

Preserving Information while Respecting Privacy through an Information Theoretic Framework for Synthetic Health Data Generation

Nadir Sella^{*1}, Florent Guinot^{*1}, Nikita Lagrange², Laurent-Philippe Albou¹, Jonathan Desponds¹, Hervé Isambert²

1. Institut Roche, Boulogne-Billancourt, France
2. Institut Curie, CNRS UMR168, PSL University, Sorbonne University, Paris, 75005, France

* These authors contributed equally to this work

Correspondence and requests for materials should be addressed to nadirsella@gmail.com, florent.guinot@roche.com .

ABSTRACT

Generating synthetic data from medical records is a complex task intensified by patient privacy concerns. In recent years, multiple approaches have been reported for the generation of synthetic data, however, limited attention was given to jointly evaluate the quality and the privacy of the generated data. The quality and privacy of synthetic data stem from multivariate associations across variables, which cannot be assessed by comparing univariate distributions with the original data. Here, we introduce a novel algorithm (MIIC-SDG) for generating synthetic data from electronic records based on a multivariate information framework and Bayesian network theory. We also propose a new metric to quantitatively assess the trade-off between the Quality and Privacy Scores (QPS) of synthetic data generation methods. The performance of MIIC-SDG is demonstrated on different clinical datasets and favorably compares with state-of-the-art synthetic data generation methods, based on the QPS trade-off between several quality and privacy metrics.

INTRODUCTION

With the significant increase in patient recruitment for clinical trials and the collection of real-life health-related datasets over the past few decades, ensuring patient privacy has become of utmost importance. Indeed, facilitating new research initiatives and promoting data sharing are necessary for pushing the boundaries of biomedical research. Machine learning and deep learning approaches offer promise in synthesizing

health data while safeguarding patient privacy. However, existing regulatory standards, including the European General Directive on Data Protection (GDPR), often impose restrictions on data sharing and secondary data usage due to concerns about data security.

The primary and secondary use of sensitive data must not only comply with local legislations, but also align with the initial patient consent to safeguard personal privacy. Data publication needs to guarantee Statistical Disclosure Control (SDC), referring to techniques able to ensure that no person is identifiable from the published data. This includes two possible cases: i) identification disclosure when an attacker is able to link some data to a specific individual and ii) attribute disclosure when the attacker is able to learn new information on the subject, by using prior knowledge and the information contained in the data.

Classical anonymization techniques like k-anonymity, first introduced by Latanya Sweeney and Pierangela Samarati¹ in 1998, protect user data by minimizing the risk of re-identification, while keeping in theory a good level of data utility. K-anonymity is obtained through data suppression and data generalization, so that each person in the collection cannot be distinguished from at least k-1 individuals by using quasi-identifier features (attributes available to an adversary). Machanavajjhala and colleagues showed k-anonymity to be vulnerable to some attacks when using background knowledge² and proposed a new privacy criteria, named l-diversity. In 2007, Li and colleagues published a novel privacy criteria named t-Closeness, with even stronger properties for privacy preservation³. In addition to being computationally expensive⁴ and requiring prior knowledge to be able to anonymize the data, these classical tools are shown to deteriorate the data distribution, making them no longer exploitable in many situations.

Together with traditional anonymization techniques, recent technological advances in artificial intelligence, notably in generative modeling, led to the development of synthetic tabular data generation (SDG) algorithms⁴. SDG is performed by training a machine learning model on a real data set to generate synthetic data that mimic the original dataset. This process is done by learning the underlying data distribution and using it to generate synthetic samples. A vast number of methods have been developed over the last decade for synthetic tabular data generation⁴. Among the algorithms that can deal with mixed type data, combining categorical, discrete and continuous variables, are Classification and Regression Trees (CARTs), Bayesian Networks⁵ and Variational Autoencoders (VAE).

In addition, recent advancements in Deep Learning have led to the emergence of a new category of methods known as Generative AI. This field encompasses various

techniques, including Generative Adversarial Networks (GANs) and Diffusion-based models. Initially developed for generating high-quality images, significant progress has been made to apply these methods to tabular data, particularly Electronic Health Records (EHRs), which offer a vast amount of data for training such models.

Notably, there are Wasserstein-GAN models, such as EMR-WGAN⁶ and medWGAN⁷ (derived from medGAN⁸ model), that are specifically designed for generating high-quality samples from electronic medical records. More recently, diffusion-based models like EHRDiff⁹ and MedDiff¹⁰ have also been adapted to generate synthetic health data, similar to the work done with GAN models.

A recent review⁴ reported multiple methods for tabular data generation with a focus on the healthcare application context showing that there is no universal method or metric to evaluate and benchmark the performance of various approaches, both in terms of quality and privacy preservation. Several algorithms^{11,12} have been reported to generate high fidelity data while showing an optimal level of privacy but looking at data quality primarily through the prism of similarity of univariate distributions between the synthetic and original data. However, the richness of information in a dataset stems from the intricate multivariate associations among variables. These associations also contribute to privacy concerns, as combinations of quantitative variables can potentially be exploited to re-identify individuals. Therefore, it is crucial to consider the trade-off between data quality and data privacy when evaluating synthetic data generation methods, particularly in the context of clinical trials for which the number of patients and controls are typically limited.

As synthetic data generation algorithms do not create anonymized versions of the original samples but rather generate entirely new data that capture the statistical properties of the original dataset, metrics suited to traditional anonymization, like k-anonymity, are therefore not applicable in this context. We thus chose to use other relevant metrics, adapted from the identifiability and the membership inference scores, to assess the effectiveness in preserving the privacy of the original data.

The evaluation of data quality in synthetic tabular datasets, highlighted in a recent survey¹³, emphasizes the significance of metrics such as inter-dimensional relationship similarity, latent distribution similarity, joint distribution similarity, and prediction similarity. Recognizing the importance of capturing these aspects comprehensively, our benchmarks incorporate different metrics across a range of variable scales. At univariate scale, we perform chi-square and Wilcoxon tests. We then consider metrics in the bivariate context, using correlation and mutual information, and metrics in the multivariate context, with Wasserstein distance, to thoroughly assess the quality of synthetic datasets.

The existence of a privacy-utility trade-off in synthetic data has already been recognized and investigated ¹⁴. We quantify this trade-off in our benchmarking by introducing a Quality-Privacy Score (QPS) which combines different quality and privacy metrics to evaluate synthetic health data.

This paper serves two purposes: First, we propose a novel synthetic tabular data generation algorithm specifically suited for health records. It is based on the results of the Multivariate Information-based Inductive Causation (MIIC) algorithm ^{15,16} that has been shown to reliably capture multivariate relations across features. Second, we use multiple quality and privacy metrics and introduce a trade-off measure between these two complementary metrics, with the aim of comparing state-of-the-art synthetic healthcare data generation algorithms.

METHODS

In this section we first describe our novel synthetic data generation method: the MIIC-SDG algorithm is able to generate synthetic data that accurately captures the underlying multivariate distribution of the original data, without duplicating the data. The method is based on the reconstruction of a bayesian network that preserves direct associations and causal relationships from the original dataset. In the second part we review existing algorithms for synthetic tabular data generation and compare their performance over two real life benchmark datasets. Finally, we discuss metrics for evaluating quality and privacy of synthetic data, individually and jointly.

The proposed method, MIIC-SDG, takes advantage of the MIIC algorithm¹⁵⁻¹⁸ (MIIC network reconstruction) which can reliably capture the set of direct associations in complex heterogeneous datasets such as healthcare medical records. MIIC-SDG expands on the MIIC algorithm by adding a new algorithm that transforms a graph into a directed acyclic graph (MIIC-to-DAG) as well as a synthesizer (MIIC synthesizer) that takes into account the joint multivariate distribution associated with the data to generate high-quality samples that mimic the original data.

MIIC-SDG algorithm is composed of three steps, as shown in Figure 1:

- A. MIIC: Inferring a graphical model associated with the original dataset using MIIC algorithm.
- B. MIIC-to-DAG: Creating a directed acyclic graph (DAG) using the previously inferred network .

- C. MIIC-synthetizer: Generating synthetic samples based on the DAG and the original data by using several approaches that depend on the nature of parents and children nodes in the graph.

The MIIC-SDG algorithm is available as an R package, allowing the user to create synthetic versions of the original data inside a secured environment, without requiring any personal data to be transmitted to a web-server.

MIIC network reconstruction. The MIIC (Multivariate Information-based Inductive Causation) is an algorithm that infers a graphical network to represent the direct and possibly causal associations between variables in a dataset ^{15,16}. The algorithm has the ability to estimate conditional mutual information, even when the dataset includes a mixture of categorical and continuous variables. MIIC does not have hyperparameters and is not sensible to the order of features in the input data. It can estimate the set of associations between variables in the presence of missing data without the need of an a priori data imputation technique. MIIC has proven to be robust to sampling noise and to reliably estimate (conditional) mutual information. These features have been demonstrated in multiple benchmarks^{16,19}.

The MIIC-generated graph is composed of both directed and undirected edges, and may contain directed cycles, a common characteristic of causal discovery constraint-based methods). The directed edges result from identifying v-structures¹⁵ ($A \rightarrow B \leftarrow C$), where two independent and thus unconnected variables are linked to a third one. Additionally, directed edges can result from the propagation of orientations from upstream v-structures. However, propagated orientations do not necessarily indicate causal associations.

MIIC is available through a web-server¹⁸ and an R package and has been recently applied to a breast cancer cohort of 1,200 patients treated at Institut Curie, Paris¹⁷, as well as a larger cohort of 400,000 breast cancer patients from the SEER database^{16,19}. MIIC provides a novel way to globally visualize, analyze, and understand the connections between well-known clinical features.

MIIC DAG generation. In this study, we expand MIIC's ability to learn unparameterized network structures by incorporating a framework capable of generating synthetic data from a MIIC reconstructed graph. This approach takes advantage of the Bayesian framework, where the starting point is a DAG that can be parametrized using the original data. In this scenario, prior to data generation, the initial graph has to be transformed into a DAG. To this end, we designed and implemented a new algorithm, MIIC-to-DAG, that orients all undirected edges and removes all directed cycles from the

original graph reconstructed by MIIC, while retaining most v-structures and propagated orientations from the original MIIC network, whenever possible. This limits the impact of MIIC-to-DAG on the causal relationships and/or associations found from the real data. In particular, the propagation of orientations used by MIIC ensures that all remaining undirected edges in the MIIC structure can be oriented in either direction without adding new v-structures in the MIIC-DAG, unless such orientations create cycles. Indeed, orientating these undirected edges does not change the actual associations between the implicated variables, despite changing their causal relations, due to global Markov equivalence between the corresponding graphs.

MIIC-to-DAG algorithm consists in two steps: first, we orient each undirected edge in the MIIC network so as to minimize the number of directed cycles and possibly avoid them. Then, we remove all directed cycles from the graph, if some are present. In order to guarantee the removal of all the cycles of the graph, MIIC-to-DAG iteratively considers the longest cycle in the graph (the one with the most edges) and flips the edge that minimizes the number of remaining cycles in the graph. Taking the longest cycle guarantees the removal of at least one cycle at each iteration and therefore convergence towards a DAG. The pseudocode of the MIIC-to-DAG procedure is presented in Supplementary Figure 9. Beyond the visualization of the MIIC reconstructed network, the MIIC web server also enables the visualization of the DAG generated by the MIIC-to-DAG algorithm, by loading the list of directed edges available in the output of the execution of the MIIC-SDG algorithm. This step allows the user to inspect the DAG and to check the associations used to generate the data for each single variable. Details are present in the manual of MIIC-SDG (code availability section). The visualization of the causal network is possible at the following address: https://miic.curie.fr/vis_NL.php.

MIIC synthesizer. In this extension of the MIIC algorithm, the data generation component leverages the DAG structure obtained in the second step. The MIIC-SDG synthesizer follows the same concept of other algorithms based on Bayesian assumptions, where the sampling is initialized from variables associated with isolated or orphan nodes and then iteratively expands to nodes whose parents have already been generated. To address the presence of mixed-type variables in the original data, specific modifications have been implemented for the data generation algorithm.

There are different scenarios depending on the nature of the parent(s) (P) variable(s) and the target (T) variable:

1. *T discrete - P discrete* : this corresponds to classical Bayesian network algorithms where a multivariate conditional probability table for the target variable (child node) is estimated from the original data and then used to sample synthetic

data based on parents values (branch *a* of Figure 1d). When the target variable is a root node or an isolated node, the sampling is done directly from the target original probability table (branch *b* of Figure 1d).

2. *T discrete - P continuous/mixed*: we implemented this scenario with two different methods (branch *c* of Figure 1d). On the one hand, if the number of continuous parents is low (less than 3), continuous distributions are discretized using an optimum discretization algorithm from the MIIC framework¹⁶ that maximizes the mutual information between each parent and the target variable. This approach has shown to reliably estimate theoretical mutual information and to be adaptive to the number of samples and multimodal continuous distributions. We then applied the same method described in the first scenario to sample the target distribution. This method circumvents the difficulty of directly predicting a strongly unbalanced discrete variable with a classification model when the number of continuous predictors is low.

On the other hand, when the number of continuous parents is higher (3 or more), we choose to resort to a random forest classification model to predict the target variable, as it is able to capture non-linear associations between multiple predictors without having to discretize continuous parents.

3. *T continuous - P discrete*: in this scenario we estimate the density of the continuous target for each combination of discrete parents and sample from the empirical estimated density (branch *d* of Figure 1d). In case the target variable is a root node or an isolated node, the sampling is done directly on the estimated density of the original variable (branch *e* of Figure 1d).
4. *T continuous - P continuous/mixed* : this case is addressed with two methods (branch *f* of Figure 1d): first, if the number of continuous parents is low (less than 2), we use the optimum discretization algorithm from the MIIC framework and then learn and reproduce the density of the continuous target for each combination of discrete or discretized parents. If the number of continuous parents is higher (2 or more), we implement a random forest regression model to predict the continuous target node.

It is important to notice that missing values are represented as an extra category for discrete and categorical variables and hence reproduced in the synthetic generated data. For continuous variables, where densities estimation or classification/regression are applied, no missing value is generated for the synthetic feature.

Benchmark algorithms

MIIC-SDG. As detailed above, MIIC-SDG is composed of three steps: (A) discovers a network structure from the input data, (B) transforms this network into a DAG using MIIC-to-DAG algorithm and (C) uses this DAG and the original data to generate synthetic samples resembling the original data. The whole algorithm is detailed in the 'Methods' section I.

Bayesian. This method builds a probabilistic graphical model (Bayesian network) that represents the joint multivariate distribution by exploiting dependencies between random variables²⁰. In this framework, a DAG and a corresponding conditional probability distribution are learned from the given data. Sampling from the model is finally performed to generate a synthetic dataset. We used the code provided in the Synthcity²¹ package that uses the pgmpy package by Ankur and Abinash²⁰. The DAG is obtained using the tree search (Chow–Liu tree) or hill climbing algorithms.

Synthpop. The Synthpop algorithm, developed in 2016 by Nowok and colleagues, is a machine learning solution designed to generate synthetic test data for users who work with confidential datasets²². The synthetic data, generated through parametric and nonparametric methods, including the classification and regression trees (CART) model, aims to mimic the original data and can be used for exploratory analyses and for testing models. However, the CART model may result in final leaves with a small number of individuals, potentially compromising the privacy of the synthesized data. The authors suggest limiting this effect by specifying a minimum size for the final node produced by the CART model, though determining the appropriate value for this parameter is challenging as it is data-dependent and the method does not offer a tuning procedure.

CTGAN. CTGAN (Conditional Tabular Generative Adversarial Networks) is a deep learning algorithm published by Xu and colleagues in NeurIPS 2019, that aims at creating a generative model suitable for tabular data. CTGAN differs from traditional GANs by adding a conditional structure to both the generator and the discriminator networks, allowing it to generate synthetic samples based on specific real-world conditions. Authors have reported CTGAN outperforming Bayesian methods on most of the real datasets they presented²³.

TVAE. Tabular Variational AutoEncoders are adapted from classical variational autoencoders (VAE)²³ to enable the generation of mixed-type tabular data. This method was also used as a benchmark in the CTGAN paper. Authors claim that CTGAN achieves competitive performance across many datasets and outperforms TVAE on some benchmarks.

PrivBayes. PrivBayes is a differentially private Bayesian network model capable of efficiently handling datasets with a large number of attributes²⁴. Authors present the package as a new implementation that requires the injection of less noise compared to other differential privacy algorithms, maintaining more signal in the synthetic data. To obtain differentially private synthetic data, PrivBayes starts by creating a Bayesian network that succinctly represents the correlations among the attributes and then injects noise into each marginal distribution to ensure differential privacy. The method finally uses these noisy marginals and the Bayesian network to generate synthetic samples. The most important parameter for the algorithm is epsilon, determining the amount of noise injected in the marginal distributions. However, the choice of epsilon is not straightforward since the level of both quality and privacy of the generated data depends on the type of distributions, number of samples and complexity of the Bayesian network. For this benchmark, we choose an epsilon equal to 1 as it showed to be the best compromise in our simulations.

MedWGAN. medWGAN⁷ is a modified version of the medGAN model⁸. The medGAN model is a generative adversarial network for generating multi-label discrete patient records. It can generate both binary and count variables (i.e. medical codes such as diagnosis codes, medication codes or procedure codes). The modified medWGAN version introduced a Wasserstein GAN with gradient penalty as alternatives to the GAN in the medGAN framework. Among other benefits, the Wasserstein GAN offers an improved training stability by addressing mode collapse and enhanced gradient flow through techniques that stabilize and maintain a consistent gradient throughout the network. These advantages make WGAN a valuable approach in generative modeling tasks, enabling the generation of diverse and high-quality samples.

Random in range. This naive approach is used as a lower bound for normalizing the other benchmark methods. The “synthetic dataset” is obtained by generating random data using uniform distributions (inside the ranges of the original data). For categorical data it corresponds to a random sampling with replacement from all possible levels of categories of each feature. For continuous variables the sampling is made using the minimum and maximum values as ranges and sampling within the range with a uniform distribution.

In choosing these benchmark algorithms, we intentionally selected ones that employ different techniques for generating synthetic tabular data. A critical common feature among these chosen methods is their use of sampling from an estimated true joint distribution for data anonymization. This shared approach provided a foundation for comparing the performance of our method with these established benchmarks.

Algorithms parameters

To compare to other approaches, we used the default parameters further detailed in Supplementary Table 1.

During the initial step of MIIC-SDG, we employed specific parameters of the MIIC algorithm to facilitate the generation of the most oriented network. This involved enabling both orientation and propagation. We deliberately avoided the exploration of latent variables, deeming them less crucial for generating new data from the resultant network. Additionally, we opted not to implement any filtering on edge confidence, aiming to capture the most information on associated variables. Finally, KL-distance was enabled during the search for confounders in the presence of missing data, a crucial consideration when dealing with a substantial amount of missing data.

Assessing quality with distance metrics and predictive performances

In order to compare the methods, we have defined and computed several metrics for different settings:

Univariate analysis: We assess whether the distribution of each feature follows the same distribution in the original and synthetic datasets. To do so, we used a chi-squared test to compare categorical variables and a Wilcoxon test to compare continuous variables, with a significance level set to 0.05.

We use the chi-square test to determine if there is a significant difference between observed and expected frequencies across different categories. This test is commonly used to check for independence or association between categorical variables, assuming the variables are independent and the expected frequencies are not too small. In our case, we aim to have similar frequencies for each category in both the synthetic and original datasets.

The Wilcoxon test, a non-parametric test, is used to compare continuous distributions, particularly when the data does not follow a normal distribution. It does not assume normality and can handle skewed or heavy-tailed distributions. In our benchmark, we use the Wilcoxon test to verify that the distributions of the synthetic continuous variables are not statistically different from those of the original variables.

Correlations: We assess whether the bivariate distributions between each pair of features are preserved in the synthetic data. We compared the correlation matrices in the original and synthetic data by computing the mean absolute difference between the

matrices. The analysis is performed on all variable pairs by calculating their correlation using two approaches. The lower triangular matrix was determined by computing Pearson's correlation coefficient between continuous variables and Cramer's V between categorical variables. The upper triangular matrix was dedicated to analyzing the relationship between continuous and discrete variables. To this end, we used the MIIC algorithm which has been shown to optimally discretize the continuous features by maximizing the mutual information for all potential cut-points on the continuous variables. If a discretization was found (there is a significant correlation between the features), Cramer's V was then evaluated between the discrete and discretized variables. Also in this case the quality is directly associated with the difference between the correlation matrices. Small differences correspond to data that reliably capture the structure present in the original data. We define the correlation distance as the mean absolute difference between Correlation matrices C^S (synthetic) and C^D (original) as equation (1):

$$C_d(C^S; C^D) = \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p \left| C_{ij}^S - C_{ij}^D \right| \quad (1)$$

with p the total number of variables .

Mutual information (MI): In order to compare bivariate associations we used mutual information²⁵, which is a measure of dependence between two variables. The concept of mutual information is linked to the entropy of random variables, rooted in information theory. MI has been shown to robustly capture the association between variables even when their relationship is nonlinear. Just like correlation matrices are obtained by computing the correlation between variables, MI matrices are obtained by computing the MI between all variables in the dataset. We estimated the MI for discrete-continuous or continuous-continuous variables through the optimum discretization algorithm implemented in the MIIC package. The quality of the generated data is directly derived by computing the mean absolute difference between the MI matrices of the original data and the one of the generated data. Small differences correspond to data that reliably capture the underlying structure present in the original data. As with correlation matrices, we define the mutual information distance as the mean absolute difference between Mutual information matrices MI^S (synthetic) and MI^D (original) as equation (2):

$$MI_d(MI^S; MI^D) = \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p \left| MI_{ij}^S - MI_{ij}^D \right| \quad (2)$$

with p the total number of variables .

Multivariate distributions: To assess whether the joint multivariate distribution is preserved in the synthetic dataset we computed the Wasserstein distance (earth mover's distance) between the original and the synthetic data, using the synthcity²¹

package. The main advantage of using Wasserstein compared to other metrics, such as the Kullback-Leibler divergence for instance, is that it is a proper distance with the associated properties such as symmetry and does not require both measures (original and synthetic) to be on the same probability space. A small Wasserstein distance corresponds to synthetic data that reliably represent the multivariate distribution.

Predictive performances: One way to evaluate the quality of a dataset is to assess if the generated data can be used to perform classical machine learning tasks such as supervised learning. We therefore chose to compare the algorithms based on their capability to build a relevant machine learning model to predict overall survival using a survival random forest model. We also evaluated whether each synthetic dataset retains robust relationships by comparing variable permutation importance ranking with the “true” ranking obtained on the original dataset. We made this comparison using survival Random Forest as the machine learning algorithm. To achieve this, a K-fold cross-validation approach (K=5) was employed, where for each fold, the model was trained on 75% of the synthetic data and evaluated on a 25% hold-on test from the real data (the samples in the hold-out test set were selected to ensure that they were not used for training the generative model).

Assessing privacy metrics by identifiability score and membership inference

In this study we implemented different approaches to identify the level of privacy for each synthetic data generation method. In 2006 the paradigm of differential privacy was introduced and is still to date one of the most used techniques to try to preserve data privacy through mathematical constraints²⁶. However, differential privacy has been shown to not fully mitigate the risk of re-identification. Stadler and colleagues have shown that under certain circumstances, neither the original implementation of PrivBayes nor PATEGAN²⁷ (Private Aggregation of Teacher Ensembles Generative Adversarial Network) reliably prevents linkage attacks, leaving some samples vulnerable to membership inference attacks²⁸. Moreover, it has been reported in many scenarios that strong differential privacy constraints lead to the generation of synthetic data exhibiting a disrupted correlation structure between features, making the resulting data problematic²⁴ (Figure 3).

Revisiting Identifiability score (IS)

To ensure data privacy, generated synthetic patient records should be “different enough” from the original patient records. Following this idea, we used a framework for the evaluation of privacy risks developed by Yoon and colleagues who proposed a new concept for identifiability²⁹. Yoon and colleagues define an identifiability property related to the minimum distance between real patients and the distance between real and synthetic samples. In order to weight each feature according to its probability of

identifying patients having the same values, authors used a weighted Euclidean distance as metric, giving more weight to features having an unbalanced distribution of events. Authors define ϵ -identifiability as the property of having less than ϵ ratio observations from the original dataset in the generated synthetic dataset that are “not different enough” from the original observations. ϵ corresponds to the defined identifiability score. In this scenario, an identifiability of zero would represent a perfectly non-identifiable (private) dataset and an identifiability of one would represent a perfectly identifiable dataset. The proposed identifiability is defined for all the samples or variables. The described identifiability distance is implemented in the Synthcity package. The derived privacy score is then defined as $1 - \text{identifiability score}$.

Adapting membership inference score (MIS)

Secondly, inspired by the work of El Emam K et al.³⁰ and J. Yoon et al.²⁹, we also proposed to compute a membership inference metric. We used the partitioning membership disclosure attack method proposed by El Emam K and colleagues where, instead of using the hamming distance between samples as a similarity measure, we used a weighted Euclidean distance where the weights are defined as the entropy of each feature, as proposed by J. Yoon et al.²⁹

Moreover, to ensure a reliable distance metric with mixed variables (a combination of variables with either categorical or continuous domains), we have taken the following additional steps:

- Apply a Multiple Factor Analysis for Mixed Data (using the FactoMineR package³¹) to reduce the dimensionality of the data.
- Compute the Euclidean distance between samples in the space of the 10 first principal components as a similarity measure. It has the benefit of keeping the notion of Euclidean distance valid and meaningful.

This score evaluates whether we are able to identify which patients were used to create the synthetic dataset by subsampling the original dataset into a training and test set, for varying sample size. The derived privacy metric is then defined as $1 - \text{membership inference score}$.

Trade off between Quality and Privacy: the Quality-Privacy Score (QPS)

Literature results³² and our findings manifestly suggest the necessity of a trade-off between data quality and data privacy. On the one hand, small modifications of the original data are directly associated to good quality scores but to poor privacy ratings, since almost all the information of the dataset is maintained. On the other hand, strong perturbations or noise addition lead to a net loss on quality of the synthetic data with a concomitant gain on privacy.

This problem is analogous to the classical machine learning dilemma of obtaining good precision and recall scores, simultaneously, which calls for defining a trade-off measure such as the F_1 score measure, defined as the harmonic mean between precision and recall scores as equation (3)

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Inspired by the F_1 score definition, we formulated a version of quality *versus* privacy trade-off by adapting the classical F_1 formula above. This requires defining data quality and privacy scores in the range $[0, 1]$, where best score values correspond to 1.

To this end, we normalized each quality and privacy metric by a reference value, namely those computed on the data generated by the Random method.

We define as normalized quality each of the previously described quality measures divided by the corresponding reference value as equation (4):

$$Q(MI^S) = 1 - \frac{MI_d(MI^S; MI^D)}{MI_d(MI^R; MI^D)}, \quad (4)$$

where $Q(MI^S)$ is the normalized quality for mutual information of synthetic dataset, MI^S , MI^D , and MI^R are the mutual information matrices of, respectively, the synthetic data, the original data, and the random data, and $MI_d(X; Y)$ is the mutual information distance between mutual information matrices X and Y , as defined in section 3.

We define similarly the normalized quality for Wasserstein distance as equation (5):

$$Q(W_p^S) = 1 - \frac{W_p(x^S; x^D)}{W_p(x^R; x^D)}, \quad (5)$$

where $Q(W_p^S)$ is the normalized quality for Wasserstein distance of the synthetic dataset, x^S , x^D , x^R are the distributions of the synthetic, original and random datasets respectively and $W_p(X; Y)$ is the Wasserstein p -distance between distributions X and Y .

Finally we define the normalized quality for correlation distance as equation (6):

$$Q(C_d^S) = 1 - \frac{C_d(C^S; C^D)}{C_d(C^R; C^D)}, \quad (6)$$

where $Q(C_d^S)$ is the normalized quality for correlation distance of the synthetic dataset, C^S, C^D, C^R are the correlation matrices of the synthetic, original and random datasets and $C_d(X; Y)$ is the correlation distance between correlation matrices X and Y , as defined in section 3.

In the same way, normalized privacy is defined as equation (7):

$$NP_{IS} = 1 - \frac{IS_S}{IS_R}, \quad (7)$$

with $IS_S \in [0, 1]$ and $IS_R \in [0, 1]$ the identifiability scores of the synthetic and random datasets respectively.

Similarly we define the normalized membership inference score, as equation (8):

$$NP_{MIS} = 1 - \frac{MIS_S}{MIS_R}, \quad (8)$$

with $MIS_S \in [0, 1]$ and $MIS_R \in [0, 1]$ the membership inference scores of the synthetic and random datasets respectively.

In this setting, we define our Quality-Privacy scores as the harmonic mean of the normalized quality and privacy scores, as equation (9):

$$QPS = 2 \times \frac{NQ * NP}{NQ + NP}, \quad (9)$$

with $NQ \in \{Q_{MI_D}, Q_W, Q_{C_d}\}$ and $NP \in \{NP_{MIS}, NP_{IS}\}$.

In addition, to facilitate benchmark comparison between different synthetic data generation methods, we have also designed two global scores to highlight the performance across the 3 quality metrics (Wasserstein, Mutual Information, correlation) and the 2 privacy metrics (Membership inference and identifiability) used in this study.

We will refer to these global scores as $MetaQPS_{am}$ and $MetaQPS_{hm}$ in the rest of the paper.

$MetaQPS_{am}$ corresponds to the F1-score between the arithmetic mean of the quality scores and the arithmetic mean of the privacy scores, defined as:

$$metaQPS_{am} = 2 \left[\frac{3}{(Q_{MI_D} + Q_W + Q_{C_d})} + \frac{2}{(NP_{MIS} + NP_{IS})} \right]^{-1} \quad (10)$$

$MetaQPS_{hm}$ corresponds to the F1-score between the harmonic mean of the quality scores and the harmonic mean of the privacy scores, defined as:

$$metaQPS_{hm} = 2 \left[\frac{(Q_{MI_D}^{-1} + Q_W^{-1} + Q_{C_d}^{-1})}{3} + \frac{(NP_{MIS}^{-1} + NP_{IS}^{-1})}{2} \right]^{-1} \quad (11)$$

$MetaQPS_{am}$ and $MetaQPS_{hm}$ could be readily adapted to include other quality and privacy scores, if needed. Hence, the rationale of the $MetaQPS$ metrics is to first take the arithmetic or harmonic means over all quality and privacy scores, separately, before quantifying the global quality versus privacy trade-off using an F1-score measure. While the $MetaQPS_{am}$ metric, based on arithmetic means tends to equally weigh the contributions of each quality and privacy scores, the $MetaQPS_{hm}$ metric, based on harmonic means, gives more weights to lower individual quality and privacy scores and is therefore expected to be more sensitive to the most discriminative quality and privacy scores.

Benchmark datasets

This paper aims to provide a comprehensive benchmark specifically focused on the anonymization of small, complex datasets with high privacy risks. To ensure the relevance and robustness of the comparison, we intentionally included a diverse range of methods including Bayesian and constraint-based approaches as well as Deep learning methods. However, it is important to note that Deep Learning methods like CTGAN, medWGAN, and TVAE require larger sample sizes for effective convergence. To address this specificity, we chose to incorporate three datasets of various sizes to assess the performance of each method on a wide range of sample sizes (from 100 to 20,000 patients).

Breast cancer (METABRIC)

The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset is a collection of over 2,000 clinically annotated primary breast cancer specimens obtained from tumor banks in the UK and Canada³³. The cohort encompasses clinical variables and genetic information including copy number alterations, copy number variations, and single nucleotide polymorphisms. The METABRIC dataset was selected for this study due to its widespread usage and validation in the literature, as well as its suitable sample size for the application of machine learning algorithms and coexistence of numerical and categorical features. The original dataset, consisting of 2491 patients and 36 variables, was pre-processed by removing patients with more than 20% missing values and variables with unique values. The resulting filtered dataset comprised 1977 patients and 29 clinical variables, with 19 of them being discrete and 10 continuous. Figure 2 reports the DAG obtained by using the MIIC-to-DAG algorithm (step 2B in Figure 1), which was applied on the network reconstructed by the MIIC algorithm, shown in Supplementary Figure 2. The graph contains 63 edges reporting direct association between the 29 variables. Unconnected nodes represent features that are not associated with any other variable in the data (following MIIC residual mutual information evaluation). It is important to remember that the MIIC algorithm does not have hyperparameters, does not need any tuning, can deal with missing data and is not sensible to the order of features in the data. The used data comes from public data available in the cBioPortal repository:

https://www.cbioportal.org/study/summary?id=brca_metabric.

Details about the METABRIC patient population, consents, approvals, tissue collection and sample processing are provided in the publication by Curtis et al.³³ All patient specimens were obtained with appropriate consent from the relevant institutional review board. The data are available at the European Genome-phenome Archive (<http://ebi.ac.uk/ega/>), which is hosted by the European Bioinformatics Institute, under accession numbers EGAS00000000083.

Bladder cancer

To extend our assessment of synthetic data generation approaches, we integrated a second dataset derived from a Roche clinical trial on bladder cancer (Phase II single-arm study IMvigor210³⁴). The study investigated metastatic urothelial cancer treatment with the anti PD-L1 agent atezolizumab and highlighted key factors influencing outcomes, such as the presence of a CD8+ T-effector cell phenotype and the burden of neoantigens and tumor mutations. These known key outcomes were then also used as a quality metric in our assessment.

IMvigor210 was a multicentre, single-arm, phase 2 trial that investigated efficacy and safety of atezolizumab in metastatic urothelial cancer. This trial was done in 47

academic medical centers and community oncology practices across seven countries in North America and Europe. All patients provided written informed consent before study entry. The study was done in accordance with the Declaration of Helsinki and International Conference of Harmonization Good Clinical Practice guidelines. The trial is registered with ClinicalTrials.gov under the following number: NCT02951767 for cohort n°1 and NCT02108652 for cohort n°2. The protocol was approved by institutional review boards or independent ethics committees at participating study sites. The data has been made available through an R package on the following website: <http://research-pub.gene.com/IMvigor210CoreBiologies/>

This addition enriches our benchmark with a second real and practical use case coming from our organization. It also helps validate the overall applicability of our approach to different sets of data, varying both in dimensionality and in the types of variables analyzed. Specifically, IMvigor210 included 310 participants with 297 known outcomes. We further restricted the benchmark to 24 features keeping a balanced mix of continuous and discrete variables.

Diabetes

Diabetes is among the most prevalent chronic diseases, impacting around 500 million people worldwide. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy.

To test our algorithm on larger sample sizes, we added a third dataset related to Diabetes. This dataset originates from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset, sourced from telephone surveys, encompassing USA residents' health behaviors, conditions, and socio economic aspects³⁵. As stated by the Behavioral Risk Factor Surveillance System, data and materials produced by federal agencies are in the public domain and may be reproduced without permission.

It is often used for machine learning purposes and is also publicly available on Kaggle, the version we used is available at this address: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>.

It contains factors influencing diabetes disease and other chronic health conditions related to diabetes. The full dataset contains 253680 samples and 22 features.

Benchmark setting

We evaluated the different algorithms by subsampling the whole 1977 sample METABRIC dataset in different subsample sizes: 50, 100, 200, 500, 1000, 1500 and

1977. This allows one to assess the performance of each method in multiple subsets along with their stability. We created 10 datasets for each sample size and each one of them was used to generate 10 synthetic datasets with a different seed (100 datasets for each sample size). It is important to note that seed effects are different in each algorithm. For the IMvigor210 study we chose to analyze the dataset in subsets of 100, 200 and 297 samples. For Diabetes, we analyzed the algorithm performances on larger sample sizes: 500, 1000, 5000, 10000 and 20000.

RESULTS

Univariate distribution comparison as a per variable analysis.

When comparing two datasets that share the same set of features, the simplest analysis we can conduct involves assessing the distribution of each variable within both the original and synthetic datasets. Table 1 presents the average count of features that exhibited statistically significant differences based on these two tests across various sample sizes (columns) and algorithms (rows). The standard deviation is provided in parentheses. The results indicate that Synthpop is the most effective method for replicating univariate distributions, followed closely by Bayesian algorithms and MIIC-SDG, which demonstrated similar performance. Conversely, the other algorithms fell short in reproducing the univariate distribution, with between 16 (CTGAN and TVAE) and 21 features exhibiting differences in the largest sample size. As expected, the random method flagged nearly all variables as different, owing to its random sampling approach within the original feature range.

Bivariate analysis to assess associations between variables

Mutual Information (MI) distance: Mutual information distances are evaluated by calculating the mutual information difference between real and synthetic data, evaluating it for all pairs of variables. The results are presented in Supplementary Figure 2. The data generated by the MIIC-SDG algorithm reproduces well the mutual information between variables present in the original data, obtaining better scores than other methods for small sample sizes (50 and 100 samples). With an increasing number of samples the mutual information is best reproduced by Synthpop, with MIIC-SDG positioning at second or third place (close to bayesian tree search). MedWGAN scores at the fourth place. CTGAN obtained the lowest score on the smallest sample size, and eventually improved its performance with larger sample sizes. TVAE and the Bayesian hill climbing technique scored closely on the largest datasets, with Bayesian hill climbing reaching the fourth or fifth position in smaller samples (< 500 samples).

Correlations (Corr) distance: Supplementary Figure 3 reports the correlation distances, defined in section 3, between synthetic data and the corresponding original data. It can be observed that the Bayesian tree search and Synthpop algorithms have comparable performances, both for low and high sample sizes. MIIC-SDG has better results than competitors on small sample sizes (< 200 samples) but the gain in performance obtained by increasing sample sizes stabilizes at higher values compared to the first two competitors. CTGAN obtained low scores for smaller sample sizes (< 500 samples), but the performance increases fast with increasing the number of samples, obtaining similar scores to MedWGAN. TVAE fails to capture the correlation structure, even if it is improved with 1500 samples. Bayesian hill climbing with BIC criterion does not obtain competitive scores, even with large sample sizes. PrivBayes exhibit similar results to the random data method, showing a complete loss of correlation patterns. Globally, MIIC-SDG, CTGAN and MedWGAN obtained similar scores at large sample sizes.

The correlation matrices for the METABRIC datasets (using 1000 samples) are shown in Figure 3. Values are obtained as a mean correlation over all executions from running the algorithms on the 1000 sample datasets (using bootstrap) and using multiple seeds.

The performance of the different methods are, in order: Bayesian with tree search algorithm, Synthpop, CTGAN, MIIC-SDG and medWGAN with the same score, TVAE and PrivBayes. As expected, the score of the random method is by far the lowest one and it is strongly dependent on the type of associations between variables that exist in the original data (correlation structure and strength). As shown in Supplementary Figure 12, we compare the correlation matrices for the Diabetes dataset in a scenario where the correlation structure is much sparser, with only a few features being correlated. In this context, using the mean correlation distance to differentiate between algorithms is less meaningful. Random values can exhibit a low mean correlation distance because most features are uncorrelated, and introducing random values does not significantly impact the overall correlation. Therefore, in such cases, distinguishing between algorithms in terms of correlation distances becomes more challenging.

Multivariate distribution comparisons as a global metric to evaluate quality

Wasserstein distance: The results of the multivariate Wasserstein distance assessments are presented in Supplementary Figure 4. In this multivariate setting the Bayesian network approach with tree search estimation reached the best scores with very small distances, followed by Synthpop and medWGAN at similar scores, MIIC-SDG at the fourth place, and Bayesian hill climbing (using BIC criterion), with

CTGAN and TVAE in similar ranges. The differential privacy bayesian approach with epsilon set to 1 generates datasets with much higher distances compared to the other benchmarked methods.

From these results, we can notice that there are large gaps between the different methods, with the Bayesian (tree search estimation) approach giving much better results than competitors. In Supplementary Figure 4, we can observe that the Bayesian tree search approach generates synthetic categorical variables on the METABRIC dataset that closely resemble the original distribution. However, the performance difference is less pronounced for continuous variables. The Bayesian approach demonstrates overall good performance, which can be attributed to the METABRIC dataset being predominantly composed of categorical variables. This gives an advantage to methods that excel in handling such variable types.

When computing the Wasserstein distance using only continuous features, we observe that Bayesian tree search obtains indeed similar scores with respect to Synthpop and MIIC-SDG. Similarly, MedWGAN obtains good scores mostly on categorical variables, but performs slightly less well on continuous variables, which might be due to the low sample size used to train the model.

Machine learning performance: Predicting Overall Survival (OS) response

The aim of this part is to evaluate the ability of synthetic data generation algorithms to preserve multivariate information for the purpose of predicting survival features. Figure 4 represents the feature importance for the prediction of Overall Survival (SO) in the original data and in all benchmark algorithms. We note that MIIC-SDG, Synthpop, CTGAN and medWGAN all predict the same three main features (Nottingham Prognostic Index, Number of positive lymphatic nodes and Tumor size) as the most important features for OS prediction. Figure 5 shows the concordance index estimates from Survival Random Forest model to predict Overall Survival. While MIIC-SDG does not outperform a number of benchmark methods in terms of OS prediction, it is important to keep in mind that the ability to predict a target variable (OS in this case) from other features can also be used as a metric for privacy risk, as it can be seen as an inference attack on sensitive attributes. Having a high concordance with the true data also correlates to a high risk in the case of inference attacks.

Re-identification as a metric to evaluate privacy level.

Identifiability score: The identifiability score corresponds to the probability of re-identification given the combination of all data on any individual patient. It is evaluated by measuring the identifiability of the finite original patient data using the finite generated synthetic data. Supplementary Figure 5 shows the privacy score of the synthetic data generation algorithms evaluated as $(1 - \text{Identifiability})$, as a high identifiability score indicates low privacy. The Bayesian algorithm with tree search has the lowest privacy scores (0.23 - 0.44), followed by the Synthpop algorithm (0.44 - 0.49), MIIC-SDG (0.45 - 0.6), TVAE (0.47 - 0.76), CTGAN (0.59 - 0.81), Bayesian with hill climbing (0.71 - 0.87), PrivBayes (0.94 - 0.99) and Random (0.97 - 1), with numbers between parenthesis corresponding to the smallest and biggest sample sizes. Interestingly, the random algorithm did not reach 1 for the smallest sample sizes.

Membership inference score: The membership inference score corresponds to the probability of identifying which patients have been used to generate the synthetic dataset. Supplementary Figure 6 shows the privacy score of the synthetic data generation algorithms evaluated as $(1 - \text{membership score})$, as a high membership score indicates low privacy. Bayesian tree search algorithm is the method with lowest privacy as it is easy to guess if a sample has been used or not to generate the synthetic data, followed by the Synthpop method, where privacy scores never increase above 0.5. MIIC-SDG remains at the third position, with privacy scores increasing well with larger sample sizes. CTGAN obtains slightly better results in the membership inference attack, together with TVAE. Bayesian hill climbing generates datasets where it is hard to guess the membership of samples in the original data, while PrivBayes and the random algorithm obtain best privacy scores, with values close to 1.

Quality-Privacy Scores (QPS) as a trade-off between quality and privacy

Quality-Privacy scores can be evaluated by using different metrics for both quality and privacy. Both dimensions have been evaluated by calculating the ratio between the value obtained using the data of each algorithm and the value obtained using the corresponding random data, so that both quality and privacy range in $[0,1]$ (normalized formula). For quality measures we focused on the normalized Mutual Information distance that reliably captures bivariate associations and clearly discriminates the different approaches. As privacy metrics we considered the results obtained with the identifiability and the membership inference scores, in their normalized version. QPS are then obtained by taking the harmonic mean of the normalized quality and privacy scores, for each combination of quality-privacy measures, as introduced in the Method section. In addition, in order to facilitate the comparison of synthetic data generation methods across multiple quality and privacy scores, we also introduced two global QPS

metrics, $\text{metaQPS}_{\text{am}}$ and $\text{metaQPS}_{\text{hm}}$, based, respectively, on the arithmetic means and harmonic means of several quality scores and privacy scores, as detailed in Methods.

In Figure 6 we show the mutual information quality metric and the two privacy scores as well as the two derived QPS (one for each privacy metric). Based on the QPS derived from mutual information distance, the MIIC-SDG method is ranked as the best algorithm in terms of the quality-privacy trade-off for both QPS metrics. It is only outperformed by Synthpop for large sample sizes when considering the Membership inference as the privacy score. However, it is important to notice that the privacy evaluated from the identifiability score is always smaller than 0.5 for Synthpop, a value much lower to the one obtained through MIIC-SDG, ranging in 0.6 - 0.7 for higher sample sizes (> 200 samples). This makes synthetic data generated with Synthpop significantly less private than synthetic data generated with MIIC-SDG. The QPS obtained using MI distances thus highlights MIIC-SDG ability to reliably generate quality synthetic data while best preserving the privacy of the original sensitive data.

We also analyzed the results obtained with the other quality distances, the two privacy measures and the corresponding QPS (4 supplementary metrics). These complete results are shown in Supplementary Figure 7. MIIC-SDG algorithm obtained the best QPS results for correlation distance for small sample sizes (<500 samples) and obtained the second or third best scores after the CTGAN and medWGAN for the largest sample sizes. Based on the QPS obtained using the Wasserstein distance and Identifiability scores, the Bayesian hill climbing algorithm emerges as the top-performing method, followed closely by CTGAN, MIIC-SDG, and TVAE, all exhibiting similar scores. In this scenario Synthpop does not show competitive results due to a poor privacy score. Bayesian tree search and PrivBayes score poorly as well, the first due to a low privacy score and the second due to a low quality of the generated data.

We then run the same pipeline on the second dataset (IMvigor210), presented in Figure 7, obtaining comparable results to the ones on the METABRIC data. Also in this case MIIC-SDG shows better results than competitors based on the mutual information QPS, reporting a good trade-off between quality and privacy scores. Supplementary Figure 10 provides detailed information on the QPS, quality, and privacy scores for the IMvigor210 dataset. In this scenario, MIIC-SDG scores rank either first or second in terms of QPS (and third when tied with Synthpop), using correlation as the distance metric. However, QPS ranks using Wasserstein distance do not effectively differentiate between the best methods, which all perform essentially the same, as observed on the METABRIC dataset.

Finally, we benchmarked the different synthetic data generation methods on a larger dataset, Diabetes (500 to 20,000 samples). The results reported in Supplementary Figure 11 show that MIIC-SDG shares, together with Synthpop, the top-ranking QPS score based on mutual information, with MedWGAN also reaching similar high scores at large sample sizes (>1000 samples). However, the three methods cannot be seen as equivalent in all contexts, as Synthpop exhibits a significantly higher quality with concomitant lower privacy, whereas medWGAN exhibits a higher privacy but lower quality overall. By contrast, MIIC-SDG appears to achieve a better balance between privacy and quality defined in terms of MI distances. Using other quality metrics, MedWGAN is shown to achieve the best QPS tradeoff based on correlation distance but the worst Wasserstein distance overall. Yet, as with the first two datasets, the Wasserstein distance did not provide a discriminative ranking of the best performing methods.

Figure 8 summarizes these results by integrating all quality and privacy scores into single metaQPS metrics for each dataset and all sample sizes analyzed in this study. This figure demonstrates the good overall performance of MIIC-SDG algorithm, particularly at small sample sizes, while other state-of-the-art methods, such as Synthpop, CTGAN and medWGAN, achieve better or similar performance at large sample sizes. Among the Bayesian methods included in this study, Bayesian hill climbing and Bayesian tree search exhibit somewhat lower performance, but significantly outperform privBayes. This suggests that the incorporation of differential privacy into privBayes adversely impacts the overall quality of the generated synthetic data. All in all, we note that the $\text{metaQPS}_{\text{hm}}$ metric, based on the harmonic means of the quality and privacy scores, provides a more stringent evaluation of synthetic data generation methods by highlighting the most discriminative quality and privacy scores integrated in the global $\text{metaQPS}_{\text{hm}}$ metric. By contrast, the arithmetic means of individual quality and privacy scores integrated in the $\text{metaQPS}_{\text{am}}$ metric tend to be less discriminative. We would therefore recommend using the $\text{metaQPS}_{\text{hm}}$ metric to integrate several quality and / or privacy scores in future comparative studies of synthetic data generation methods.

DISCUSSION

Over the past few decades, there has been a significant increase in the recruitment of patients for clinical trials and the collection of real-life health-related datasets. This surge has been witnessed in both public institutions and private companies, resulting in the accumulation of vast amounts of patient information. As the number of collected studies continues to grow, it becomes increasingly urgent to explore effective solutions for harnessing this wealth of data. This entails facilitating new research initiatives and

promoting data sharing, all with the ultimate goal of pushing the boundaries of medical research. In recent years, various machine learning and deep learning approaches have been employed to synthesize health data. These approaches hold the promise of enabling data sharing while safeguarding patient privacy. Regulatory standards such as the European General Directive on Data Protection (GDPR) mandate that data holders implement robust measures to ensure data security and prevent potential data breaches, often leading to restrictions on data sharing and secondary data usage. However, few established standards exist to guarantee adequate data anonymization and data security. Previously proposed methods like k-anonymity, l-diversity, and t-closeness have limitations when it comes to preserving privacy while maintaining sufficient data quality for research purposes. Therefore, the development of new quantitative standards is imperative to facilitate data anonymization through the generation of synthetic data and assess the level of risk associated with data publication. Synthetic data generation offers several opportunities that can be categorized as follows: firstly, as a tool for collaborative projects thanks to the straightforward and time-efficient sharing of data; secondly, as a pre-production platform for development; and finally, for deriving insights from data. The first two application types typically do not necessitate the faithful reproduction of complex statistical features found in the original datasets. However, the generation of insights relies on the synthetic data generator's ability to preserve intricate data patterns, including those not yet identified in the original dataset. Indeed, the estimation of the joint multivariate distribution of clinical trials is a hard task due to the limited sample size of such datasets. It is therefore crucial to identify synthetic data algorithms that are able to operate on such a scale and provide meaningful results.

In order to identify a suitable algorithm for synthetic data generation applicable to biomedical/clinical data, it is essential to consider the preservation of data quality and the protection of privacy simultaneously. To address this challenge, we conducted a comprehensive evaluation of the various state-of-the-art algorithms across multiple scenarios, examining data quality and privacy both separately and more importantly in combination. To evaluate the impact of sample size on each algorithm's performance, we used datasets of varying sizes: small (ranging from 50 to 200 samples), medium (500 to 2000 samples), and large (5000 to 20000 samples).

Our approach involved the following steps:

1. Defining Quality Metrics: we first established various metrics for assessing the preservation of data quality. These metrics were designed to gauge the ability of the methods to generate data that closely resembles the original dataset. One of the most used methods to compare data have been performed through the

Pearson correlation coefficients. However, Pearson correlation is also known to be very sensitive to outliers, which may explain some of the apparent good relative rankings of certain methods under correlation scores, while they exhibit poorer performance under more robust statistical criteria such as MI, which only depends on the ranks (not the specific values) of the variables of interest. We hence focused our analysis on MI, but we also presented results using the more classical correlation concept.

2. Privacy Considerations: we then focused on the privacy of the original sensitive data. Our aim was to prevent the re-identification of patients and safeguard against the disclosure of sensitive patient information.
3. Defining trade-off metrics: To provide a comprehensive assessment of the sensitive synthetic data generated, we proposed to combine the quality and privacy metrics. The resulting combination of metrics introduced in the paper defines novel pairwise and global quality-privacy scores, the QPS and metaQPS metrics, which we used to rank all the algorithms included in our benchmark.

Our novel SDG method (MIIC-SDG) provides good overall performance in terms of quality-privacy tradeoff, especially at small sample size, while other state-of-the-art methods, such as Synthpop, CTGAN, and medWGAN, achieve better or similar performance at large sample sizes, Figure 8. MIIC-SDG, which integrates graph learning and information theoretic approaches, retains complex association patterns, including those that remain undiscovered in the original dataset, making it suitable for extracting novel insights from sensitive biomedical data. Other methods approximating the full joint distribution, such as Synthpop, are also capable of preserving these complex association patterns, although they typically present a higher risk of patient re-identification. By contrast, methods that focus on preserving more limited aspects of the data, like first-order correlations, are not well-suited for novel insight generation, although they may suffice for pre-production algorithmic development purposes. Furthermore, it is important to keep in mind that SDG methods utilizing GAN or VAE approaches, such as CTGAN, medWGAN and TVAE, require large sample sizes to achieve stable results. This characteristic explains why they tend to achieve lower performance on small datasets, such as Metabric and IMvigor210. On the other hand, on larger datasets, such as Diabetes, GAN approaches achieve performance at par with the best SDG methods in terms of QPS ranking.

These results confirm that Deep Learning approaches are definitely relevant when the sample size is large enough to properly train them, as already evidenced by their success in generating synthetic Electronic Health Records (EHR) from datasets with

tens of thousands of records^{6,7,12}. However, in the specific context of small and complex biomedical datasets, typically found in clinical studies, the relevance and benefits of Deep Learning approaches is not as apparent, especially in terms of information preservation.

MIIC-SDG stands out particularly well for generating synthetic datasets that contain mixed data types and a low number of samples (< 1000), a characteristic typically observed in actual clinical datasets. The method is also well-suited for handling longitudinal data, where measurements are taken at various time points.

This paper also introduces novel pairwise and global quality-privacy scores (QPS and metaQPS), which aim to quantify the tradeoff between quality and privacy measures in the evaluation of synthetic health data generation methods. These scores are designed to be bounded between 0 and 1 such that a QPS = 0 represents a randomized dataset. However, while reaching the theoretical maximum value of 1 may be difficult or impossible in practice, these scores can still be used to compare and rank different SDG methods, similar to how relative F1-scores are used to compare and rank ML methods.

An important feature of pairwise QPS metrics is their focus on a specific pair of quality and privacy scores. Indeed, SDG methods need to be assessed in context and certain quality or privacy criteria might be preferable to others in different contexts. For instance, while Synthpop, MIIC-SDG and medWGAN shared top-ranking QPS scores based on mutual information on the Diabetes dataset (Supplementary Figure 11), the three methods cannot be seen as equivalent in all contexts. Indeed, Synthpop exhibits a significantly higher quality with concomitant lower privacy, whereas medWGAN exhibits a higher privacy but lower quality overall (Supplementary Figure 11). Hence, MIIC-SDG appears to achieve a better balance between privacy and quality defined in terms of MI distances, which likely stems from the information-based approach employed by MIIC.

By contrast, MedWGAN was found to achieve the best QPS tradeoff based on correlation distance but the worst Wasserstein distance overall (Supplementary Figure 11). As correlation measures are well known to be highly sensitive to outliers, this may explain the favorable results of GAN methods with correlation quality metrics in general. Instead, MI is only sensitive to the rank (not the specific values) of variables, making it a much more robust quality score to outliers in the data. Finally, quality metrics derived from Wasserstein distances were found to be less discriminative than MI distances for ranking the best performing methods, which all reach similar Wasserstein scores in general.

To facilitate the overall comparison of SDG methods, we also introduced global performance scores integrating multiple quality and privacy scores. These metaQPS metrics provide a global indicator of each algorithm's performance across all relevant quality and privacy scores and health datasets included in this study, Figure 8. The two metaQPS metrics come with advantages and disadvantages. The metaQPSam metric, based on the arithmetic means of individual quality and privacy scores, equally weighs the contributions of each quality and privacy metrics. This tends to overlook the differences between SDG methods by masking their limitations on the most stringent quality or privacy measures. By contrast, the metaQPShm metric, based on the harmonic means of the quality and privacy scores, gives more weight to the most stringent quality and privacy scores. Hence, the metaQPShm metric is more conservative and better suited to highlight the differences between alternative SDG methods, due to its sensitivity to the most discriminative quality and privacy scores. MetaQPShm should therefore be favored in future comparative studies of SDG methods integrating several quality and / or privacy scores.

This study comes with limitations. First, we performed benchmark comparisons across a wide range of SDG methods and could not therefore include all available approaches beyond a few representatives of each class of SDG methods with open access and readily usable codes. Second, these benchmark comparisons are limited to three datasets and might not generalize to all possible health data, although we have tried to cover a wide variety of health datasets, including a broad range of sample sizes and mixed-type (continuous and categorical) variables. Third, this study primarily focused on a quantitative evaluation of the quality versus privacy tradeoff of SDG methods, with the introduction of novel pairwise and global QPS metrics. Yet, we recognized that other indicators, such as the overall survival prediction, are also important to assess the relevance of synthetic health data generation methods. Finally, the membership inference score, which requires a hold-out set for computation, could not be estimated for the entire datasets. This limitation could be overcome by employing a k-fold cross-validation procedure, as implemented in the TensorFlow Privacy Membership Inference Attack (MIA) Python library³⁶.

Likewise, the novel MIIC-SDG method reported in this paper has notable limitations, despite its good performance in terms of global quality-privacy scores, Figure 8, especially at small sample sizes. First, MIIC-SDG does not rank among the best methods in predicting overall survival (OS) from a Survival Random Forest model, which suggests potential limitations in predicting specific tasks. Second, while MIIC-SDG achieves very good performance in terms of quality-privacy tradeoff based on mutual information (MI) distances, it does not display the best performance across all quality metrics. For instance, other methods such as medWGAN excel when correlation

scores are used. Similarly, MIIC-SDG's well-balanced performance in terms of quality-privacy trade-off implies that MIIC-SDG offers mid-range privacy scores and might not be the method of choice for applications requiring the highest level of data protection.

Considering these limitations and to better reflect the importance of quality versus privacy criteria in different contexts, a possible extension of the F1-based QPS scores, introduced here, would be to define F_{α} -based QPS scores. This extension would give different relative weights between quality and privacy criteria, enabling a stronger emphasis on either quality ($\alpha < 1$) or privacy ($\alpha > 1$) of synthetic data. Another possible modification to enhance the importance of privacy considerations would be to include other privacy scores, such as the nearest neighbor adversarial accuracy risk developed by Yale A et al³⁷, in the global metaQPS metrics.

All in all, MIIC-SDG is particularly effective in generating synthetic data from biomedical datasets, which typically include a limited number of patients (< 1000) and a complex multivariate joint distribution. In this context, MIIC-SDG tends to outperform or be comparable to other state-of-the-art methods (Figure 8) with similar execution times (Supplementary Figure 8). However, to generate synthetic health data from very large datasets, such as electronic health records, deep learning approaches may be more suitable.

DATA AVAILABILITY

No new samples or human data were collected for the purpose of this study. All data used is publicly available online.

The METABRIC data comes from public data available in the cBioPortal repository: https://www.cbioportal.org/study/summary?id=brca_metabric.

The IMvigor210 dataset³⁴ is publicly available, through an R package, at the following address: <http://research-pub.gene.com/IMvigor210CoreBiologies/>.

The DIABETES dataset is publicly available on the CDC website: https://www.cdc.gov/brfss/annual_data/annual_2015.html or on kaggle: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>.

CODE AVAILABILITY

The MIIC-SDG algorithm is available on github at the following address <https://github.com/miicTeam/miic-sdg> as an R package, named `miicsdg`. It generates the synthetic data as a data frame, and the DAG that is used to sample the synthetic data. The DAG can be visualized with the tool available in the MIIC web server https://miic.curie.fr/vis_NL.php to better appreciate the relationships between variables.

ACKNOWLEDGEMENTS

Not applicable. No funding was granted for the study.

AUTHOR CONTRIBUTIONS

All authors contributed to the concept and outline of the manuscript. N.S. and F.G. drafted the paper. N.S. and F.G. are co-first authors and contributed equally. N.S. designed and implemented the data generation algorithm, N.S, H.I. and N.L. designed and implemented the MIIC-to-DAG algorithm; N.S. and F.G. performed research and comparative analyses. N.S., F.G., H.I., N.L., L-P.A. and J.D. contributed to data interpretation. All authors participated in revising the manuscript and approved the completed version.

COMPETING INTERESTS

The authors declare no competing interests

REFERENCES

1. Samarati, P. & Sweeney, L. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.
2. Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkatasubramanian, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**, 3-es (2007).
3. Li, N., Li, T. & Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. in *2007 IEEE 23rd International Conference on Data Engineering* 106–115 (2007). doi:10.1109/ICDE.2007.367856.
4. Synthetic data generation for tabular health records: A systematic review | Elsevier Enhanced Reader.
<https://reader.elsevier.com/reader/sd/pii/S0925231222004349?token=A1A2F0F547312B6E8B6E57513E22AE6A24312CBAA3E0AC15EF5856CFE5E45CAA9ACF0F8DAD5150741B62329381769E48&originRegion=eu-west-1&originCreation=202209061>

50235 doi:10.1016/j.neucom.2022.04.053.

5. Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *Npj Digit. Med.* **3**, 1–13 (2020).
6. Zhang, Z., Yan, C., Mesa, D. A., Sun, J. & Malin, B. A. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 99–108 (2019).
7. Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc. JAMIA* **26**, 228–241 (2018).
8. Choi, E. *et al.* Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. Preprint at <https://doi.org/10.48550/arXiv.1703.06490> (2018).
9. Yuan, H., Zhou, S. & Yu, S. EHRDiff: Exploring Realistic EHR Synthesis with Diffusion Models. Preprint at <https://doi.org/10.48550/arXiv.2303.05656> (2024).
10. He, H., Zhao, S., Xi, Y. & Ho, J. C. MedDiff: Generating Electronic Health Records using Accelerated Denoising Diffusion Model. Preprint at <https://doi.org/10.48550/arXiv.2302.04355> (2023).
11. Guillaudeux, M. *et al.* Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *Npj Digit. Med.* **6**, 1–10 (2023).
12. Yoon, J. *et al.* EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *Npj Digit. Med.* **6**, 1–11 (2023).
13. Ghosheh, G. O., Li, J. & Zhu, T. A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Comput. Surv.* **56**,

- 147:1-147:34 (2024).
14. Yan, C. *et al.* A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* **13**, 7609 (2022).
 15. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Comput. Biol.* **13**, e1005662 (2017).
 16. Cabeli, V. *et al.* Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Comput. Biol.* **16**, e1007866 (2020).
 17. Sella, N. *et al.* Interactive exploration of a global clinical network from a large breast cancer cohort. *Npj Digit. Med.* **5**, 1–10 (2022).
 18. Sella, N., Verny, L., Uguzzoni, G., Affeldt, S. & Isambert, H. MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* **34**, 2311–2313 (2018).
 19. Ribeiro-Dantas, M. da C. *et al.* Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients. *iScience* **27**, (2024).
 20. Ankan, A. & Panda, A. pgmpy: Probabilistic Graphical Models using Python. in 6–11 (Austin, Texas, 2015). doi:10.25080/Majora-7b98e3ed-001.
 21. Qian, Z., Cebere, B.-C. & van der Schaar, M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. Preprint at <https://doi.org/10.48550/arXiv.2301.07573> (2023).
 22. Nowok, B., Raab, G. M. & Dibben, C. synthpop: Bespoke Creation of Synthetic Data in R. *J. Stat. Softw.* **74**, 1–26 (2016).

23. Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling Tabular data using Conditional GAN. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
24. PrivBayes: Private Data Release via Bayesian Networks: ACM Transactions on Database Systems: Vol 42, No 4. <https://dl.acm.org/doi/abs/10.1145/3134428>.
25. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
26. Dwork, C., McSherry, F., Nissim, K. & Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. in *Theory of Cryptography* (eds. Halevi, S. & Rabin, T.) 265–284 (Springer, Berlin, Heidelberg, 2006). doi:10.1007/11681878_14.
27. Jordon, J., Yoon, J. & Schaar, M. van der. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. in (2018).
28. Stadler, T., Oprisanu, B. & Troncoso, C. Synthetic Data -- Anonymisation Groundhog Day. Preprint at <http://arxiv.org/abs/2011.07018> (2022).
29. Yoon, J., Drumright, L. N. & van der Schaar, M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE J. Biomed. Health Inform.* **24**, 2378–2388 (2020).
30. El Emam, K., Mosquera, L. & Bass, J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J. Med. Internet Res.* **22**, e23139 (2020).
31. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
32. van Breugel, B., Sun, H., Qian, Z. & van der Schaar, M. Membership Inference

Attacks against Synthetic Data through Overfitting Detection. Preprint at <https://doi.org/10.48550/arXiv.2302.12580> (2023).

33. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
34. Mariathasan, S. *et al.* TGF β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
35. CDC - 2015 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2015.html (2022).
36. [privacy/tensorflow_privacy/privacy_tests/membership_inference_attack at master · tensorflow/privacy](https://github.com/tensorflow/privacy/tree/master/tensorflow_privacy/privacy_tests/membership_inference_attack). *GitHub*
https://github.com/tensorflow/privacy/tree/master/tensorflow_privacy/privacy_tests/membership_inference_attack.
37. Yale, A. *et al.* Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* **416**, 244–255 (2020).
38. TGF-b attenuates tumor response to PD-L1 blockade by contributing to exclusion of T cells. <http://research-pub.gene.com/IMvigor210CoreBiologies/>.

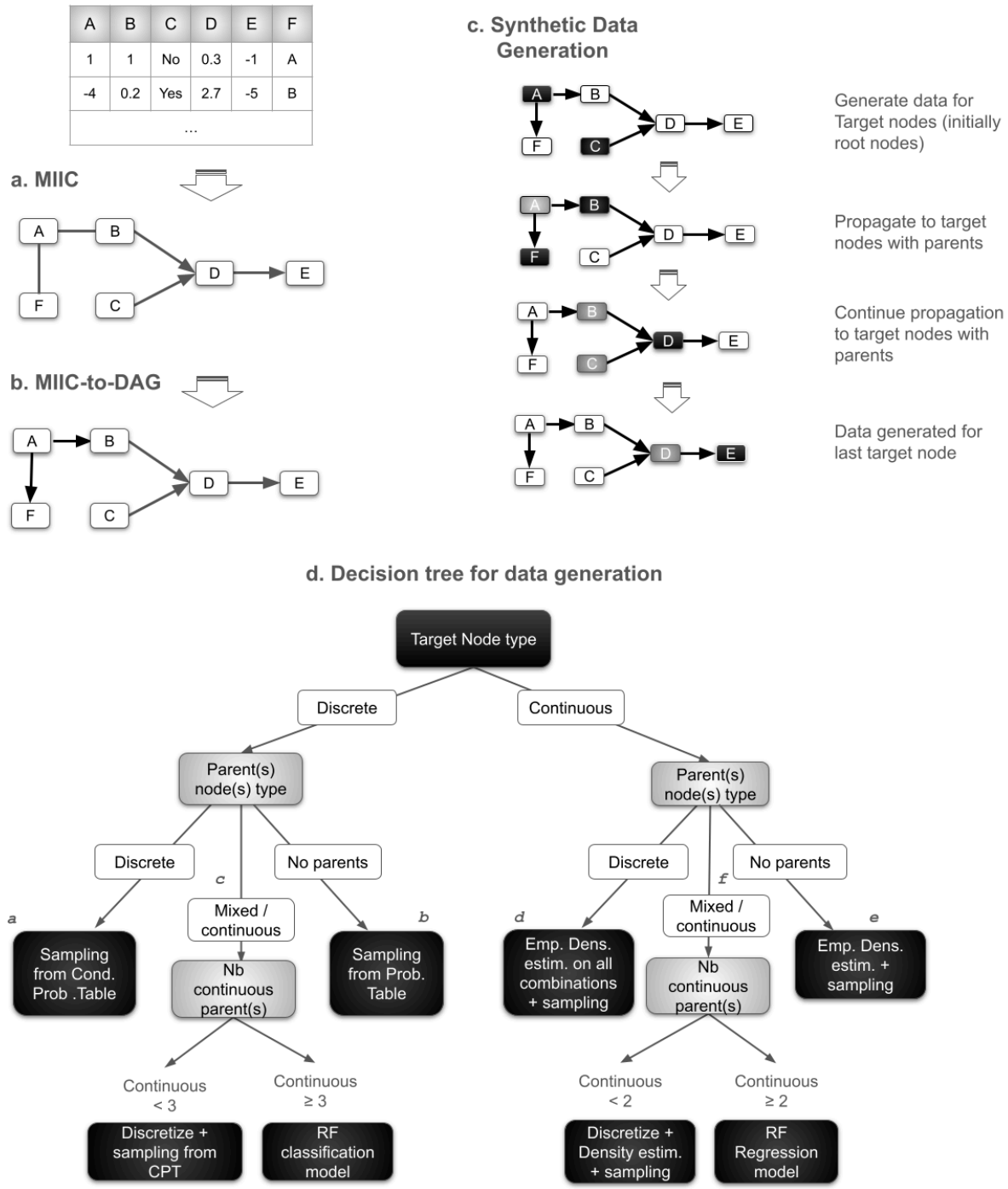


Figure 1. MIIC-SDG pipeline. This illustration shows the complete data generation process, from the original data table to the generated data, following the 3 main steps described in section I of Methods. a) Execution of the MIIC algorithm from the original data table. This step generates a graph where nodes represent the variables of the data matrix and edges represent direct associations between variables. b) Transformation of

the graph into a directed acyclic graph (DAG) through the MIIC-to-DAG algorithm. c) Generation of the data using the original data table and the reconstructed DAG. d) Details on the data generation step: each scenario takes into account the variable type of the target and parents nodes to adapt the sampling procedure (branches a, b, c, d, e, f are described in the section C of Methods. RF: Random Forest; Cond. Prob. Table: Conditional Probability Table; Prob. Table: Probability Table; Emp. Dens.: Empirical density).

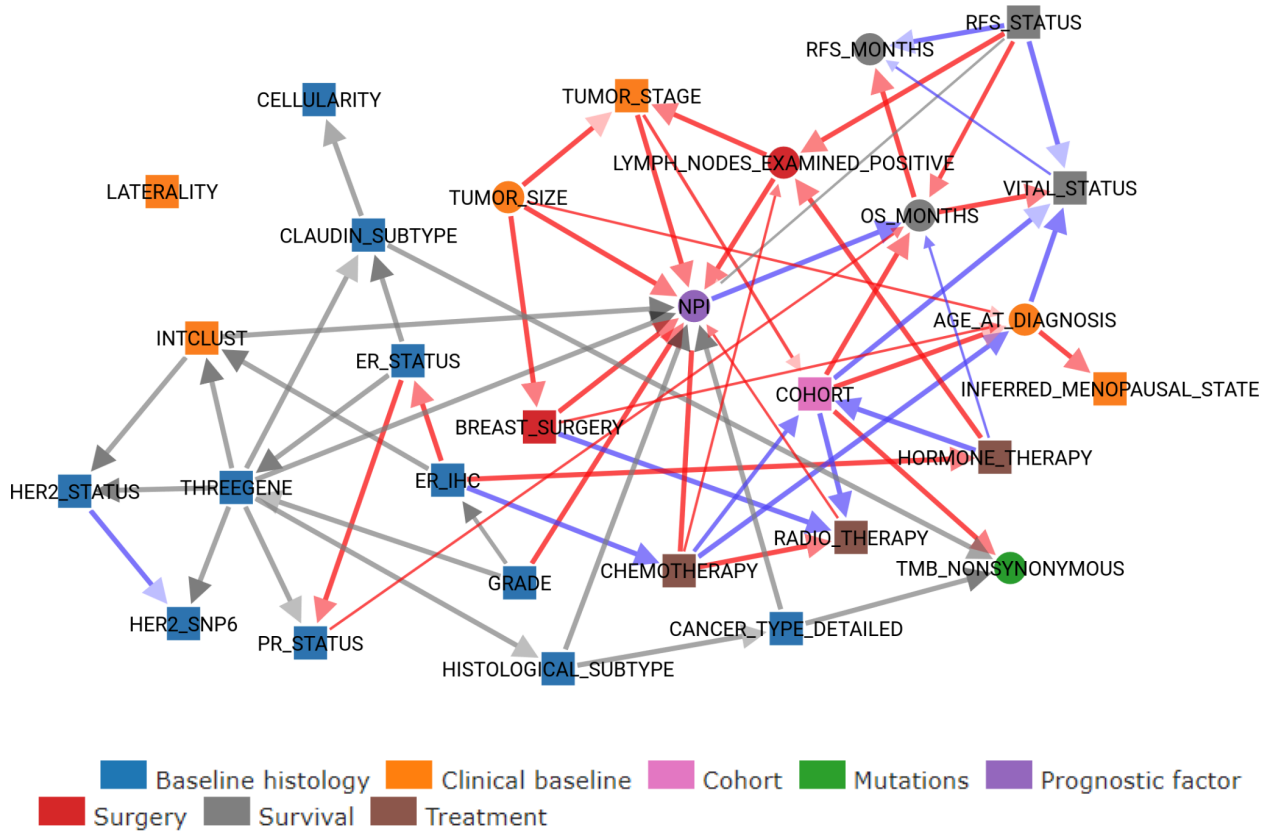


Figure 2. Network reconstructed by MIIC-to-DAG starting from the network obtained by MIIC from the full METABRIC dataset. The network is learned with the MIIC-to-DAG algorithm (step B of MIIC-SDG) starting from the network obtained by MIIC, Supplementary Figure 1, with the parameters defined in Supplementary Table1. This network corresponds to the directed acyclic graph used to generate the synthetic data for the step C of MIIC-SDG. This network is visible at the following address: https://miic.curie.fr/job_results_NL.php?id=METABRIC_DAG

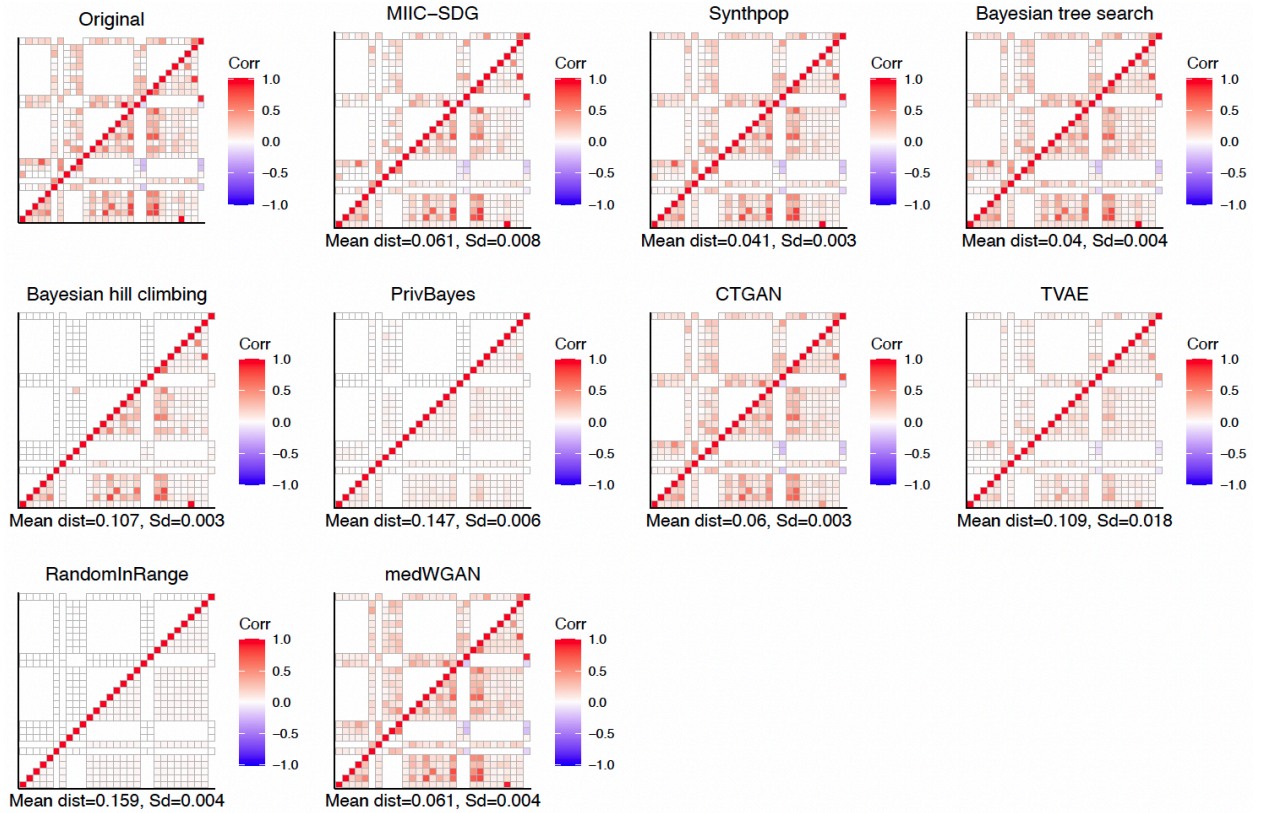


Figure 3. Correlation matrices evaluated on 1000 samples for METABRIC dataset. Correlation for each x,y combination is evaluated as the mean value over all executions with the same sample size.

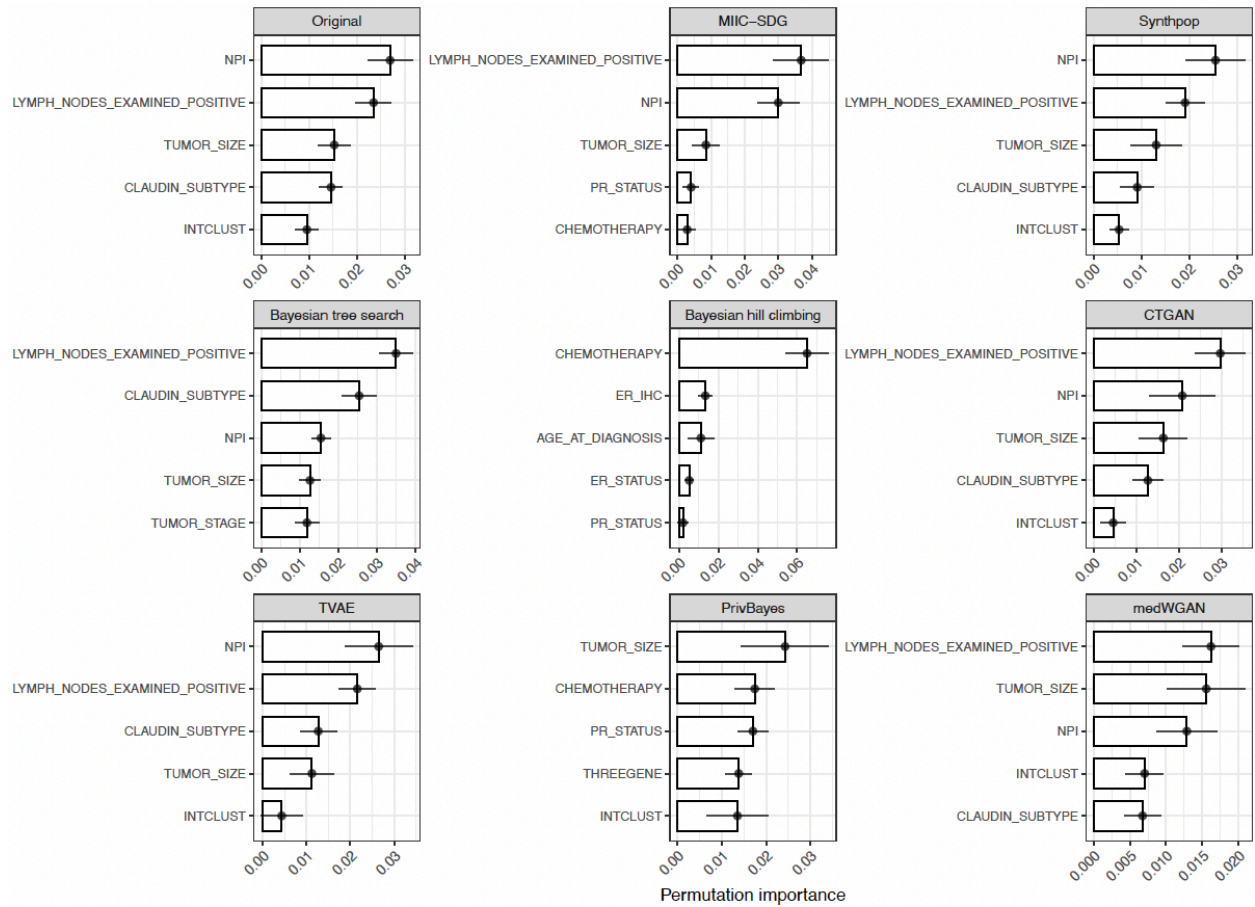


Figure 4. Features permutation importance to predict overall survival. We used a Survival Random Forest model fitted on a set of 1977 patients from the METABRIC dataset.

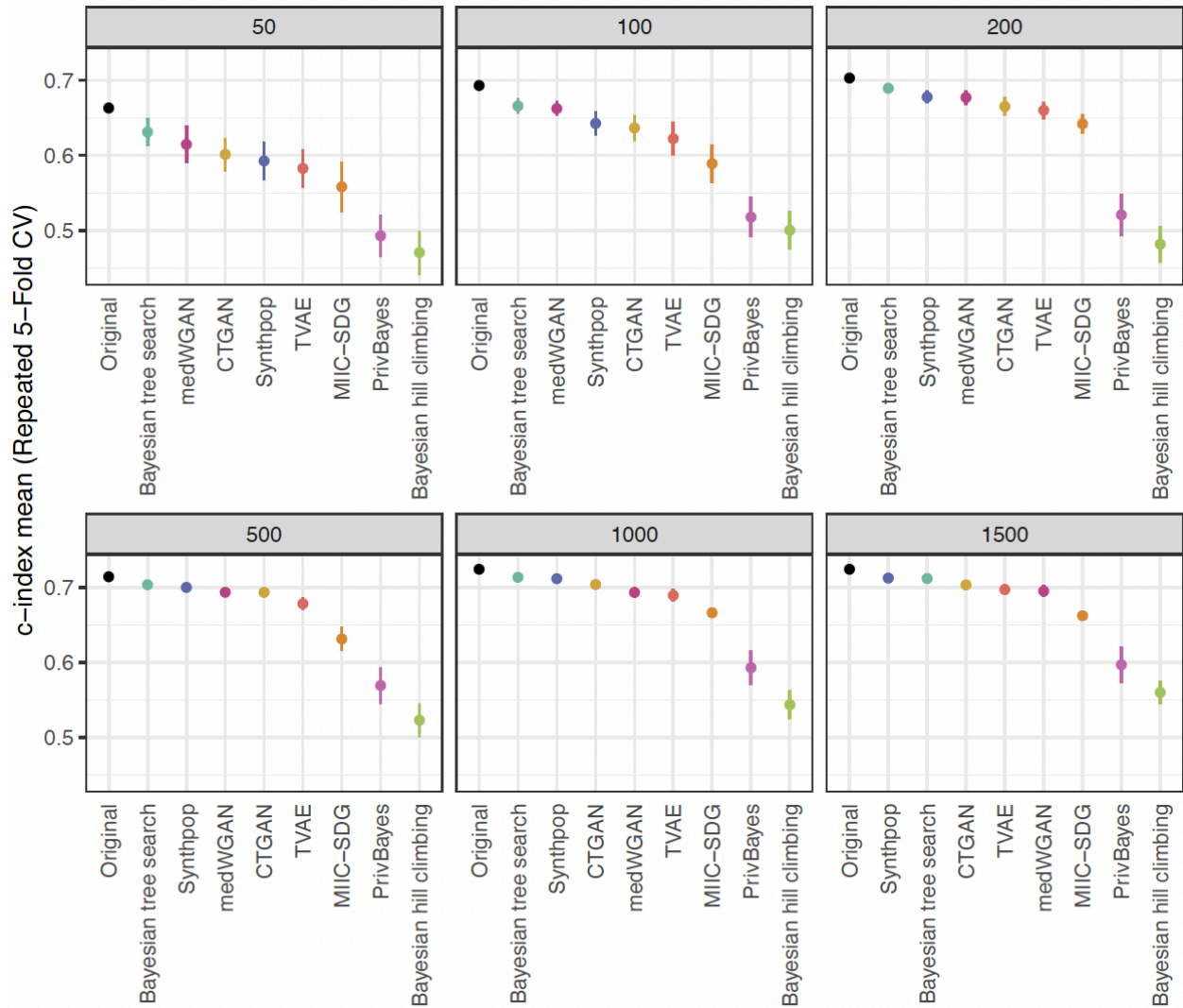


Figure 5. K-Fold Cross-validated c-index estimates. This analysis was made using a Survival Random Forest model to predict Overall Survival in the METABRIC dataset in function of sample size (K=5).

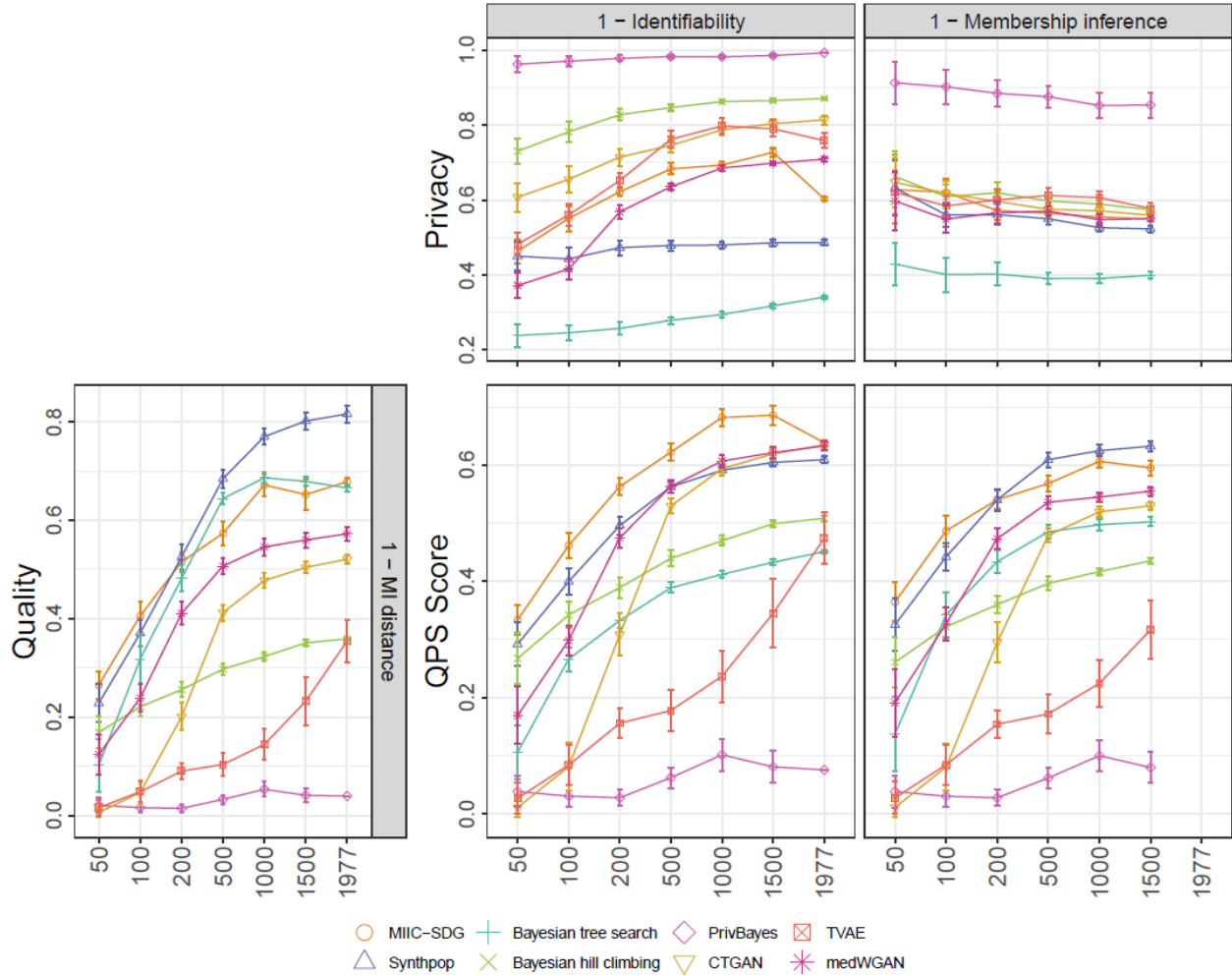


Figure 6. Quality, privacy and quality-privacy scores (QPS). This comparison is made using Mutual Information distance as quality measure and privacy evaluated using identifiability and membership inference scores.

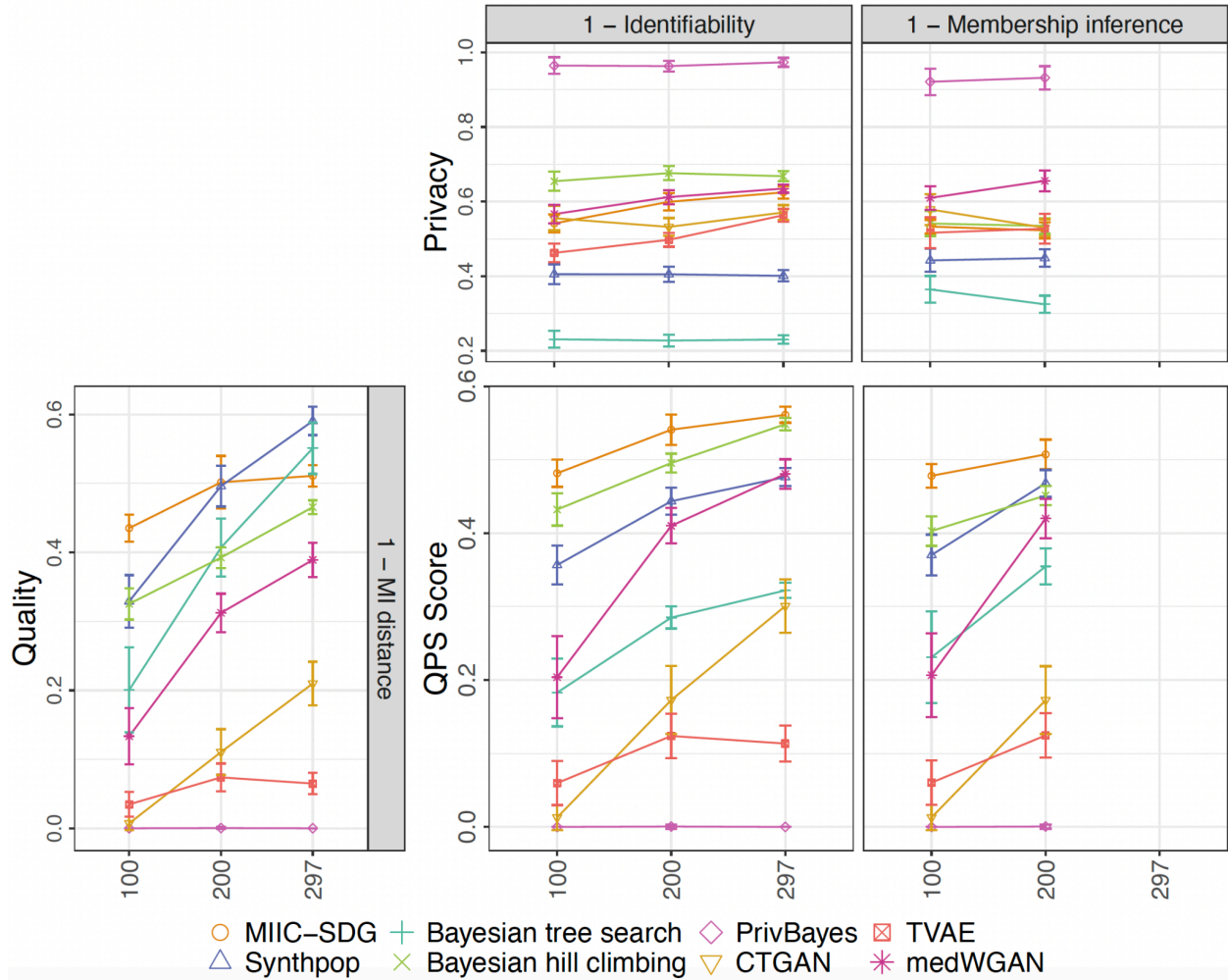


Figure 7. Quality, privacy and quality-privacy scores (QPS) on IMvigor210. This comparison is made using Mutual Information distance as quality measure and privacy evaluated using identifiability and membership inference scores. Data from the IMvigor210 trial (Bladder cancer).

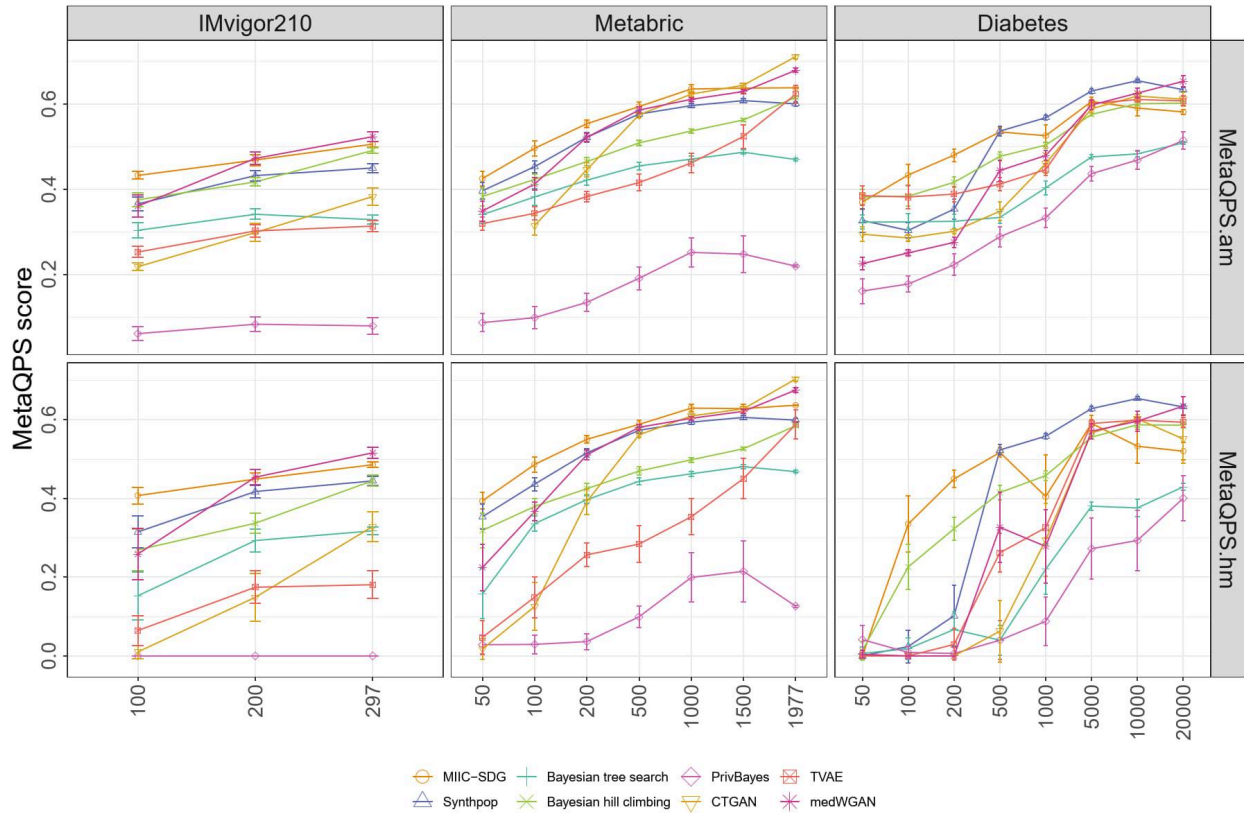


Figure 8: Meta-QPS scores of each dataset (IMvigor210, Metabric and Diabetes). These plots summarize the benchmark results by integrating all quality and privacy scores into single metaQPS metrics for each dataset and all sample sizes analyzed in this study. MetaQPS.am corresponds to the F1-score between the arithmetic mean of quality scores and the arithmetic mean of the privacy scores. MetaQPS.hm corresponds to the F1-score between the harmonic mean of quality scores and the harmonic mean of the privacy scores (see details in Methods).

	N = 50	N = 100	N = 200	N = 500	N = 1000	N =1500	N = 1977
MIIC-SDG	1.3 (1.3)	0.8 (0.9)	1.1 (0.9)	1.7 (1.6)	2.5 (1.6)	4.3 (2.7)	3.8 (1.2)
Synthpop	0.3 (0.5)	0.2 (0.4)	0.3 (0.7)	0.8 (0.8)	0.3 (0.7)	0.3 (0.7)	0.5 (0.7)
Bayesian tree search	0.1 (0.3)	0.9 (0.9)	1.7 (0.5)	2.0 (0)	2.4 (0.5)	2.6 (1.3)	2.0 (0)
Bayesian hill climbing	0.3 (0.5)	0.9 (0.9)	1.8 (0.4)	2.1 (0.3)	2.0 (0)	2.9 (0.6)	2.0 (0)
PrivBayes	11.3 (2.8)	16.2 (1.1)	18.1 (2.2)	17.2 (4.4)	15.8 (2.1)	20.7 (3.6)	21.0 (0)
CTGAN	6.8 (2.7)	8.5 (3.4)	10.9 (2.5)	12.8 (4.2)	15.6 (4)	16.0 (2.9)	16.3 (2.7)
TVAE	4.5 (1.8)	5.1 (2.4)	10.1 (3.2)	14.6 (4.7)	19.9 (4)	16.4 (3.7)	15.9 (2.1)
RandomInRange	13.6 (2)	19.2 (1)	22.9 (1.5)	26.0 (0.9)	26.6 (0.7)	27.2 (0.8)	27.5 (0.5)
medWGAN	5.4 (1.3)	7.0 (0.9)	7.8 (1.1)	11.0 (2.3)	14.5 (2.1)	16.5 (2.4)	16.7 (1.9)

Table 1. Average number of significantly different features between real and synthetic distribution on METABRIC dataset. This analysis was made using chi-squared or Wilcoxon tests, with a 0.05 p-value threshold. Standard deviation is reported in parentheses. Top method for each sample size is reported in bold font.