j.cosrev.2023.100546.
22. Yan C, et al. A multifaceted benchmarking of synthetic electronic health record generation models. Nat Commun. Dec. 2022;13(1):7609. https://doi.org/10.1038/s41467-022-35295-1.
23. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility Metrics for Evaluating Synthetic Health Data Generation methods: Validation Study. JMIR Med Inf. Apr. 2022;10(4):e35734. https://doi.org/10.2196/35734.
24. Arnold C, Neunhoeffer M. Really Useful Synthetic Data -- A Framework to Evaluate the Quality of Differentially Private Synthetic Data, 2020, https://doi.org/10.48550/ARXIV.2004.07740
25. Dankar FK, Ibrahim MK, Ismail L. A multi-dimensional evaluation of Synthetic Data generators. IEEE Access. 2022;10:11147–58. https://doi.org/10.1109/ACCESS.2022.3144765.
26. Santangelo G, Nicora G, Bellazzi R, Dagliati A. SynthCheck: A Dashboard for Synthetic Data Quality Assessment, in *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 246–256. https://doi.org/10.5220/0012558700003657
27. Jordan PW, Thomas B, McClelland IL, Weerdmeester B, Industry, editors. 0 ed., CRC, 1996, 207–12. doi: https://doi.org/10.1201/9781498710411-35.
28. The Synthetic Data Vault. Put synthetic data to work! Accessed: Sep. 26, 2023. [Online]. Available: https://sdv.dev/
29. Task C, Bhagat K, Howarth G. Natl Inst Stand Technol Mar. 2023;13. https://doi.org/10.18434/MDS2-2943. SDNist v2: Deidentified Data Report Tool.
30. Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A Unified Framework for Quantifying Privacy Risk in Synthetic Data, *Proc. Priv. Enhancing Technol.*, vol. 2023, no. 2, pp. 312–328, Apr. 2023, https://doi.org/10.56553/popets-2023-0055
31. SynthGauge. Data Science Campus. Nov. 01, 2023. Accessed: Nov. 22, 2023. [Online]. Available: https://github.com/datasciencecampus/synthgauge
32. Raab GM, Nowok B, Dibben C. Assessing, visualizing and improving the utility of synthetic data. arXiv Nov. 2021;13. https://doi.org/10.48550/arXiv.2109.12717.
33. Noruzman A, Ghani NA, Zulkifli N. Gretel.ai: open-source Artificial Intelligence Tool to generate New Synthetic Data. Mar. 2022.
34. SynthEval. schneiderkamplab, May 16, 2023. Accessed: Nov. 23, 2023. [Online]. Available: https://github.com/schneiderkamplab/syntheval
35. Qian Z, Cebere B-C, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv. 2023. https://doi.org/10.48550/ARXIV.2301.07573.
36. plotly/dash., Plotly, May, 07. 2024. Accessed: May 07, 2024. [Online]. Available: https://github.com/plotly/dash
37. Vallevik VB, et al. Can I trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. Int J Med Inf. May 2024;185:105413. https://doi.org/10.1016/j.ijmedinf.2024.105413.
38. Patki N, Wedge R, Veeramachaneni K, The Synthetic Data Vault, in. 2016 *IEEE International Conference on Data Science and Advanced* Analytics *(DSAA)*, Montreal, QC, Canada: IEEE, Oct. 2016, pp. 399–410. https://doi.org/10.1109/DSAA.2016.49
39. Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012, presented at the Computing in Cardiology, 2012, pp. 245–248.
40. Johnson A. challenge2012. Apr. 26, 2023. Accessed: Sep. 28, 2023. [Online]. Available: https://github.com/alistairewj/challenge2012
41. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw. Dec. 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03.
42. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. Elsevier Science, 2016. [Online]. Available: https://books.google.it/books?id=USPfCQAAQBAJ
43. Hyzy M et al. Aug., System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis, JMIR Mhealth Uhealth, vol. 10, no. 8, p. e37290, 2022, https://doi.org/10.2196/37290
44. Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. J Biomed Inf. Mar. 2022;127:103996. https://doi.org/10.1016/j.jbi.2022.103996.
45. Stadler T, Oprisanu B, Troncoso C. Synthetic Data - Anonymisation Groundhog Day, presented at the Proceedings of the 31st USENIX Security Symposium, Security 2022, 2022, pp. 1451–1468.
46. MIMIC II Databases. Accessed. May 28, 2024. [Online]. Available: https://archive.physionet.org/mimic2/

Publisher's note