Received: 7 September 2019 | Revised: 30 September 2020 | Accepted: 10 November 2020

DOI: 10.1111/coin.12427

SPECIAL ISSUE ARTICLE

Computational Intelligence WILEY

Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy

Zhenchen Wang[1] | **Puja Myles**[1] | **Allan Tucker**[2]

[1]CPRD, Medicines and Healthcare Products Regulatory Agency, London, UK

[2]Department of Computer Science, Brunel University London, London, UK

Correspondence

Zhenchen Wang, CPRD, Medicines and Healthcare Products Regulatory Agency, London, UK.
Email: zhenchen.wang@mhra.gov.uk

Funding information

Department for Business, Energy and Industrial Strategy, Grant/Award Number: 104676; Innovate UK; Pioneer Fund

Abstract

Electronic healthcare record data have been used to study risk factors of disease, treatment effectiveness and safety, and to inform healthcare service planning. There has been increasing interest in utilizing these data for new purposes such as for machine learning to develop predictive algorithms to aid diagnostic and treatment decisions. Synthetic data could potentially be an alternative to real-world data for these purposes as well as reveal any biases in the data used for algorithm development. This article discusses the key requirements of synthetic data for multiple purposes and proposes an approach to generate and evaluate synthetic data focused on, but not limited to, cross-sectional healthcare data. To our knowledge, this is the first article to propose a framework to generate and evaluate synthetic healthcare data with the aim of simultaneously preserving the complexities of ground truth data in the synthetic data while also ensuring privacy. We include findings and new insights from synthetic datasets modeled on both the Indian liver patient dataset and UK primary care dataset to demonstrate the application of this framework under different scenarios.

KEYWORDS

electronic healthcare records, privacy, synthetic data, synthetic data evaluation, synthetic data generation

1 INTRODUCTION

Electronic healthcare record (EHR) data are a rich source of clinical symptoms, diagnoses, investigations, and treatments. The recent development of intelligent applications makes these data attractive for the application of various data mining and machine learning algorithms. For example, in Khalid et al.'s work, the data are used to yield new insights into drug use patterns[1] and Ravizza et al. proposed an application for diagnosis and prediction of diseases.[2] However, there is a need for developing a synthetic dataset that would complement such rich real-world data for various reasons outlined below.

(1) Ease of access—Access to individual record level real-world data, even in pseudonymized format with strong personal identifiers removed, tends to be strictly regulated to control the risk of inadvertent patient reidentification according to Sebastian et al.'s work.[3] Moreover, the legal bases for sharing of these routinely collected data may present restrictions on use that need to be monitored by data controllers. The ability to streamline data access approvals with synthetic datasets could increase the speed of research innovation.

(2) Cost-efficiency—In the context of healthcare data collection, using a synthetic data generation model for benchmarking and validation is significantly more cost-efficiency than expanding the population coverage of real-world data due to the cost of scaling-up collection and processing pipelines.

(3) Test efficiency—Lee and Whalen's work identified that using a synthetic data generation model can efficiently improve algorithms or functions in an information system[4] by generating desired data on-the-fly. For example, one typical scenario could be to test the scalability and robustness of an algorithm.

(4) Patient privacy protection—The social-demographic and health-related content in the healthcare data makes patient identification more likely and therefore a fully synthetic approach can better mitigate this risk according to Park and Ghosh.[5]

(5) Completeness—It can be difficult to conduct unbiased data research if there are inherent biases in the data. Nowok et al. argue that synthetic data can supplement real data by either filling gaps or enlarging a subgroup dataset.[6] In addition, Wu et al. discovered that anonymization measures for real data may compromise data utility due to information loss.[7]

(6) Benchmarking and validation capabilities—This is useful when comparing different machine learning methods against a standardized dataset while focusing on a specific set of diseases (e.g., cardiovascular diseases). Synthetic data could be generated to be intentionally distinct from real data to reveal biases in algorithmic performance.

Despite all the advantages outlined above, there is no synthetic data generation and evaluation approach that can be applied to healthcare data to ensure that the generated data preserves key ground truth characteristics (such as sensible biological relationships between variables) while ensuring privacy.

In this article, we propose a framework to focus on synthetic data generation ensuring data utility, that is, clinically meaningful, and the preservation of patient privacy. The following four additional key requirements are envisioned to be critical for making synthetic data usable:

Preservation of biological relationships. The chosen data variables and target studies should preserve the correct underlying biological relationships (e.g., female specific diseases should not include male patients) or well-established clinical symptom-diagnosis pairs should be preserved (e.g., excessive thirst because of diabetes).

Univariate distance. Each concerned variable in the synthetic data should have similar fundamental aggregated statistics to the ground truth (e.g., similar distributions) for both continuous variables and categorical variables.

Multivariate distance. Data with multiple dimensions often have correlated structure between data variables. Retaining such correlational structure in synthetic data is crucial to ensure that it is a truthful representation of the real world. In some studies, initial knowledge (e.g., clinical expertise or familiarity with real-world distributions) may be required to inform the development but this would only be feasible when the number of concerned variables is relatively small. However, when the variable numbers are large such as in Ravizza et al.'s work,[2] a manual approach will be extremely resource-intensive and challenging.

Preservation of patient privacy. Privacy might be a concern even in the case of fully synthetic data. This can occur when synthetic data generation produces a very similar dataset in terms of aggregated characteristics to real-world data. For example, a small number (i) of rare disease cases in the synthetic data occur in patients with similar ages living in the same geographical region. Thus, when i is sufficiently small, there is a chance that these outliers can be identified from the data. Essentially, this occurs when a synthetically generated patient shares the same (or similar) characteristics as a real patient purely by chance. As a result, it is important to consider mechanisms which can be put in place to protect against such a scenario.

The article is organized as follows. Section 2 reviews existing approaches and challenges to generating synthetic datasets, privacy concerns relating to generated data and multidimensional complexities in EHR. Section 3 presents the proposed framework followed by two case studies in Section 4, and Section 5 concludes the article by discussing the use and outlook of synthetic data.

2 | RELATED WORKS

There are researches concerning synthetic data generation, while preserving privacy in the data. There are also many studies that address the complexities of healthcare data such as in Mandal et al.[8] and Chen et al.'s works.[9] The objective of this section is to review previous work in these areas and to discover the lessons learned that can contribute to the proposed framework.

2.1 | Synthetic data generation methods

In general, synthetic data generation methods can be categorized into three groups.

2.1.1 | Group 1: Generating synthetic data based on some statistical properties of the real-world data

This approach is useful when real data are difficult to access, for example, the data are transient and difficult to collect, or the data are scarce such as in the case of rare diseases, or the distribution of events is highly imbalanced, such as in the case of outliers in Robnik's work.[10] In Ruscio and Kaczetow's work,[11] data are sampled from the population distribution. In another practical example showcased by Lee et al., the velocity property is used to generate synthetic data for various tactical moving objects in the military context.[12] In Buczak et al.'s work,[13] care patterns discovered from real patients are used for synthetic patients whereas Riano and

Fernandez-Perez[14] incorporate both statistical information and rules based on expert knowledge to simulate episodes of care.

2.1.2 | Group 2: Adding noise

This approach based on adding noise to a small sample of data, which is particularly useful when only part of the real-world data needs to be regenerated such as in Syahaneim et al.,[15] Iftikhar et al.,[16] and Cano and Torra's works.[17] A relevant technique often used is data imputation, which addresses missing values in the data such as in Kontopantelis et al.[18] and Caiola and Reiter's works.[19]

2.1.3 | Group 3: Using machine learning techniques to generate/extend the dataset using prediction and inference

These techniques can be applied to generate both semisynthetic data and fully synthetic data. In Yang et al.'s work,[20] a hidden Markov model is used to discover the hidden properties of data and generate the semisynthetic medical process data. In Patki et al.'s work,[21] generative models are used to create both semi- and fully synthetic data in relational databases in replace of real data.

Despite the available methods, generating synthetic healthcare data requires additional consideration due to its longitudinal nature and long data format (e.g., selecting a set of variables from the whole dataset for a specific study) before using these approaches.

2.2 | Privacy preservation

Privacy concerns have heightened with the advent of the General Data Protection Regulation (GDPR) within the EU especially when sensitive patient data are involved.[22] A range of approaches to mitigate the risk of patient privacy disclosure such as perturbation, condensation, randomization, and fuzzy based methods have been described in Langarizadeh et al.'s work[23] to eliminate links between identifiable data and the data subject defined by ISO (International Organization for Standardization)[24] and ICO (Information Commissioner's Office).[25]

However, anonymization of data can compromise its utility according to Wu et al.[7] if important information is removed. In addition, the anonymized data may still present a residual risk to privacy. In Sweeney's work,[26] it was found that 87% of the population in the United States had reported characteristics that made them unique based only on five-digit ZIP, gender, and date of birth. About half of the population are likely to be uniquely identified by only gender, date of birth, and city/town/municipality in which the person resides.[26] In Narayanan and Shmatikov's work,[27] a robust algorithm is further proposed to deanonymize large sparse datasets.

A fully synthetic dataset can effectively tackle the privacy issue in a sense that all data values can be completely different from real-world data. However, in the context of clinical studies it may be necessary to rely on a restricted set of values to ensure clinical meaningfulness. The risk of privacy disclosure might still exist when outliers in synthetic data are the same as those in real-world data by chance.

2.3 | Complexity in electronic healthcare data

Complexity can come from two levels: first, access to heterogenous data sources according to Bache et al.'s work[28] and second, the complex relationships inherent in the data. The term "heterogenous" in this context, is more about data representations. This could mean cross-sectional data in which each record summarizes a longitudinal history versus the longitudinal data in which each time-stamped record forms the history of a healthcare path. Cross-sectional data are often favored by researchers. However, the conversion from longitudinal to cross-sectional is often a complicated data processing task. The term "relationships" can be defined as relationships within the data model of different entities such as patient entity, clinical observation entity, drug record entity, and so on, and of different variables within each entity, and therefore the complexity is partially a result of not isolating these relationships. As a result, a comprehensive data model is required to take multiple variables into account.

In Pedersen and Jensen's work,[29] an example is given to demonstrate a multidimensional model and nine requirements of data modeling on patient diagnosis episodes, for example, commonly occurring many-to-many relationships between patients and diagnoses should be handled by the model. In addition, the underlying many-to-many biological relationships are derived from empirical studies and observations, for example, a predefined set of variables such as cardiovascular risk factors described in Hippisley-Cox et al.'s work.[30] Learning these many-to-many relationships within and across the data entities remains a challenge so that the fidelity of the clinical information can be retained.

3 | SYNTHETIC DATA FRAMEWORK

3.1 | Overview

A synthetic data framework is proposed with the objective of enabling the synthesis of healthcare data while ensuring a high degree of similarity (i.e., biological relationships and biological values) between ground truth and the synthetic data as well as preserving privacy.

The framework (see Figure 1) outlines a set of processes including the ground truth selection process as input, the synthetic data generation process, the evaluation process, and ultimately the generation of the sensible synthetic data selection as an output.

The ground truth selection process entails two linked tasks which are a privacy sensitive variable identification task and a biological relationship definition task. The results from these two tasks can contribute to the study data model definition, which in turn determines the suitable synthetic data generation model.

The synthetic data generation process aims to produce a set of synthetic data candidates and ensure the data quality, that is, data values and format of the resultant data are consistent. The evaluation process focuses on assessing the four key criteria between the selected ground truth and the generated synthetic datasets: biological relationship preservation, univariate distance, multivariate distance, and privacy preservation.

The sensible synthetic data selection is a recursive process which ends when all evaluation criteria are met.

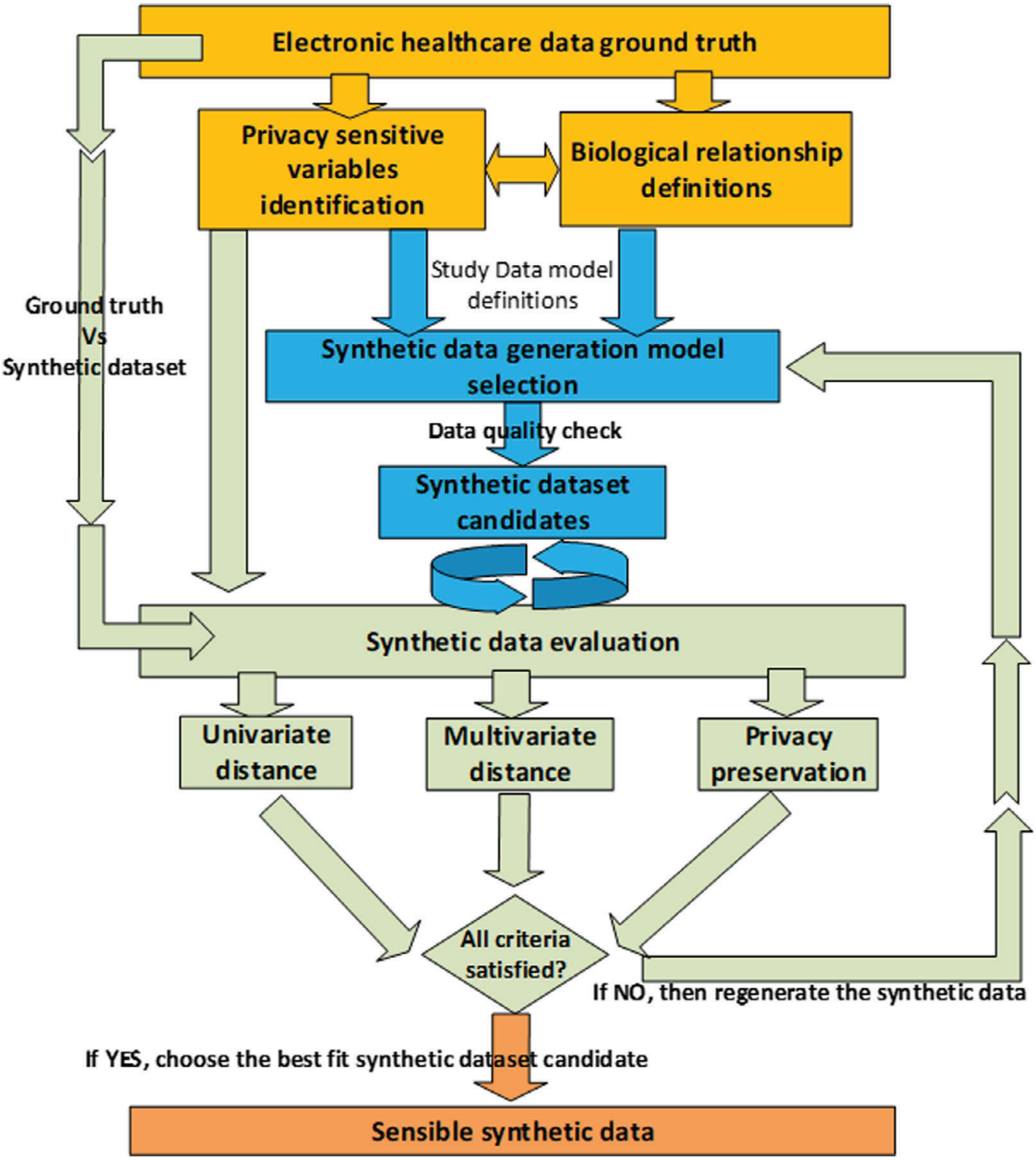

Electronic healthcare data ground truth
Privacy sensitive variables identification
Biological relationship definitions
Study Data model definitions
Ground truth Vs Synthetic dataset
Synthetic data generation model selection
Data quality check
Synthetic dataset candidates
Synthetic data evaluation
Univariate distance
Multivariate distance
Privacy preservation
All criteria satisfied?
If NO, then regenerate the synthetic data
If YES, choose the best fit synthetic dataset candidate
Sensible synthetic data

FIGURE 1 Synthetic data framework overview [Color figure can be viewed at wileyonlinelibrary.com]

3.2 | Ground truth selection process

The ground truth is closely dependent on the study context. In this process, the selected variables need to go through two tasks: namely *privacy sensitive variable identification* and *biological relation definition*.

In most cases, each variable alone may not cause privacy issues, but a combination of variables (tuple) may present an increased risk of reidentification. However, without knowing the distributions of these tuples it is usually impossible to identify these variables. In these cases, referencing relevant official guidance and standards from bodies like the UK's Information Commissioner's Office (ICO*) to predefine a set of sensitive variables can be a good solution.

Once the individual sensitive variables are listed, the researchers can further define variable combinations among them that are deemed to be privacy sensitive. Examples of variables defined as sensitive patient data include familial relationships, marital and occupational status, sexuality, eligibility for social benefits.

Well-established biological relationships between variables and the study target also need to be defined at this stage. For example, the condition lupus can be relevant when studying hypertension.

3.3 | Synthetic data generation process considerations

In this work we also propose some of the considerations for using previously described synthetic data generation approaches in Section 2.1 with respect to (a) the identification of biological relationships between variables and (b) the privacy sensitive nature of variables as well as (c) the number of variables.

Adding noise can be a good solution when full synthetic data are not required, that is, most of the real-world data can be released. A common challenge for state-of-the-art methods such as differential privacy[31] is mainly on how to balance the added noise with the utility of the data. More noise usually means less privacy risk, but also means less utility on the data. Preserving biological relationships between variables may not be a concern when a small amount of noise is added. However, a large amount of noise will require additional checks if the original correlations between variables are preserved to ensure the data utility.

An expert-driven approach is especially useful and efficient when data are difficult to access. Given key statistical information for each variable, synthetic data can be created by combining the randomly generated data for each variable based on statistical metrics such as probability distribution functions and maximum/minimum values. These can often be extracted from the existing literature. When the data can be accessed, an algorithm such as Copula[32] can be an effective synthetic data generation method as it can effectively capture the dependencies between multiple variables. However, when the total number of variables becomes large this may not be able to capture the hidden patterns among variables.

Machine learning algorithms can be an option when requirements on both data utility and privacy are stringent while biological relationships between variables are numerous and complex. This is because these algorithms are often good at discovering hidden patterns in the data. Despite these advantages, choosing a good algorithm is often dependent on the requirements on the transparency of the algorithm. When transparency is required, then the interpretation of an algorithm's model architecture and intermediary results become important. Building a "black box" type machine learning model may result in synthetic data where the underlying logic are not fully understood or trusted. As a recommendation, when there are alternatives, a more transparent algorithm such as Bayesian networks[33] should always be in considered first.

Once synthetic data are generated, a data quality check may be required for variables with continuous values so that they are biologically plausible, Table 1 lists the common data quality checks required and the corresponding actions proposed.

3.4 | Evaluation process

As synthetic data generation models can produce a set of synthetic data candidates, the evaluation process must be able to differentiate these candidates. Here, we use the "distance" to quantify

TABLE 1 Biological data quality checklist

Biological value checklist	Actions
Negative biological values	Remove instances, for example, glucose < 0
Minimum/Maximum value bound	Remove instances beyond the bounds, for example, age > 300
Decimal places	Format to ground truth's decimal places, for example, 1 decimal place to BMI value
Minimum/maximum pairwise difference	Remove instances beyond the bounds, for example, systolic and diastolic difference > 100

the difference between synthetic dataset and the ground truth via a general function $dis(X^s, X^g)$, where the X^s and X^g represent the input spaces of synthetic data and ground truth, respectively. In general, the closer the distance, more similar two datasets are.

3.4.1 | Univariate distance

Univariate distance can be defined based on two types of variables: discrete variables D, the number of which is denoted by I and continuous variables C, the number of which is denoted by K. For the discrete variables, the probability distribution p is used to represent the discrete ith individual variable $\boldsymbol{d_i^s}$ and $\boldsymbol{d_i^g}$, $i = [1, I]$; for the continuous variables, the density function f is used to represent the kth individual variable $\boldsymbol{c_k^s}$ and $\boldsymbol{c_k^g}$, $k = [1, K]$; The univariate distance can be defined as two Equations (E1) and (E2):

$$\boldsymbol{E_1} = \boldsymbol{dis}(\boldsymbol{p}(\boldsymbol{d_i^s}), \boldsymbol{p}(\boldsymbol{d_i^g})), \text{where } \{\boldsymbol{d_i^s}, \boldsymbol{d_i^g}\} \in \boldsymbol{D}, \tag{E1}$$

$$\boldsymbol{E_2} = \boldsymbol{dis}(\boldsymbol{f}(\boldsymbol{c_k^s}), \boldsymbol{f}(\boldsymbol{c_k^g}), \text{where } \{\boldsymbol{c_i^s}, \boldsymbol{c_i^g}\} \in \boldsymbol{C}. \tag{E2}$$

During the univariate distance evaluation, each variable is independently assessed and the *dis* function needs to be consistent for a variable type. Here, the probability difference is used as a *dis* function for $\boldsymbol{E_1}$, and a hypothesis test results, that is, p-value, for example, Kolmogorov–Smirnov test (KS test), can be used as a *dis* function for $\boldsymbol{E_2}$.

3.4.2 | Multivariate distance

The multivariate analysis is required to compare the interrelationship and patterns of data instances within two datasets $\{\boldsymbol{s_1^s}, \dots, \boldsymbol{s_i^s}\} \in \boldsymbol{S^s}$ for synthethic dataset and $\{\boldsymbol{s_1^g}, \dots, \boldsymbol{s_j^g}\} \in \boldsymbol{S^g}$ for ground truth dataset, where lowercase $\boldsymbol{s}$ represents the data instance, that is, total of i generated synthetic data instances and total of j ground truth data instance used to generate the synthetic data, and uppercase $\boldsymbol{S}$ represents the collection of data instances.

The purpose of the *dis* function in the multivariate distance test is to measure the Euclidean distance in a, possibly transformed, input space $X \in \mathbb{R}^k$, where n dimensions can be reduced and projected to shared k dimensions $(0 < k < n)$ via $\boldsymbol{f_n^k}$ and n is the total number of variables in data space $\mathbb{R}^n$. Existing multivariate analysis methods such as ordination techinques including nonmetric multidimensional scaling (NMDS) proposed by Kruskal[34] can be used. The multivariate distance function between two datasets thus can be defined as an Equation (E3):

$$\boldsymbol{E}_3 = \boldsymbol{dis}(\boldsymbol{f}_{\boldsymbol{n}}^{\boldsymbol{k}}(\boldsymbol{S}^g \in \mathbb{R}^{\boldsymbol{n}}), \boldsymbol{f}_{\boldsymbol{n}}^{\boldsymbol{k}}(\boldsymbol{S}^s \in \mathbb{R}^{\boldsymbol{n}})). \tag{E3}$$

In addition, the *dis* function can also be a pairwise comparison measurement for continuous variables. A correlation matrix hence can be used to compare ground truth and synthetic data. Let the matrix $M^g = [br_{ij}^g]$, where br_{ij} is the correlation between ith and jth variables in the ground truth. As a result, given the calculated correlation matrix $M^s = [br_{ij}^s]$ for the synthetic dataset. The distance function can be also defined as Equation (E4)

$$\boldsymbol{E}_4 = \boldsymbol{dis}\left(\boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{g}}, \boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{s}}\right) = \begin{cases} \mathbf{1} & \boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{s}} \neq \boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{g}} \\ \mathbf{0} & \boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{s}} \approx \boldsymbol{br}_{\boldsymbol{ij}}^{\boldsymbol{g}} \end{cases}. \tag{E4}$$

The researchers need to define ≈ (almost equal) or ≠ (not equal), in the form of some threshold value. Here is an example to showcase how $\boldsymbol{E}_4$ can be used:

Given two variables A and B, where the correlation coefficient between them in ground truth data is 0.5, and in the synthetic data is 0.4. The difference between the correlations is 0.1. If a researcher defines the correlation difference range to be [0, 0.3] to assume almost equal correlations and beyond this range to be considered unequal, then for A and B the distance will be 0 according to Equation (E4) because 0.1 falls into the range of [0, 0.3].

3.4.3 | Privacy preservation

Privacy preservation is needed when both of the following conditions are met at the same time:

1. both datasets (ground truth and synthetic dataset) have some identical data instances and
2. some or all these data instances are "rare" in the ground truth, that is, outliers.

The sensitive variables or their possible combinations can be identified in the ground truth selection process from the input spaces. For a set of sensitive variables in the synthetic dataset, $S_k^s = \{s_1, \dots, s_k\} \in \mathbb{R}^k$, where $1 \le k \le$ total number of variables for a data instance. The distance function here is to calculate the distances among targeting data instances of sensitive variable(s) s_n to the other $\widehat{s_n}$ in space $\mathbb{R}^k$; the ones with larger distance to the rest groups of the data instances are viewed as outliers using Equation (E5).

$$\boldsymbol{E}_5 = \boldsymbol{dis}(\boldsymbol{s}_{\boldsymbol{n}}, \widehat{\boldsymbol{s}_{\boldsymbol{n}}}), \text{where } \mathbf{1} \le \boldsymbol{n} \le \boldsymbol{k},\ \ \boldsymbol{s}_{\boldsymbol{i}}\ \mathbf{U}\ \ \widehat{\boldsymbol{s}_{\boldsymbol{i}}} = \boldsymbol{S}_{\boldsymbol{k}}^{\boldsymbol{s}}. \tag{E5}$$

Density-based clustering approaches, such as DBSCAN proposed by Martin et al.,[35] can be used in this test. Once the outliers from synthetic data and ground truth are identified, then an exhaustive comparison among these outliers can be carried out to test if same data instances exist. If no identical instances are found, then we consider the privacy to be preserved.

3.4.4 | Clinical evaluation

One of the coauthors (P.M.) has a clinical background and reviewed the graph structures representing the relationships learned to assess whether the learned relationships between variables were recognized in medical research or if were clinically plausible. We undertook a further clinical evaluation test whereby two independent medical assessors reviewed a sample ($n = 100$) containing randomly selected records for equal number of synthetic and real patients. The assessors were blinded to which records related to synthetic data patients and which related to real patients. Assessors were asked to select at least 20 patient records from this test dataset and categorize them as synthetic or real based on the clinical characteristics. Assessors were free to either use a random selection approach or select patients that stood out because of their clinical characteristics. The research team then calculated the percentage of total records, synthetic patient records and real patient that were correctly classified by each assessor.

3.5 | Sensible synthetic data selection process

Based on the evaluation methods, we here define a *sensible synthetic dataset(s)* as that which meets all the evaluation criteria, that is, univariate distance comparison, multivariate distance comparison, and privacy preservation. In order to achieve this, the following algorithm (see Table 2) of a set of sequential and recursive steps is provided to enable this decision process given the input of synthetic and ground truth datasets.

The whole selection process is followed by the synthetic data generation process with a set of predefined global variables:

SynGen = the synthetic data generation process that triggers the sensible synthetic data selection process.

S = the output of SynGen, a list of generated synthetic datasets, total number ≥ 1.

S' = the ground truth, a single dataset.

S^c = the list of sensible synthetic data candidates, empty when initialized.

N = total number of variables in S'.

4 | CASE STUDIES

Two case studies demonstrate the use of the proposed framework. Two different synthetic dataset generation methods, that is, Copula and Bayesian networks, were applied followed by the synthetic data selection processes based on the data generation methods. The potential benefits of using synthetic data are also demonstrated which include data size enlargement, and performance when used as alternative to ground truth in AI training, testing and prediction applications.

4.1 | Case study I: Indian liver patient dataset

The Indian liver patient dataset (ILPD) is a dataset that is open to be downloaded[†] and studied in different researches such as in Venkata et al.'s work.[36] While this is not sourced from EHR data, the data representation reflects an EHR-sourced cross-sectional dataset. The purpose of this case here is to demonstrate how the proposed synthetic data framework can be applied to this dataset

TABLE 2 Algorithm to select the sensible synthetic dataset

```
Step 1. Initialisation
    SET $S$ = the list of generated synthetic datasets, total number $\geq 1$.
    SET $S'$ = the ground truth, a single dataset.
    SET $S^c$ = the list to hold sensible synthetic data candidates, initialise to an empty list.
    SET $N$ = total number of variables in $S'$.

Step 2. Univariate comparison
    SET C1 = minimum distance condition for discrete variables, i.e. probability difference, a number.
    SET C2 = minimum distance condition for continuous variables, i.e. p-value threshold, a number.
    SET C = total number of variables meeting distance conditions, initialise a value to 0.
    FOR EACH $s \in S$:
      SET $sim_{un}$ = outputs of E1 for a discrete variable and E2 for a continuous variable.
      FOR EACH sim $\in sim_{un}$:
        IF sim is calculated from a discrete variable and sim $\leq$ C1
          C = C+1.
        ELSE
        IF sim is calculated from a continuous variable and sim $\geq$ C2 (it could be sim $\leq$ C2 upon
        hypothesis test)
          C = C+1.
        END IF
      END FOR
      IF C = N
        Add s to $S^c$.
      END IF
      C = 0
    END FOR
    IF $S^c$ is empty
      CALL: SynGen
    ELSE
      RETURN $S^c$
    END IF
Step 3. Multivariate distance comparison
    SET C3 = minimum distance condition, a number upon a dimension reduction method, e.g. 0.2 for
    NMDS.
    SET C4 = minimum distance condition using correlation condition, a threshold value $\in [-2,2]$.
    SET C = total number of variables meeting distance conditions using E4, initialise a value to 0.
    FOR EACH $s \in S^c$:
      SET $sim_{mu}$ = outputs of E3 or E4.
      IF $sim_{mu}$ > C3 given using E3
        Remove $s$ from $S^c$
      ELSE
      IF using E4
        FOR EACH sim $\in sim_{un}$:
          IF sim = 0
          C = C+1
          END IF
        END FOR
        IF C! = N
          Remove $s$ from $S^c$
        END IF
        C = 0
```

(Continues)

TABLE 2 (Continued)

```
            END IF
        END FOR
        IF $S^c$ is empty
            CALL: SynGen
        ELSE
            RETURN $S^c$
        END IF
Step 4. Privacy preservation
    SET $p_r^k$ = outliers found from each set from $S^c$ , initialise to an empty list.
    SET $P_r^{k'}$ = outliers found from $S'$
    FOR EACH $s \in S^c$:
        IF s $\in (P_r^{k'} \cap p_r^k)$
            Remove $s$ from $S^c$
        END IF
    END FOR
    IF $S^c$ is empty
        CALL: SynGen
    ELSE
        RETURN $S^c$
    END IF
Step 5. Output
    RETURN $S^c$
```

for (1) generating synthetic data in which no ground truth should appear in these synthetic data and (2) for showing how synthetic data can help scale data when there is a limited amount of data and the impacts of such scaling on research.

4.1.1 | Ground truth description

This dataset contains of a total of 583 people with 416 liver patient records and 167 nonliver patient records. The dataset was collected from north east of Andhra Pradesh, India and it contains 441 male patient records and 142 female patient records. Table 3 summarizes the ILPD variables and their statistics.

4.1.2 | Using Copula to generate synthetic data

Copula is a multivariate cumulative distribution function that can be used to understand the dependency structure among different variables and as a result is widely used to model multivariate datasets when the underlying dependency is essential such as in Sklar's work[37] and Kao et al.'s work.[38] Since copula by default models correlation structure and the marginals/distributions for each variable, part of the evaluation of the proposed synthetic data framework, that is, univariate distance test (see Table 4), multivariate distance (see Figure 2) test using pairwise correlation, can be done during the data generation process. In this case, the t-Copula[39] is used to model the data. Table 4 lists the approximated univariate distributions based on the skewness and kurtosis estimated by bootstrap. Figure 3 shows the generated copulas visual representations based on the ground truth single variable distributions and multivariate correlations.

TABLE 3 Variables used in ILPD study with summary statistics; the last variable is the dependent variable

#	Variables (acronyms)	Variable type (available values)	Summary statistics
1	Age (age)	Numeric	Mean (SD): 44.7 (16.2) Min < med < max: 4 < 45 < 90
2	Gender (gender)	Categorical (0: female or 1: male)	0:24.4% 1:75.6%
3	Total bilirubin (tb)	Numeric (1 decimal place)	Mean (SD): 3.3 (6.2) Min < med < max: 0.4 < 1 < 75
4	Direct bilirubin (db)	Numeric (1 decimal place)	Mean (SD): 1.5 (2.8) Min < med < max: 0.1 < 0.3 < 19.7
5	Alkaline phosphatase (alp)	Numeric (0 decimal place)	Mean (SD): 290.6 (242.9) Min < med < max: 63 < 208 < 2110
6	Alanine aminotransferase (alt)	Numeric (0 decimal place)	Mean (SD): 80.7 (182.6) Min < med < max: 10 < 35 < 2000
7	Aspartate aminotransferase (asat)	Numeric (0 decimal place)	Mean (SD): 109.9 (288.9) Min < med < max: 10 < 42 < 4929
8	Total proteins (tp)	Numeric (1 decimal place)	Mean (SD): 6.5 (1.1) Min < med < max: 2.7 < 6.6 < 9.6
9	Albumin (alb)	Numeric (1 decimal place)	Mean (SD): 3.1 (0.8) Min < med < max: 0.9 < 3.1 < 5.5
10	Albumin/globulin ratio (a/g)	Numeric (2 decimal places)	Mean (SD): 0.9 (0.3) Min < med < max: 0.3 < 0.9 < 2.8
11	Liver patient (lp)	Categorical (0: liver patient or 1: not liver patient)	0:71.4% 1:28.6%

4.1.3 | Evaluation and sensible synthetic dataset selection

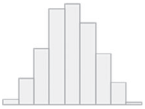
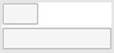
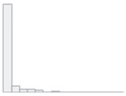
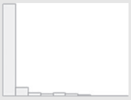
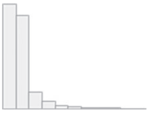
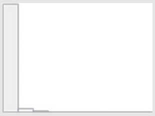
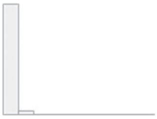
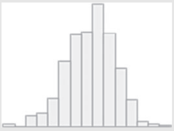
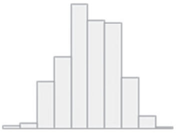
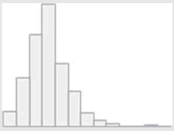
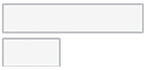
The Copula-based approach is based on univariate distribution (see Table 4) and multivariate correlation (see Figure *2*) obtained from ground truth and therefore, in this case the sensible synthetic data selection stage will be mainly focused on ensuring there is no duplication between synthetic data and ground truth. Figure 4 illustrates the generated synthetic ILPD pairwise variable distributions using the copulas generated in Figure 3 versus ground truth pairwise variable distributions, where no duplicated record exists between both datasets.

In addition to the pairwise comparison, Nonmetric Multidimensional Scaling (NMDS) is also used here to explore how data instances are clustered in the multidimensional space. Two datasets with similar clusters are mapped to a 2-dimensional space (stress = 0.12) as shown in Figure 5. The similarity value obtained from an analysis of similarity (ANOSIM) ($p = .001$) further suggests that there is no significant difference between groups.

4.1.4 | Synthetic dataset applications

Two experimental tasks are done here to demonstrate (1) synthetic data can help increase the data size when more data are required.(2) Synthetic data can have similar performance in predicting task as ground truth.

TABLE 4 ILPD variable distribution modeling with the estimated distribution and graphical representations

#	Variables (acronyms)	Estimated distribution	Graphical representation
1	Age (age)	Normal (mean: 44.7, SD: 16.2)	
2	Gender (gender)	Binomial (size = 1, prob = 0.24)	
3	Total bilirubin (tb)	Negative binomial (size = 0.30, mean = 3.30)	
4	Direct bilirubin (db)	Gamma (shape = 0.28, rate = 0.189)	
5	Alkaline phosphatase (alp)	Gamma (shape = 1.43, rate = 0.004)	
6	Alanine amino-transferase (alt)	Gamma (shape = 0.20, rate = 0.002)	
7	Aspartate aminotrans-ferase (asat)	Negative binomial (size = 0.13, mean = 109.94)	
8	Total proteins (tp)	Normal (mean: 6.5, SD: 1.1)	
9	Albumin (alb)	Normal (mean: 3.1, SD: 0.8)	
10	Ratio albumin and globulin ratio (a/g)	Gamma (shape = 8.24, rate = 8.76)	
11	Liver disease patient (lp)	Binomial (size = 1, prob = 0.71)	

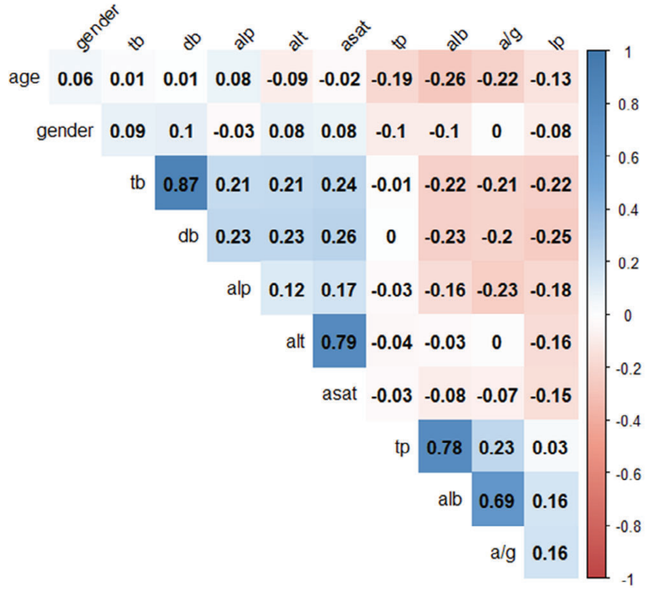


FIGURE 2 ILPD (583 people) ground truth Spearman's correlation coefficients [Color figure can be viewed at wileyonlinelibrary.com]

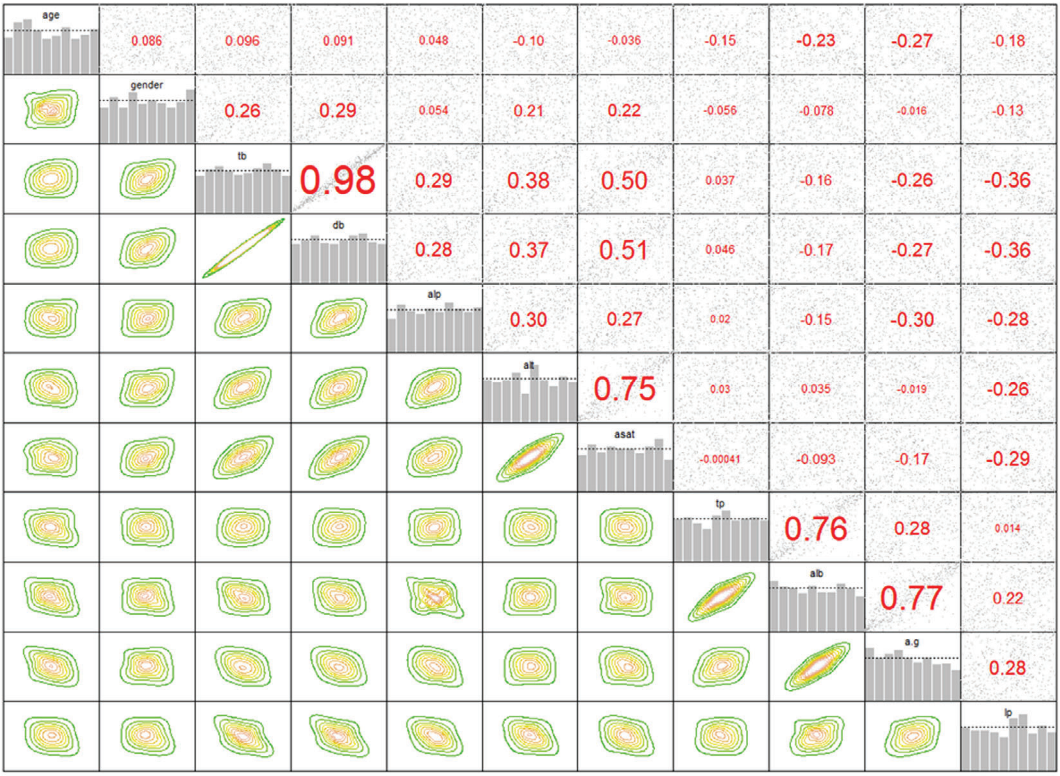


FIGURE 3 Generated ILPD (583 people) copulas visual representations with Spearman's correlation coefficients [Color figure can be viewed at wileyonlinelibrary.com]

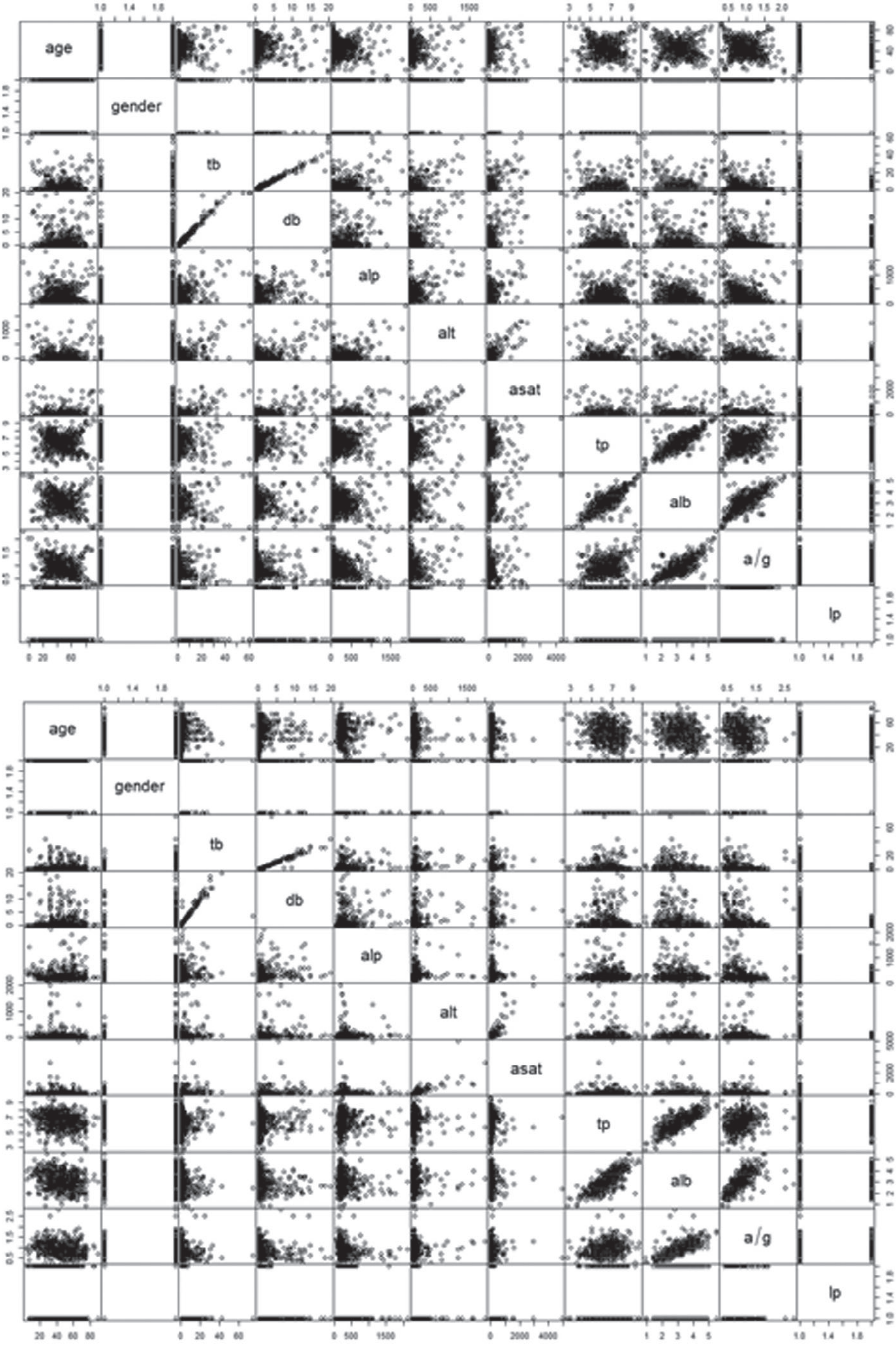


FIGURE 4 Synthetic (top) versus ground truth (bottom) ILPD (583 people) pairwise variables distributions visual representations

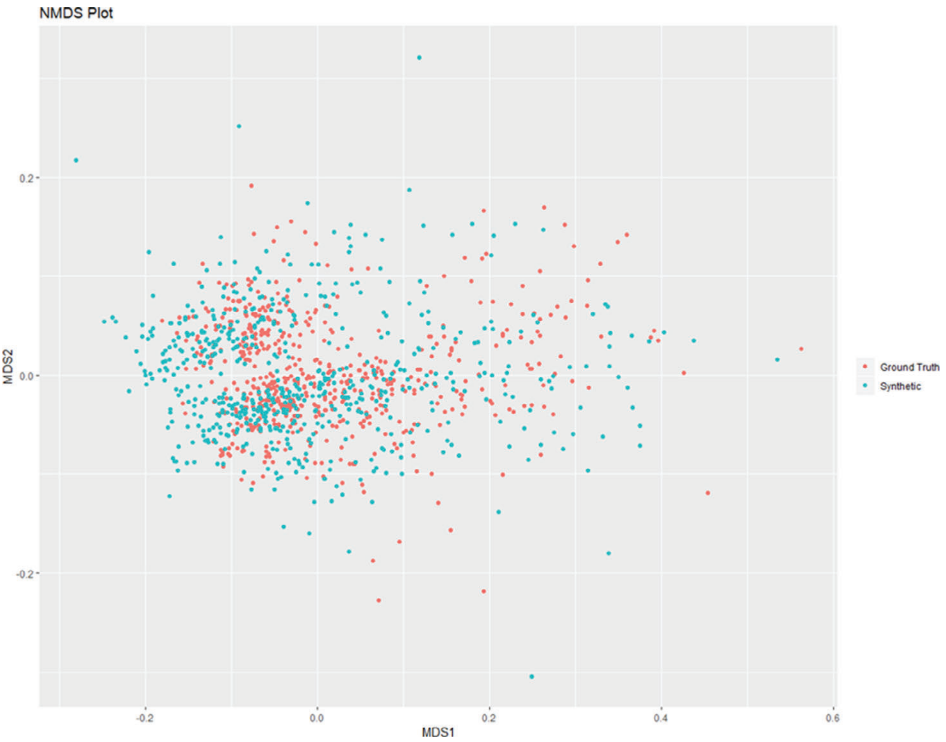


FIGURE 5 Representation of datasets in 2-dimension space using NMDS [Color figure can be viewed at wileyonlinelibrary.com]

Scaling up the data size

This task is to enlarge population size by 10 times, that is, from 583 people to 5830 people without duplicated results between two datasets. Figure 6 shows the generated copulas visual representations, which can be compared with copulas with 583 people in Figure 3, and also Figure 2 in terms of correlation coefficients. Among the generated 5830 records, 4033 people are liver patients and 1797 people do not have liver disease. The results clearly show that the correlational direction between variables is well preserved despite the scaled data size.

Predicting liver disease

Prediction is one of the most common objectives for many researches. Here, the task is to predict if a patient has liver disease, so the outcome variable will be lp (see Table 3).

The first experiment will use the multiple liner regression to formulate a linear relationship between predictor variables and outcome variable. The result from using ground truth of original 583 patients is compared with the result from using 5830 synthetic patients. Table 5 shows the coefficient of determination R^2 for both datasets which indicate that in both cases, the multiple linear regression does not fit the data well as R^2 values are close to 0.

The second experiment uses linear discriminant analysis (LDA)[40] with a leave-one-out cross-validation (LOOCV) to predict liver patients by distinguishing people with liver disease from healthy people. Tables 6 and 7 show the confusion matrices for ground truth and synthetic data, respectively. From the tables, the accuracy of the prediction by applying the same LDA can

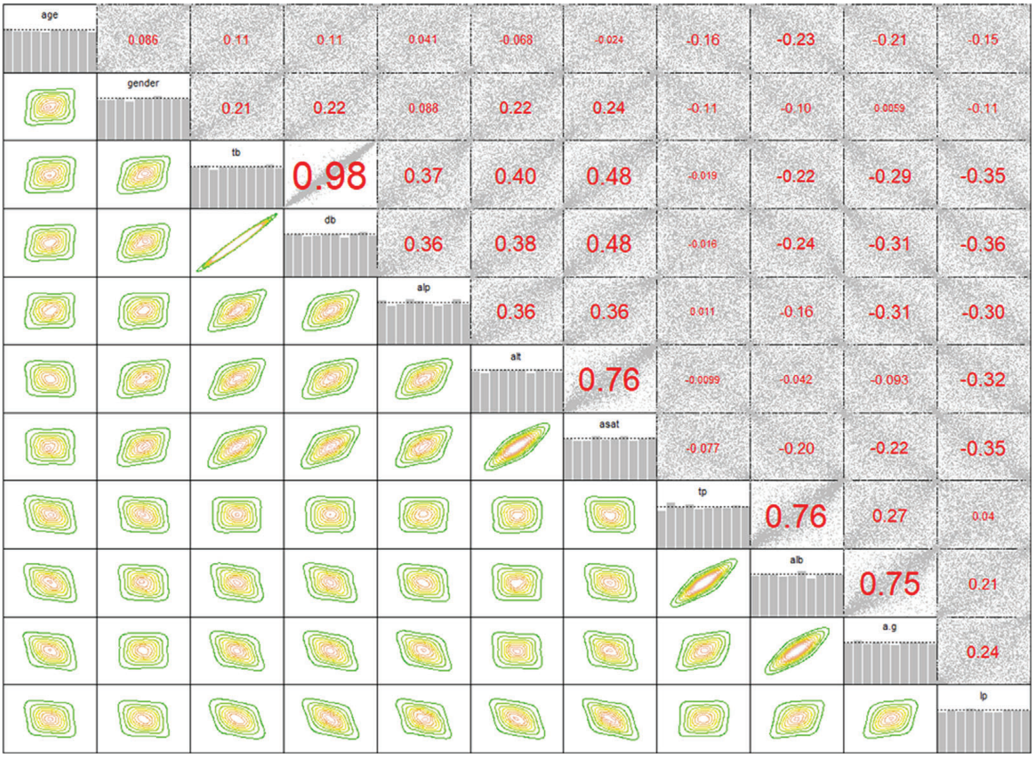


FIGURE 6 Generated ILPD (5830 patients) copulas visual representations with Spearman's correlation coefficients [Color figure can be viewed at wileyonlinelibrary.com]

Dataset	R^2
Ground truth (583 people)	0.12
Synthetic (5830 people)	0.11

TABLE 5 The coefficient of determination values for both ground truth and synthetic data using multiple linear regression

	Predicted (percent)	
Actual	**Yes**	**No**
Yes	398 (0.957)	18 (0.043)
No	154 (0.922)	13 (0.078)

TABLE 6 Confusion matrix of ILPD ground truth

be obtained, which is 71.51% for synthetic data and 70.32% for ground truth. As a result, the performance of using two datasets is very similar.

4.2 | Case study II: Cardiovascular disease prediction

A common epidemiological research area concerns cardiovascular disease (CVD). Here, for the first time, we use 57,397 patient instances to demonstrate how the proposed framework can be

TABLE 7 Confusion matrix of ILPD synthetic data

	Predicted (percent)	
Actual	**Yes**	**No**
Yes	3874 (0.961)	159 (0.039)
No	1502 (0.836)	295 (0.164)

used to generate and evaluate the synthetic data for a CVD study from a longitudinal clinical dataset sourced from the Clinical Practice Research Datalink (CPRD) Aurum database. This case study represents a typical scenario where healthcare data access is restricted. CPRD Aurum is a UK primary care database covering over 20% of the UK population as of May 2020 and includes 17,400 clinical event types across patients with 25% of the patient data tracing back at least 20 years. A detailed data specification can be found online on the CPRD website‡ and in the published data resource profile.[41]

4.2.1 | Ground truth selection and description

This case study focuses on cardiovascular risk factors which are well characterized in the clinical research literature,[30,42,43] widely applied in clinical practice§ as well as supported by clinical guidelines. The variables used are listed in Table 8, the first 22 variables are defined as independent/predictor variables and the last variable, that is, stroke/heart attack is the dependent/outcome variable. We adopted variable definitions and code lists as described in the literature[30] with validation by a clinical expert.

For demonstration purposes, all variables are predefined as privacy sensitive variables because these variables in combination are thought to be more likely to identify a patient than when only a few of them are used. For this case study we only included patients over 16 years to reflect the typical epidemiological profile of cardiovascular disease risk.

4.2.2 | Variable selection process

The generation of the ground truth dataset goes through three procedures, that is, (1) medical code mapping, (2) cohort selection and clinical records extraction, and (3) longitudinal data to cross-sectional data conversion.

In the medical code mapping procedure, clinical categorical variables (exclude ethnicity, gender, and region) in Table 8 are mapped to a set of medical codes in CPRD Aurum. Table 9 demonstrates the medical codes mapped to the chronic kidney disease variable.

In the cohort selection and clinical records extraction, the cohort is randomly sampled from the population. A total of 57,397 data instances (patients) are sampled, and the summary statistics are described in Table 8. As one of the purposes of this study is to showcase an application of using AI algorithms to predict the CVD risk within 10 years' time, the baseline for each patient starts from 10 years before his/her first stroke/heart attack incident, or 10 years before October 1, 2018, that is, October 1, 2008, with no prior stroke/heart attack incident before this date throughout the clinical record.

TABLE 8 Variables used in CVD study with their summary statistics; the last variable is the dependent variable

#	Variables (acronyms)	Variable type (available values)	Summary statistics
1	Age (age)	Numeric	Mean (SD): 71.7 (14.3) Min < med < max: 26 < 73 < 109
2	Ethnicity (ethr)	Categorical	White or not stated: 28.4% Indian: 1.5% Pakistani: 0.6% Bangladeshi: 0.2% Other Asian: 0.9% Black Caribbean: 0.7% Black African: 0.3% Chinese: 0.1% Other: 67.2%
3	Gender (gender)	Categorical (male or female)	FEMALE: 51.4% MALE: 48.6%
4	BMI (bmi)	Numeric (1 decimal place)	Mean (SD): 28.1 (6.1) Min < med < max: 8.3 < 27.2 < 149.7
5	Cholesterol/HDL ratio (choleratio)	Numeric (1 decimal place)	Mean (SD): 3.8 (1.2) Min < med < max: 0.1 < 3.6 < 9.9
6	Systolic blood pressure (sbp)	Numeric (0 decimal place)	Mean (SD): 133.9 (16.4) Min < med < max: 62 < 134 < 240
7	Systolic blood pressure SD (sbps)	Numeric (2 decimal places)	Mean (SD): 14 (6.5) Min < med < max: 0.4 < 13.3 < 70
8	Family history of CHD (fh_cad)	Categorical (TRUE or FALSE)	TRUE: 3.0% FALSE: 97%
9	Smoking status (smoking)	Categorical	Nonsmoker/unknown: 67.7% Ex-smoker: 21.9% Light smoker: 4.4% Moderate smoker: 3.9% Heavy smoker: 2.1%
10	On hypertension treatment (treathyp)	Categorical (TRUE or FALSE)	TRUE: 6.1% FALSE: 93.9%
11	Chronic kidney disease (ckidney)	Categorical (TRUE or FALSE)	TRUE: 11.1% FALSE: 88.9%
12	Rheumatoid arthritis (ra)	Categorical (TRUE or FALSE)	TRUE: 5.7% FALSE: 94.3%
13	Atrial fibrillation (af)	Categorical (TRUE or FALSE)	TRUE: 6.3% FALSE: 93.7%
14	On atypical antipsychotic medication (atyantip)	Categorical (TRUE or FALSE)	TRUE: 0.3% FALSE: 99.7%
15	Migraines (migr)	Categorical (TRUE or FALSE)	TRUE: 5.7% FALSE: 94.3%

(Continues)

TABLE 8 (Continued)

#	Variables (acronyms)	Variable type (available values)	Summary statistics
16	Systemic lupus erythematosus (sle)	Categorical (TRUE or FALSE)	TRUE: 0.1% FALSE: 99.9%
17	Severe mental illness (semi)	Categorical (TRUE or FALSE)	TRUE: 10.3% FALSE: 89.7%
18	Type 2 diabetes (type2)	Categorical (TRUE or FALSE)	TRUE: 15.3% FALSE: 84.7%
19	Type 1 diabetes (type1)	Categorical (TRUE or FALSE)	TRUE: 1.4% FALSE: 98.6%
20	On regular steroid tablet (steroid)	Categorical (TRUE or FALSE)	TRUE: 2.6% FALSE: 97.4%
21	Region (region)	Categorical	London: 9.9% East of England: 3.4% North East: 7.4% North West: 20.2% Yorkshire and the Humber: 4.5% East Midlands: 3.0% South Central: 15.6% South West: 8.3% West Midlands: 14.4% South East Coast: 13.3%
22	Erectile dysfunction (impot)	Categorical (TRUE or FALSE)	TRUE: 4.7% FALSE: 95.3%
23	Stroke/heart attack (strokeha)	Categorical (TRUE or FALSE)	TRUE: 17.5% FALSE: 82.5%

When converting the longitudinal data to cross-sectional data, the extracted clinical records are "rolled up" for each variable so that for each patient's 10 years of clinical data can be summarized as a single record of the variables/columns defined in Table 8. For the variable "age," each patient's baseline age was used, and an average value was used for variables "choleratio," "sbp," and "sbps." For the rest of the variables, each earliest value was used.

4.2.3 | Using Bayesian networks to generate synthetic data

Bayesian networks (BNs) are a type of probabilistic graphical model that models the joint distribution of a domain using a directed acyclic graph structure and local conditional distributions associated with each node. Inference algorithms can be used to learn these graphical structures and to predict variable values based on their posterior distributions given some evidence. In the clinical study context, BNs are often used to discover new relationships among multiple factors such as in Fuster-Parra et al.'s work[33] and to identify key factors that contribute to certain outcomes in CVD studies as found by Multani et al..[44]

Figure 7 shows the inferred BN structure from the CVD data in the form of a directed acyclic graph (DAG) between variables.

Medical code	Clinical term description
2773184015	Chronic kidney disease stage 3
304071000000115	Chronic kidney disease stage 3
304091000000116	Chronic kidney disease stage 4
304111000000114	Chronic kidney disease stage 5
557811000000119	Chronic kidney disease stage 3A
557831000000110	Chronic kidney disease stage 3B
595811000000117	Chronic kidney disease stage 3 with proteinuria
595871000000110	Chronic kidney disease stage 3 without proteinuria
595931000000116	Chronic kidney disease stage 3A with proteinuria
595991000000115	Chronic kidney disease stage 3A without proteinuria
596131000000114	Chronic kidney disease stage 3B without proteinuria
596261000000111	Chronic kidney disease stage 4 without proteinuria

TABLE 9 Chronic kidney disease medical code mapping

The accompanying conditional probabilities for each interlinked variable are also learned which serve the basis for making inference based on evidence and sampling new data, that is, synthetic data from BNs learnt from the ground truth. Taking the smoking status for example, a probability table for white male patients can be learned (see Table 10). Synthetic data can be generated via random sampling from the learned Bayesian network.

4.2.4 | Evaluation and sensible synthetic dataset selection

The evaluation starts by sampling the same number of patients from both synthetic data and ground truth. In this case study, we considered a sample size of 800 with one set of ground truth of 800 patients, and five sets of synthetic datasets (800 sample in each set) as the candidate sensible synthetic datasets. The following sections report the analysis processes of one of the candidate datasets that best meets all our criteria, that is, (1) high similarity in terms of results of univariate and multivariate analyses; (2) no duplicate records in ground truth and synthetic data in terms of outliers in both datasets.

Univariate distance analysis

For continuous type variables, KS test is used (see Table 11). For discrete type variables, the probabilities of available values are compared, and a probability difference smaller than 2% is acceptable in this case (see Table 12).

Multivariate distance analysis

A correlation analysis is initially used to determine if the generated synthetic data captures the interrelationships as would be expected in the ground truth. Figure 8 shows the test that can confirm that one of the synthetic datasets can capture the significant correlations from ground truth. The correlation between BMI and weight in this case is the only pair that is significantly correlated, that is, 0.82 for synthetic data and 0.86 for ground truth.

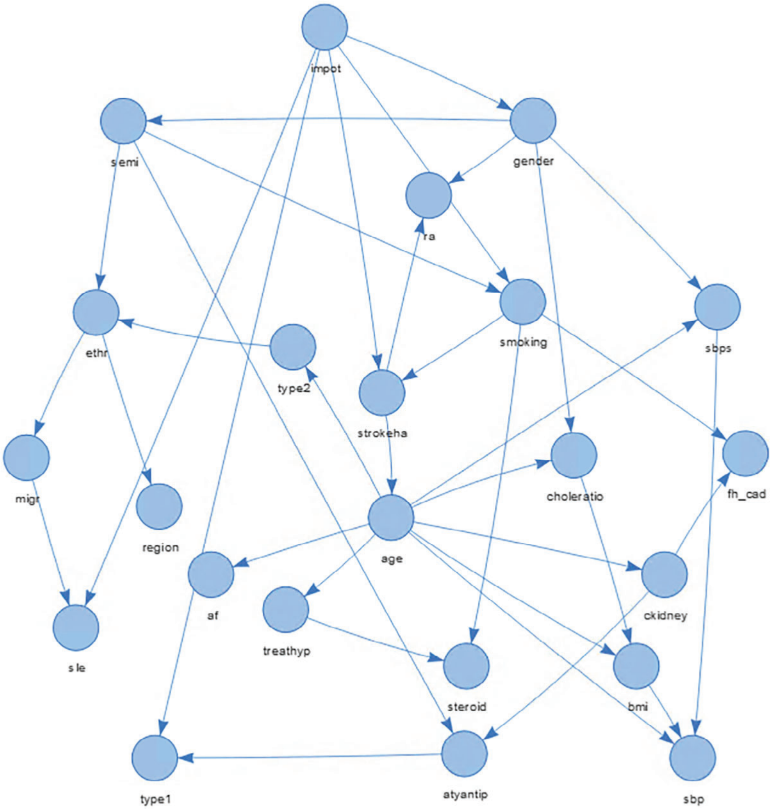


FIGURE 7 Inferred DAG of internal relationships between CVD variables (annotated by acronyms from Table 8) using BN structure learning [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 10 Learned male and female smoking status probability TABLE

ID	Smoking status	Male	Female
1	Nonsmoker/unknown	0.30	0.36
2	Ex-smoker	0.28	0.16
3	Heavy smoker	0.04	0.02
4	Light smoker	0.04	0.04
5	Moderate smoker	0.04	0.04

Like the previous case study with NMDS, data are mapped to a 2-dimensional space (stress = 0.18) as shown in Figure 9. Both datasets illustrate similarities in terms of clusters and distribution such as a clear split of population horizontally in lower data dimension. The similarity value from ANOSIM is $p = .001$, which suggests that there is no significant difference between groups.

Outlier detection and comparison

Instead of doing an exhaustive pairwise comparison, the outlier approach can be an alternative approach of eliminating duplicated outlier records which are sometimes important for a clinical study yet with high privacy risks. Here, the DBSCAN[35] is used to identify the outliers in the dataset, 36 outliers are identified from ground truth, and 21 outliers from synthetic data, respectively (see Figure 10).

TABLE 11 KS test for continuous variables

Variable	*p*-Value
Age	.74
BMI	.65
Cholesterol/HDL ratio	.59
Systolic blood pressure (sbp)	.72
Systolic blood pressure SD (sbps)	.55

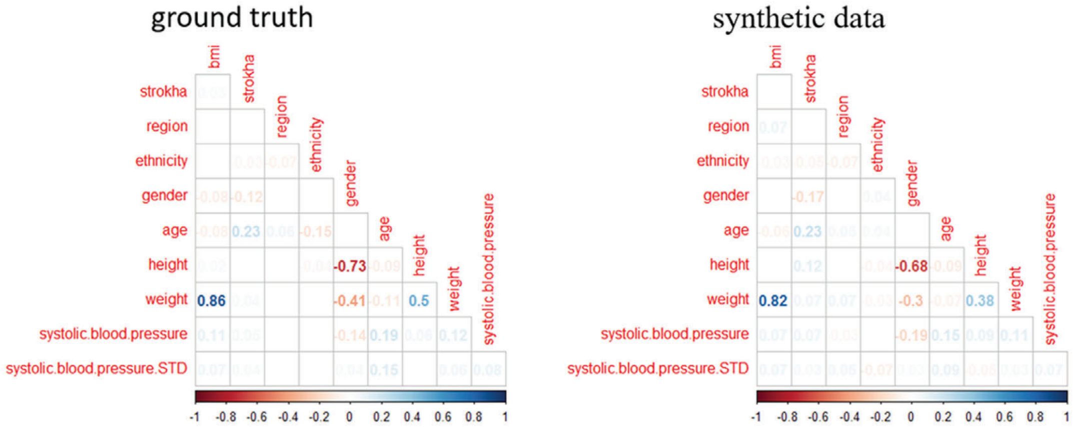


FIGURE 8 The partial correlation matrix for both ground truth (left) and synthetic data (right), all other insignificant correlations are removed ($P > .05$) [Color figure can be viewed at wileyonlinelibrary.com]

Clinical validation

The initial clinical evaluation of the BN learned variable relationships found that almost all relationships were well recognized in the medical research and where this was not the case, there was a possible explanation for the observed relationships in the ground truth data. Expert 1 initially reviewed a random sample of 20 records and then chose another four records for patients who had the outcome of interest (heart attack or stroke). Expert 2 reviewed a random sample of 25 records and another 19 records for patients who had the outcome of interest (heart attack or stroke). The overall results indicate that while both experts could correctly identify at least 50% of the sampled records (62.5% for expert 1 and 50% for expert 2) as being either synthetic or real, the accuracy was markedly different for real records (92.8% for expert 1 and 84.2% for expert 2) and synthetic records (20% for expert 1 and 6.7% for expert 2). Table 13 summarizes the clinical evaluation results.

4.2.5 | Synthetic dataset applications

Two experiments are carried out in the context of common research activities to demonstrate (1) synthetic data can be an alternative to ground truth when data access is restricted and (2) synthetic data can be a good alternative to machine learning training/testing dataset.

TABLE 12 Probability differences via probability of ground truth − probability of generated synthetic data

Variable	Probability difference (ground truth - synthetic)
Ethnicity (ethr)	White or not stated: 1.2% Indian: 0.8% Pakistani: 0.2% Bangladeshi: −0.1% Other Asian: 0.2% Black Caribbean: −0.2% Black African: 0.1% Chinese: 0% Other: −2.2%
Gender (gender)	TRUE: 1.4% FALSE: −1.4%
Family history of CHD (fh_cad)	TRUE: −0.4% FALSE: 0.4%
Smoking status (smoking)	Nonsmoker/unknown: −1.2% Ex-smoker: 0.8% Light smoker: 0.2% Moderate smoker: 0% Heavy smoker: 0.2%
On hypertension treatment (treathyp)	TRUE: 0.02% FALSE: −0.02%
Chronic kidney disease (ckidney)	TRUE: 1.1% FALSE: −1.1%
Rheumatoid arthritis (ra)	TRUE: 1.3% FALSE: −1.3%
Atrial fibrillation (af)	TRUE: 0.4% FALSE: −0.4%
On atypical antipsychotic medication (atyantip)	TRUE: 0.5% FALSE: −0.5%
Migraines (migr)	TRUE: 1.4% FALSE: −1.4%
Systemic lupus erythematosus (sle)	TRUE: 1.1% FALSE: −1.1%
Severe mental illness (semi)	TRUE: 0.3% FALSE: −0.3%
Type 2 diabetes (type2)	TRUE: 1.5% FALSE: −1.5%
Type 1 diabetes (type1)	TRUE: 1.7% FALSE: −1.7%
On regular steroid tablet (steroid)	TRUE: 0.9% FALSE: −0.9%

(Continues)

TABLE 12 (Continued)

Variable	Probability difference (ground truth - synthetic)
Region (region)	London: −0.4% East of England: −0.2% North East: −0.1% North West: 1.2% Yorkshire and the Humber: 0% East Midlands: −0.8% South Central: 1.3% South West: 0.6% West Midlands: −1.2% South East Coast: −0.4%
Erectile dysfunction (impot)	TRUE: 1.7% FALSE: −1.7%
Stroke/heart attack (strokeha)	TRUE: 1.5% FALSE: −1.5%

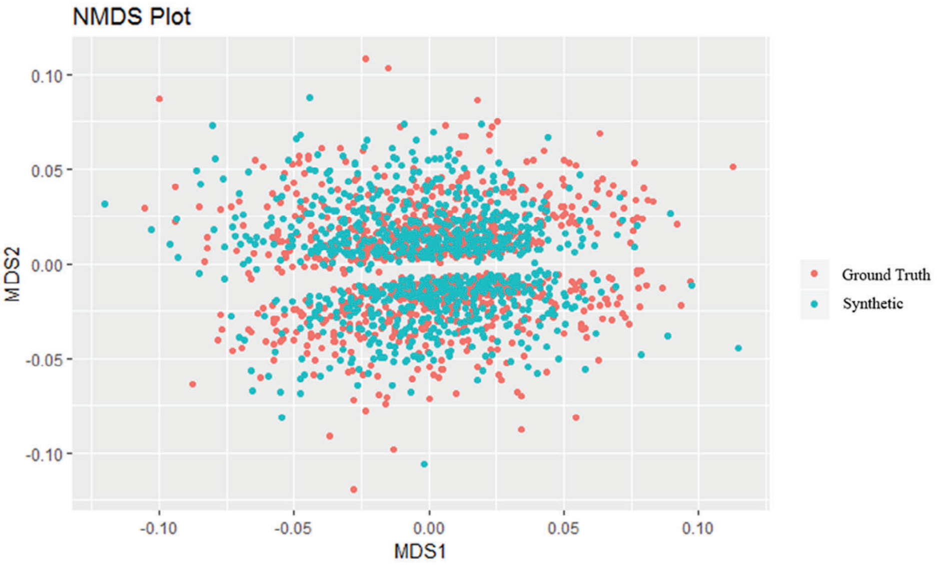


FIGURE 9 Representation of datasets in 2-dimension space using NMDS [Color figure can be viewed at wileyonlinelibrary.com]

Alternative to ground truth

The cardiovascular disease (CVD) risk calculator[30] is used to predict 10-year CVD risk based on the features described in Table 8. It is applied to both ground truth and synthetic datasets. Figure 11 shows the ROC curves for both ground truth and synthetic data. Figure 12 shows the synthetic data ROC curve, and the difference between two ROCs is trivial (p = .52077). In terms of the AUCs, there is 0.9% difference, that is, synthetic dataset is 0.9% higher (73.6% vs. 72.7%). The best fit threshold values are 24.4 for ground truth and 20.5 for synthetic dataset.

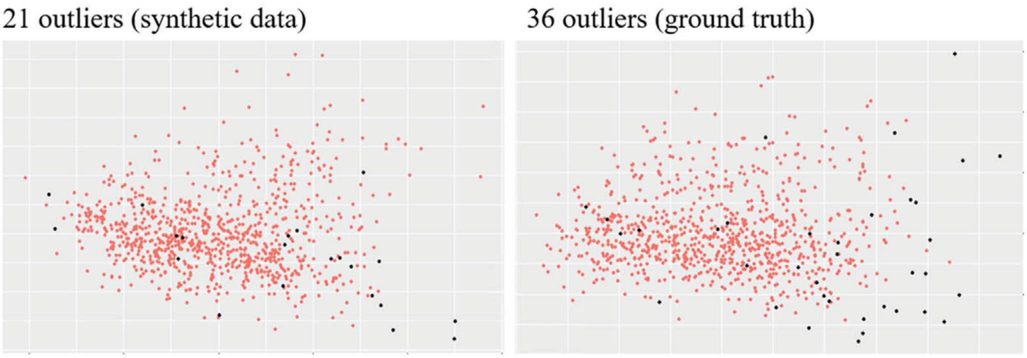


FIGURE 10 The detected outliers in both ground truth (36 data instances) and generated synthetic data (21 data instances). Darkest points are outliers [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 13 Clinical evaluation results

	Expert 1 results		Expert 2 results	
Type of records evaluated	Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
Total records	15/24 (62.5%)	9/24 (37.5%)	17/34 (50%)	17/34 (50%)
Synthetic records	2/10 (20.0%)	8/10 (80.0%)	1/15 (6.7%)	14/15 (93.3%)
Real records	13/14 (92.9%)	1/14 (7.1%)	16/19 (84.2%)	3/19 (15.8%)

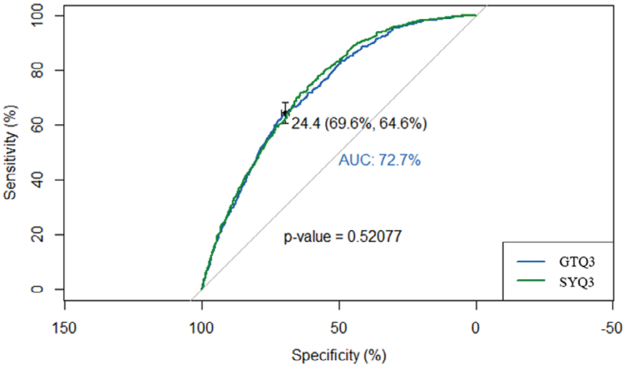


FIGURE 11 ROC curves of ground truth (GTQ3) and synthetic data (SYQ3) for CVD risk calculator. The Delong[45] method is used to compare the AUC of both GTML and SYML and the *p*-value is .52077. AUC of GTQ3 is 72.7% and the associated best fit threshold value is 24.4 [Color figure can be viewed at wileyonlinelibrary.com]

Machine learning algorithms training/testing

The ground truth and synthetic datasets are compared in terms of the results when applying same sets of machine learning algorithms to predict CVD risk. In this experiment, stacked ensembles[46] is used for the purpose of showcasing how a combination of machine learning can be applied in parallel to achieve similar results given this imbalanced clinical data, for example, the True-to-False rate of stroke/heart attack is <20% and type 1 diabetes is <2%. Six algorithms are included namely,

1. Least absolute shrinkage and selection operator (LASSO)[47]
2. Classification and regression training (CARET)[48]

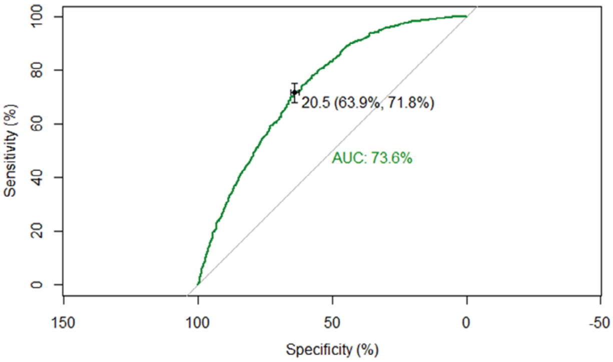


FIGURE 12 ROC curve for synthetic data. AUC is 73.6% and the best fit threshold value is 20.5 [Color figure can be viewed at wileyonlinelibrary.com]

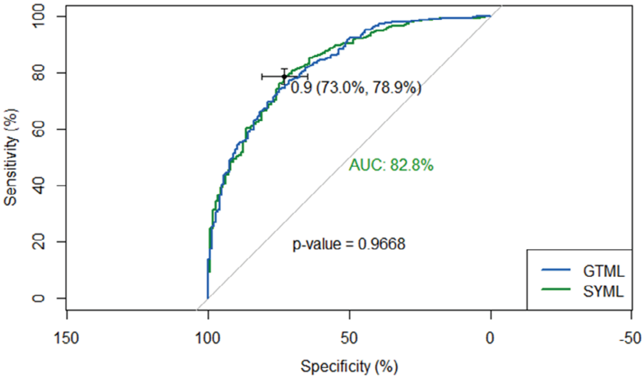


FIGURE 13 ROC curves of ground truth (GTML) and synthetic data (SYML) with the same set of machine learning algorithms. The Delong[45] method is used to compare the AUC of both GTML and SYML and the p-value is .9668. AUC of SYML is 82.8% and the associated best fit threshold value is 0.9 [Color figure can be viewed at wileyonlinelibrary.com]

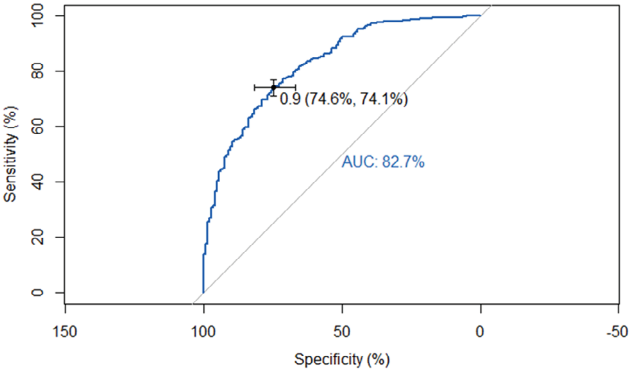


FIGURE 14 ROC curve for ground truth. AUC is 82.7% and the best fit threshold value is 0.9 [Color figure can be viewed at wileyonlinelibrary.com]

3. Extremely randomized trees[49]
4. Feed-forward neural networks[50]
5. Nonnegative least squares[51]
6. Random forest[52]

Figure 13 shows the ROC curves for both ground truth and synthetic data, while Figure 14 shows the ground truth ROC curve, and the difference between two ROCs is trivial (p = .9668). In terms of the AUCs, there is 0.1% difference, that is, synthetic dataset is 0.1% higher (82.8% vs. 82.7%).

Best fit threshold prediction value is at 0.9 for both datasets, we also can further obtain the following results.

1. For synthetic dataset—true positive rate (TPR) = 78.9% and false positive rate (FPR) = 27%
2. For ground truth—TPR = 74.1% and FPR = 25.4%

5 | CONCLUSIONS

A framework to generate and evaluate synthetic electronic healthcare data is proposed with the aim of balancing synthetic data utility and preserving patient privacy. Our definition of data utility is that it should be ultimately clinically meaningful.

5.1 | Outcomes

These were assessed via two case studies to demonstrate the framework's applicability to different scenarios and capability to deal with heterogeneity by:

1. using an open cross-sectional dataset and a licensed longitudinal database as ground truth data sources;
2. review of two different synthetic data generation methods including the copulas and inferred Bayesian structure and review of the generated synthetic data by clinical experts in the author team and project steering group;
3. comparison of (a) univariate and multivariate distance and (b) of performance of different algorithms including machine learning in the ground truth and synthetic datasets as presented in both studies.

For both studies, we included statistical tests as well as clinical tests with success in the latter being defined as "the inability to distinguish between the ground truth and synthetic data by a clinical expert." In our clinical evaluation we were able to demonstrate that while clinical experts were able to classify real patient records correctly with a high degree of accuracy, they tended to misclassify synthetic records as being real. These results strongly suggest that our synthetic data were able to reproduce clinically meaningful relationships.

The studies suggest that the proposed synthetic data generation approach allows for a high degree of fidelity between the synthetic and ground truth data, thus enabling the execution of complex machine learning algorithms in the synthetic data as if in the ground truth. The lifecycle of the synthetic data production from ground truth selection to synthetic data generation, evaluation, and sensible synthetic data output can also be fit into the specific data generation methods according to different privacy requirements.

5.2 | Discussion and limitations

We have demonstrated the applicability of our proposed synthetic data generation and evaluation framework using different healthcare datasets and scenarios. This framework is flexible enough to allow for different approaches to synthetic data generation while allowing researchers to demonstrate that they have balanced data utility with patient privacy needs. There has been recent interest in employing deep learning approaches of multilayer neural networks such as generative adversarial networks (GANs)[¶] for synthetic data generation with the belief that these may be well suited to capture complex features in the data. These approaches could be applied within the proposed framework; however, a unanimous voice from regulators and the public is that a more transparent approach is required.[53] This is partly because of the high sensitivity of the EHR data which is usually represented by high risks, it is associated with in case of privacy being compromised or wrong clinical outcomes. As a result, when choosing the data generation methods, we encourage the use of more transparent instead of "black box" type algorithms because of the extra overhead on interpretation model architectures and intermediary results. There are other challenges during synthetic data generation such as data missingness, complex interactions between variables and sensitivity analysis statistics from machine learning classifiers. These were addressed in our other work.[54]

We have not included data management in our proposal to reshape the longitudinal ground truth data from a long to a wide format, that is, converting the longitudinal data to cross-sectional data (a process we refer to as data "roll-up"). However, it should be noted that data management prior to synthetization means that the end users of the synthetic data, would have to rely on assumptions made by the team generating the synthetic data. Additional methodological works can be undertaken to explore whether the methods included in this work can be applied to longitudinal data source without the need for overt manual data management, while still preserving clinical meaningfulness.

5.3 | Outlook

In summary, among those benefits discussed at the beginning of this article, the authors believe synthetic data can be an effective alternative to real-world data in different cases due to its unique advantages: when access to the ground truth is restricted, for example, because of privacy issue, and when the real data size is not "big" enough or when lacking of machine learning/AI training/testing datasets. Our case studies demonstrated that using the proposed framework, synthetic healthcare data can be successfully generated for these scenarios. These results provide a successful proof of concept which can be extended to many other clinical research scenarios. In addition, the work here also provides a methodological template to encourage new insights on the use of synthetic data in the age of AI.

ACKNOWLEDGMENTS

This work is undertaken under the Regulator's Pioneer Fund programme sponsored by Department for Business, Energy and Industrial Strategy (BEIS) and supported by both MHRA and CPRD. The MHRA have been awarded a grant from the £10m Regulators' Pioneer Fund launched by BEIS and administered by Innovate UK. The fund enables UK regulators to develop innovation-enabling approaches to emerging technologies and unlock the long-term economic opportunities identified in the government's modern Industrial Strategy. The authors also thank Dr. Anu Jain (Medical Assessor, Licensing Division, MHRA) and Dr. Shirley Hopper (Lead Senior Medical Assessor, Licensing Division, MHRA) for providing independent clinical validation of the generated synthetic data. The Open access fees have been covered by a grant from NHS X.

ENDNOTES

[*] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/.

[†] https://archive.ics.uci.edu/ml/index.php.

[‡] https://www.cprd.com/primary-care.

[§] https://www.qrisk.org/three/index.php.

[¶] https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/

DATA AVAILABILITY STATEMENT

The Indian liver patient dataset (ILPD) that supports the findings of this study is deposited in UCI machine learning repository (https://archive.ics.uci.edu/ml/index.php). The Cardiovascular disease dataset that supports the findings of this study is available on request from CPRD (https://www.cprd.com/primary-care). The data are subject to a full licence agreement containing detailed terms and conditions of use.

ORCID

Zhenchen Wang https://orcid.org/0000-0003-4710-0298

REFERENCES

1. Khalid S, Ali MS, Prieto-Alhambra D. Cluster analysis to detect patterns of drug use from routinely collected medical data. Paper presented at: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad; 2018:194–198.
2. Ravizza S, Huschto T, Adamov A, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat Med*. 2019;25(1):57-59.
3. Sebastian Vollmer, Bilal A. Mateen, Gergo Bohner, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. arxiv.org/abs/1812.10404; 2018.
4. Lee ES, Whalen T. Synthetic designs: a new form of true experimental design for use in information systems development. *Perform Eval Rev*. 2007;35(1):191-202.
5. Park Y, Ghosh J. PeGS: perturbed Gibbs samplers that generate privacy-compliant synthetic data. *Trans Data Privacy*. 2014;7(3):253-282.
6. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R. *Stat J IAOS*. 2017;33(3):785-796.
7. Wu L, He H, Zaïane OR. Utility of privacy preservation for health data publishing. Paper presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; 2013:510–511.
8. Mandal S, Greenblatt AB, An J. Imaging intelligence: AI is transforming medical imaging across the imaging Spectrum. *IEEE Pulse*. 2018;9(5):16-24.
9. Chen Z, Salazar E, Marple K et al. An AI-based heart failure treatment adviser system. *IEEE J Transl Eng Health Med*. 2018;6:1-10.
10. Robnik-Šikonja M. Data generators for learning systems based on RBF networks. *IEEE Trans Neural Netw Learn Syst*. 2016;27(5):926-938.
11. Ruscio J, Kaczetow W. Simulating multivariate nonnormal data using an iterative algorithm. *Multivar Behav Res*. 2008;43(3):355-381.
12. Lee J, Hong J, Hong B, Ahn J. A generator of test data set for tactical moving objects based on velocity. Paper presented at: 2016 IEEE International Conference on Big Data (Big Data), Washington, DC; 2016:4011–4013.
13. Buczak AL, Babin S, Moniz L. Data-driven approach for creating synthetic electronic medical records. *BMC Med Inform Decis Mak*. 2010;10:1.
14. Riano D, Fernandez-Perez A, HEC International Joint Workshop on Knowledge Representation for Health Care, KR4HC/ProHealth 2016. Simulation-based episodes of care data synthetization for chronic disease

patients. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Berlin/Heidelberg, Germany: springer science+business media; 2017:36-50.

15. Syahaneim RA, Hazwani N, Wahida SI, Shafikah Z, Ellyza PN. Automatic artificial data generator: framework and implementation. Paper presented at: 2016 International Conference on Information and Communication Technology (ICICTM); Kuala Lumpur, 2016:56–60.
16. Iftikhar N, Liu X, Nordbjerg FE, Danalachi S. A prediction-Based Smart Meter Data Generator. Paper presented at: 19th International Conference on Network-Based Information Systems (NBiS), Ostrava; 2016:173–180.
17. Cano I, Torra V. Generation of synthetic data by means of fuzzy c-regression. Paper presented at: 2009 IEEE International Conference on Fuzzy Systems, Jeju Island; 2009:1145–1150.
18. Kontopantelis E, Parisi R, Springate DA, Reeves D. Longitudinal multiple imputation approaches for body mass index or other variables with very low individual-level variability: the mibmi command in Stata. *BMC Res Notes*. 2017;10(1):1-21.
19. Caiola G, Reiter JP. Random forests for generating partially synthetic, categorical data. *Trans Data Privacy*. 2010;3(1):27-42.
20. Yang S, Zhou Y, Guo Y, Farneth RA, Marsic I, Randall BS. Semi-synthetic trauma resuscitation process data generator. Paper presented at: IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT; 2017: 573.
21. Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. Paper presented at: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC; 2016:399–410.
22. EU General Data Protection Regulation. https://www.eugdpr.org. Accessed January 20, 2019.
23. Langarizadeh M, Orooji A, Sheikhtaheri A, 12th Annual Conference on Health Informatics Meets eHealth, eHealth 2018. Effectiveness of anonymization methods in preserving patients' privacy: a systematic literature review. *Stud Health Technol Inform*. 2018;248:80-87.
24. ISO. Health informatics—pseudonymization; 2017.
25. ICO. Anonymisation: managing data protection risk code of practice; 2015.
26. Sweeney L. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh; 2000.
27. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. Paper presented at: 2008 IEEE Symposium on Security and Privacy (SP 2008), Oakland, CA; 2008:111–125.
28. Bache R, Taweel A, Miles S, Delaney BC. An eligibility criteria query language for heterogeneous data warehouses. *Methods Inf Med*. 2015;54(1):41-44.
29. Pedersen TB, Jensen CS. Multidimensional data modeling for complex data. Paper presented at: Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337), Sydney, NSW, Australia; 1999:336–345.
30. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:1-21.
31. Begum SH, Nausheen F. A comparative analysis of differential privacy vs other privacy mechanisms for Big Data. Paper presented at: 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore; 2018:512–516.
32. Kumar P. Copula functions as a tool in statistical modeling and simulation. Paper presented at: 2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS), Delhi; 2009:1–5.
33. Fuster-Parra P, Tauler P, Bennasar-Veny M, Ligeza A, Lopez-gonzalez A, Aguilo A. Bayesian network modeling: a case study of an epidemiologic system analysis of cardiovascular risk. *Comput Methods Programs Biomed*. 2016;126:128-142.
34. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29(1):1-27.
35. Ester M, Kriegel H-P, Sander J, Xiaowei X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad U, eds. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. Palo Alto, CA: AAAI Press; 1996:226-231.

36. Venkata RB, Babu MSP, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. *Int J Database Manag Syst*. 2011;3(2):101-114.
37. Sklar M. Fonctions de repartition an dimensions et leurs marges. *Publ Inst Statist Univ Paris*. 1959;8:229-231.
38. Kao SC, Kim HK, Liu C, Cui X, Bhaduri BL. Dependence-preserving approach to synthesizing household characteristics. *Transp Res Rec*. 2012;2302(1):192-200.
39. Demarta S, McNeil AJ. The t copula and related copulas. *Int Stat Rev*. 2007;73(1):111-129.
40. Fisher RA. The statistical utilization of multiple measurements. *Ann Eugen*. 1938;8:376-386.
41. Wolf A, Dedman D, Campbell J, et al. Data resource profile: clinical practice research datalink (CPRD) aurum. *Int J Epidemiol*. 2019;48(6):1740.
42. Collins GS, Altman DG. *Predicting the 10 Year Risk of Cardiovascular Disease in the United Kingdom: Independent and External Validation of an Updated Version of QRISK2*. London, UK: BMJ Publishing Group Ltd.; n.d.
43. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open*. 2014;4(8):e005809-e005809. http://dx.doi.org/10.1136/bmjopen-2014-005809.
44. Multani P, Niemann U, Cypko M, Kühn J-P, Völzke H, Oeltze-Jafra S, Spiliopoulou M. Building a Bayesian network to understand the interplay of variables in an epidemiological population-based study. Paper presented at: Proceedings of the 31th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS18)'; 2018:88–93.
45. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
46. Wolpert D. Stacked generalization. *Neural Netw*. 1992;5(2):241-259.
47. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58(1):267-288.
48. Loh W-Y. Classification and regression trees. *Wiley Interdiscip Rev Data Mining Knowl Discov*. 2011;1:14-23.
49. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63:3-42.
50. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85-117.
51. Chen D, Plemmons RJ. *Nonnegativity Constraints in Numerical Analysis*. Hackensack, NJ: World Scientific; 2009:109-139.
52. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20(8):832-844.
53. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. (2020). *BMJ*.
54. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digit Med*. 2020;3:147.

How to cite this article: Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*. 2021;1–33. https://doi.org/10.1111/coin.12427