Journal of the American Medical Informatics Association, 28(4), 2021, 801–811
doi: 10.1093/jamia/ocaa303
Advance Access Publication Date: 23 December 2020
Research and Applications

Research and Applications

Application of Bayesian networks to generate synthetic health data

Dhamanpreet Kaur [1], Matthew Sobiesk,[1] Shubham Patil,[2] Jin Liu,[3] Puran Bhagat,[3] Amar Gupta,[1] and Natasha Markuzon[3]

[1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA[2]Rochester Institute of Technology, Rochester, New York, USA, and [3]Clinical Informatics, Philips Research North America, Cambridge, Massachusetts, USA

Corresponding Author: Dhamanpreet Kaur, BS, Massachusetts Institute of Technology, 32 Vassar St # 32-256, Cambridge, MA 02139, USA (dkaur@mit.edu)

Received 17 August 2020; Editorial Decision 9 November 2020; Accepted 16 November 2020

ABSTRACT

Objective: This study seeks to develop a fully automated method of generating synthetic data from a real dataset that could be employed by medical organizations to distribute health data to researchers, reducing the need for access to real data. We hypothesize the application of Bayesian networks will improve upon the predominant existing method, medBGAN, in handling the complexity and dimensionality of healthcare data.

Materials and Methods: We employed Bayesian networks to learn probabilistic graphical structures and simulated synthetic patient records from the learned structure. We used the University of California Irvine (UCI) heart disease and diabetes datasets as well as the MIMIC-III diagnoses database. We evaluated our method through statistical tests, machine learning tasks, preservation of rare events, disclosure risk, and the ability of a machine learning classifier to discriminate between the real and synthetic data.

Results: Our Bayesian network model outperformed or equaled medBGAN in all key metrics. Notable improvement was achieved in capturing rare variables and preserving association rules.

Discussion: Bayesian networks generated data sufficiently similar to the original data with minimal risk of disclosure, while offering additional transparency, computational efficiency, and capacity to handle more data types in comparison to existing methods. We hope this method will allow healthcare organizations to efficiently disseminate synthetic health data to researchers, enabling them to generate hypotheses and develop analytical tools.

Conclusion: We conclude the application of Bayesian networks is a promising option for generating realistic synthetic health data that preserves the features of the original data without compromising data privacy.

Key words: synthetic data, health data, data dissemination, disclosure risk, Bayesian networks

INTRODUCTION

Background and significance

Research surrounding healthcare data is incredibly useful for improving patient care through many different avenues—enhancing the interoperability of the medical system, aiding in clinical decision support, and providing feedback on effective practices.[1] However, data privacy concerns and patient confidentiality regulations continue to pose a major barrier to researchers seeking to access health data.[2,3] The process of anonymizing data is often tedious and costly, and can distort important features from the original dataset.[4] Moreover, these data are susceptible to reidentification attacks even when deidentified in accordance with existing standards.[5,6]

An alternative approach to sharing data while protecting privacy involves the generation of synthetic data.[7] A synthetic dataset preserves the user's ability to draw valid inferences, without an explicit mapping to the real data.[8] Several methods of partial and fully synthetic data generation have been proposed, including the use of random forests, support vector machines, generative adversarial networks, Gibbs sampling, decision trees, Bayesian bootstrapping and Bayesian mixture models.[9–16] Nonetheless, the application of these approaches to simulating health data has not been well studied, and in some cases, the underlying assumptions do not fit medical data very well. For example, Bayesian mixture models have only been illustrated for use with categorical data, and the hierarchical Bayesian model applied to marked point processes is specialized to work on spatial data.[14,17]

Using synthetic patient records has gained attention with the introduction of Synthea,[18] a software for simulating life spans of synthetic patients, but it uses existing public data and is not adaptable to clinical entities distributing a synthetic version of their own health data. A study on the validity of data generated by Synthea found that the calculated clinical quality measures differ substantially from those observed in the real data.[19] Deep learning methods—medGAN and its successor medBGAN—are predominantly used to synthesize health data.[15,20] However, they are limited to binary and count variables, which represent a small subset of potential medical data types.[20] While machine learning and deep learning methods are capable of detecting latent relationships in health data, these models generally lack the transparency and interpretability that is especially crucial in health analytics applications.[21–23]

Synthetic data generation

This project seeks to investigate the use of Bayesian networks to generate synthetic health data. Bayesian networks have been found to be increasingly useful over other machine learning methods for medical data in terms of classification and prediction[24,25] and can produce accurate topology, relating diseases and symptoms to facilitate clinical decision support.[26,27] In this study, we aim to simulate realistic synthetic data based on the relationships captured in learning a Bayesian network from the original data. Synthetic data are of growing importance in several domains including: (i) assisting data scientists in their endeavors to build their own AI-based models—especially in situations where complete and consistent datasets are not available; (ii) supporting research in enhancing quality of healthcare and patient safety and other areas that require analysis of quality data; (iii) providing a powerful mechanism to surmount the barrier of deidentification of patient data; and (iv) facilitating the creation of case studies that can be used for education of medical professionals and others.

The probabilistic graphical structure of Bayesian networks takes the form of a directed acyclic graph in which the nodes represent variables and the edges indicate a dependency between variables.[28] The network can be specified by user input, learned from the data, or result from a hybrid of user input and data.[29,30] Early research concluded that prior knowledge about the network was required to generate realistic synthetic data.[31] More recent methods, such as PrivBayes, showed promising results on using Bayesian networks to generate synthetic census datasets.[32] The application of Bayesian networks to generate synthetic health data has not yet been investigated to the best of our knowledge.

Synthetic data generation methods must be evaluated by their ability to preserve analytical validity and minimize disclosure risk.[33] Previous research has verified the validity of synthetic data via graphical comparisons, statistical tests, and machine learning inference.[34,35] Studies have assessed disclosure risk via attribute disclosure.[15]

OBJECTIVE

We hypothesized that Bayesian networks could offer transparency and computational efficiency over existing methods as well as possess the capability to handle various data types. We compare this method to medBGAN,[20] the predominant existing method of synthetic health data generation, using metrics of validity and disclosure risk. We provide details of the proposed method of generating synthetic data from a real dataset that can be employed by hospitals to distribute patient data to researchers, reducing the need for access to real data.

MATERIALS AND METHODS

In this section, we describe the medical datasets used in this study, then discuss the use of Bayesian networks to create synthetic data, and lastly outline the approaches for validating the generated data.

Description of datasets

We generated and evaluated synthetic datasets for several publicly available datasets containing health records that vary in dimensionality, data type, parameter distributions, and other characteristics. The first was the Medical Information Mart for Intensive Care (MIMIC-III) database, comprising deidentified electronic health records (EHRs) associated with 60K patient admissions to critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.[36] We extracted the dataset of patient diagnoses (DIAGNOSES_ICD)—943 variables taking on the value of the number of times the patient was diagnosed with the given code—and generated binary and count versions of this dataset as done by Baowaly et al.[20]

The other two datasets were from the UCI Machine Learning Repository.[37,38] One is the heart disease dataset from the Cleveland database, comprising 303 patients and 14 variables,[39] which tested the ability of our method to train accurately in spite of small numbers of patient records. The other is the UCI diabetes dataset, which includes 47 features representing patient and hospital outcomes over 10 years (1999–2008) of clinical care at 130 US hospitals and integrated delivery networks.[40] Data were limited to diabetic encounters of 1–14 days during which laboratory tests were performed and medications were administered. This tested the method's ability to handle incomplete data (assumed to be missing at random) and multiple visits from the same patient. The characteristics of these datasets are summarized in Table 1.

Using Bayesian networks

Bayesian networks are probabilistic graphical structures used to represent a set of variables and their conditional dependencies. They are defined by a directed acyclic graph $G = (V, E)$, where $v_i \in V$ corresponds to a random variable X_i, and a global probability distribution X with parameters Θ. These networks possess the local Markov property, which means the probability distribution of a variable X_i can be determined given the values of all variables X_j such that there exists an edge $e_{ji} \in E$. Thus, we can factorize the global probability distribution into local distributions as follows:

Table 1. Characteristics of datasets: dimensionality, missing values, and data types

Dataset	Heart Disease, UCI	Diabetes, UCI	MIMIC-III (binary)	MIMIC-III (count)
Number of patient records	303	101,766	46,520	46,520
Total number of variables	14	47	943	943
Variable types	4 binary 5 categorical 5 numeric	9 binary 32 categorical 6 numeric	943 binary	943 numeric
Variable descriptions	Demographics, physiological parameters	Demographics, diagnoses, medications	Diagnoses	Diagnoses
Average % of records with missing data per variable	0	4%	0	0

Abbreviations: UCI, University of California Irvine; MIMIC, Medical Information Mart for Intensive Care.

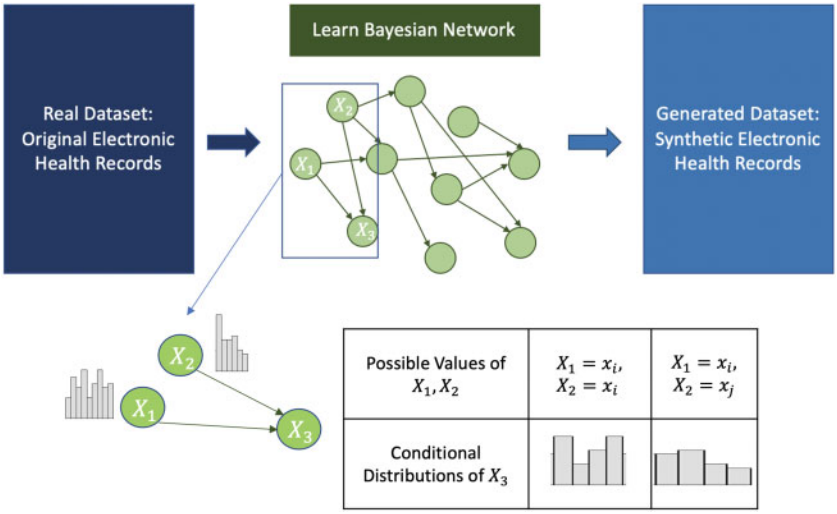

Figure 1. Local probability distributions in network structure used to simulate synthetic data.

$$P(X) = \prod_{i=1}^{N} P(X_i \quad \{parents\ of\ X_i\}; \Theta_{X_i})$$

We created a Bayesian network model for the data in which the conditional distribution of a variable is captured at each node and the edges represent probabilistic dependencies between them. Learning the network model M for a given set of data D was a 2-part process: learning the structure G given the data and learning the parameters given the data and the structure.

$$P(M|D) = P(G|D) * P(\Theta \quad G, D)$$

To first learn the structure, the score-based algorithm assigned each candidate network a goodness of fit and took this as the objective function to optimize. Since it is not computationally feasible to search the entire space of graphs on the set of nodes, a hill climbing approach was used.[41] An initial graph was chosen (usually with no edges), and for each edge deletion, reversal or addition that did not create a cyclic network and increased the score of the graph, we modified the graph accordingly. We used the Bayesian Information Criterion as the scoring function for comparison of networks during hill climbing.[42] Knowing the structure of the graph, the parameters Θ of the global distribution X were determined by estimating the parameters for the local distributions using maximum likelihood estimators.

Given a Bayesian network, we can simulate random samples to create a synthetic dataset as follows. By the definition of a directed acyclic graph, there must exist a topological ordering of nodes in the network.[43] For the set of nodes with no incoming edges, we sampled from their unconditional distributions. We then proceeded to sample from the conditional distributions of their children given the previously sampled values and iterated until values had been sampled for all nodes.[44] This constituted one row of data in the synthetic dataset. The process was repeated to independently sample the desired number of rows, which we set to be the size of the original dataset. Figure 1 illustrates the process: in this case, X_1 and X_2 are nodes without incoming edges so their values are simulated first independently, and since they are the only parents of node X_3, the value for X_3 is then generated from distributions conditional on the values of X_1 and X_2. We accomplished this process through a suite of functions encompassed by the *bnlearn* software package in R, which offers customization of the learning algorithm and scoring parameters.[29]

Validation of synthetic data

The validity of the generated synthetic data was determined through several metrics which test if the information learned from the real data can also be learned from the synthetic data. We applied these

Table 2. Validations performed on each dataset

Validation	MIMIC Binary	MIMIC Count	UCI Heart Disease	UCI Diabetes
Graphical comparison			X	X
Kolmogorov–Smirnov test	X	X	X	X
Dimension-wise means	X	X		
Rare data	X	X		
Association rule mining	X	X	X	X
Correlation matrices	X	X	X	X
Machine learning prediction	X	X	X	
Attribute disclosure risk	X			
Discriminator metric	X	X	X	X

evaluation measures as suitable for each dataset (Table 2). Data are often analyzed via summary-level statistics so we confirmed univariate similarity between the variables by comparing dimension-wise probabilities for categorical variables and dimension-wise averages for numerical variables. We employed the two-sample Kolmogorov–Smirnov test to compare the probability distribution functions of each variable between real and synthetic datasets. Since association rule mining is widely used on health data to identify the cooccurrence of clinical variables,[45] we used this approach to determine if the rules present in the original data were preserved in the synthetic data. To further test the preservation of multivariate relationships, we compared correlation matrices between the real and synthetic data.[46] Lastly, machine learning methods are commonly employed in inferring relationships between variables. We tested dimension-wise prediction by excluding an outcome variable, training a machine learning model on the remaining variables, and comparing the predictive capacities between the real and synthetic datasets.

To determine if a machine learning algorithm can discriminate between the real and synthetic datasets, we created a merged dataset that includes equal proportions of real and synthetic data and then introduced an additional variable that labels the two types of data. We randomly sampled 75% of this dataset to create our training data and trained a neural network as a binary classifier to predict whether the data were real or synthetic. We then tested the trained neural network on the remaining 25% of the dataset and reported the classification accuracy. If the synthetic data were sufficiently similar to the real data, the neural network would have difficulty in distinguishing between them and the accuracy would be around 50%—essentially just guessing which of the data were real versus synthetic. However, if the synthetic data failed to capture the underlying relationships between variables, the neural network would be able to distinguish the real data in the merged set and the accuracy would be much greater than 50%.

Privacy consideration

Although the risk of disclosure is sometimes dismissed in the case of synthetic data,[20] one must take into consideration that the best synthetic data generation method via the previous metrics would be one that exactly copies the real data and consequently defeats the entire purpose of the method. Attribute disclosure occurs when an adversary discovers crucial information about a target patient P based on some known attributes of patient P and observation of similar patients in the synthetic dataset.[47] Based on the method for the attribute disclosure analysis of the binary data described by Choi et al,[15] 1% of the real data were randomly sampled as compromised records R. For each compromised record r, the attacker was assumed to know S random attributes out of C total attributes. Then, the k-nearest neighbors' classifier was used to find the k-nearest neighbors of each compromised record r from the synthetic dataset. We assigned the most frequent value from the k-nearest neighbors for each of unknown $C - S$ attributes and calculated the average precision and average sensitivity over all compromised records R. Disclosure risk would be minimal if both the precision and sensitivity are zero.

RESULTS

We evaluated our approach using datasets of varying size, data type, and complexity: the MIMIC-III binary dataset, the MIMIC-III count dataset, the UCI heart disease dataset, and the UCI diabetes dataset.

Bayesian network construction

The networks were learned without the input of any prior knowledge through the *bnlearn* package in R.[29] The output, as shown for the diabetes dataset in Figure 2, is a graphical structure in which the nodes represent all of the variables present in the original dataset, and the edges indicate the path of dependencies for the data simulation step. We include networks for other datasets in Supplementary Figures 1–3.

Graphical comparisons

A visual comparison through histograms of the variable distributions in the real data and those generated in the synthetic data revealed that the method closely captured the variety of variable distributions found in real data (Figure 3). Examples include distributions with multiple peaks and skewed distributions comprising continuous numerical, discrete numerical, and categorical data types.

Kolmogorov–Smirnov test results

We used the Kolmogorov–Smirnov test to compare the one-dimensional probability distributions of variables between the real and synthetic datasets. Using a P value of .05, there were no variables in the MIMIC binary, MIMIC count, or heart datasets that showed statistically significant differences in their distributions between the real and synthetic data (Supplementary Table 1).

Dimension-wise means/probabilities

The dimension-wise means were compared between the real and synthetic data for the MIMIC binary and count data (Figure 4). The x-coordinate of each data point signifies the mean of a variable in real data and its y-coordinate signifies the mean of that variable in synthetic data. In the case of MIMIC binary data, this mean is just a Bernoulli probability of the variable. The diagonal line indicates the

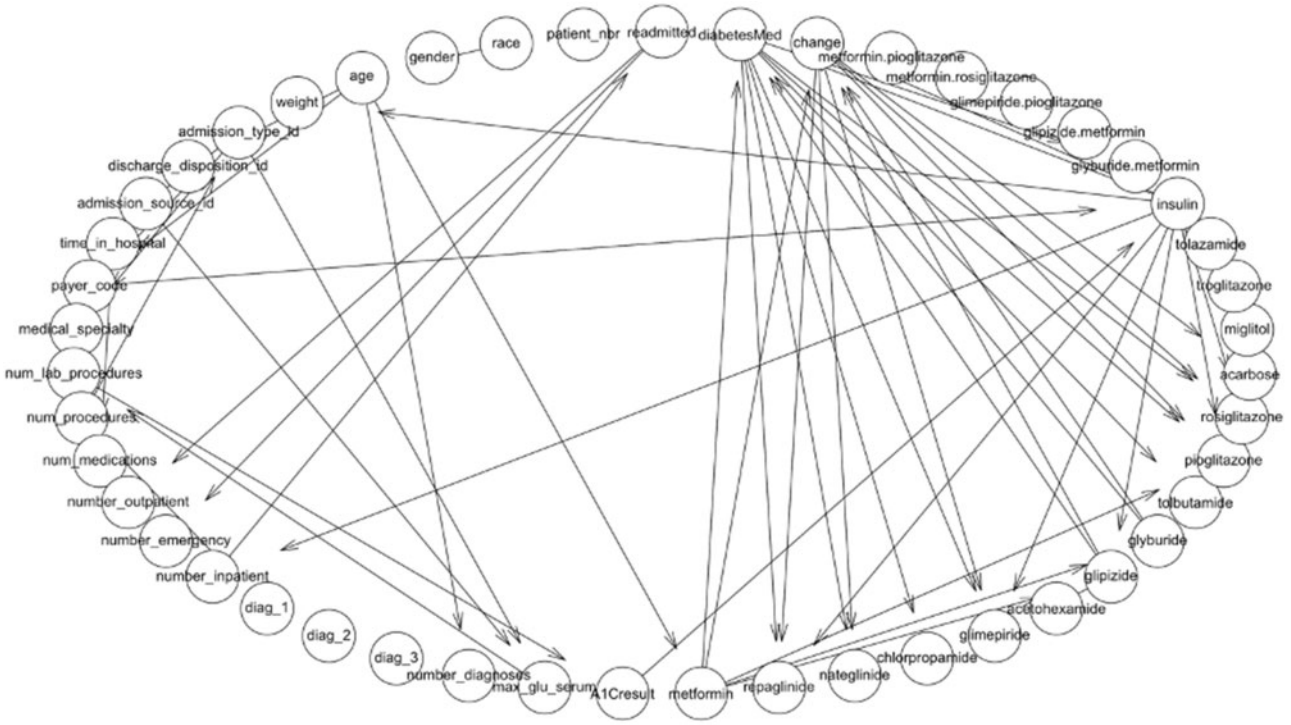

Figure 2. Bayesian network for diabetes dataset with each node representing a variable in the dataset and edges representing the path of data simulation.

ideal result in which the dimension-wise means between the real and synthetic data are identical. The correlation coefficients (0.99992 for MIMIC binary and 0.99981 for MIMIC count) indicate these results were remarkably close to 100% correlation.

Capturing rare data

The sparsity of the MIMIC datasets enabled the evaluation of the method's ability to capture rare variables, which can be difficult while minimizing disclosure risk. Of the diagnosis codes present in original data, 1.91% did not show up in the generated MIMIC-III binary data and 2.33% did not show up in the generated MIMIC-III count data (Table 3). The number of rare variables in the original data was obtained by selecting those with mean < 0.0001. For binary MIMIC data, this means a variable was considered rare if less than 0.01% of patients had that diagnosis code. The percentage of these rare variables that appear in the synthetic data was calculated. Of the 129 rare variables in MIMIC binary data, 82.9% were retained, and of the 124 rare variables in MIMIC count data, 86.3% were retained.

Multivariate relationships

The Apriori algorithm for association rule mining with parameters $support = 0.05,\ confidence = 0.2,\ minimum\ length = 2, maximum\ length = 2-10$ was used to determine if relationships between variables were preserved between the real and synthetic datasets.[40] Setting the maximum length to two variables evaluated preservation of pairwise associations, whereas setting the maximum length to ten variables evaluated preservation of multivariate associations involving up to ten variables. After extracting the rules independently in the real and synthetic datasets, the two were compared in terms of precision and recall (Table 4). A high degree of precision was observed in the multivariate rules extracted from the synthetic data across all datasets with values ranging from 0.8637 to 1.000. The pairwise associations showed stronger preservation than the multivariate relationships. Further, a high degree of recall was observed for the diabetes and heart datasets, though somewhat lower recall was observed in MIMIC-III binary data and much lower recall was observed in MIMIC-III count data.

We constructed correlation matrices for each set of real and synthetic data and computed the difference between the correlation matrices in a third matrix. For illustration purposes on the binary MIMIC data, we selected the 15 most common diagnoses codes. The difference in correlation coefficients is less than 0.1 across these diagnosis variables (Figure 5). The correlation matrices for the other datasets are included in Supplementary Materials.

Machine learning prediction

The dimension-wise machine learning prediction performances were compared between real and synthetic datasets through 3 models: logistic regression, support vector machine, and random forest—columns 1, 2, and 3, respectively of Figure 6. Each point corresponds to 1 variable in the dataset, its x-coordinate indicating the F1-score achieved for its prediction in the real data and its y-coordinate indicating the F1-score achieved for its prediction in the synthetic data. The hyperparameters for each model were kept constant between real and synthetic versions of each dataset. The diabetes dataset was omitted from this analysis because certain variables were not appropriate for comparing prediction using these models. The heart disease dataset showed high correlation between F1-scores generated from real and synthetic data across all models (CC = 0.9714, 0.9754, 0.9768). The MIMIC binary data showed slightly lower levels of correlation (CC = 0.9331, 0.9103, 0.9485), as did the MIMIC count data (CC = 0.8230, 0.8288, 0.8405).

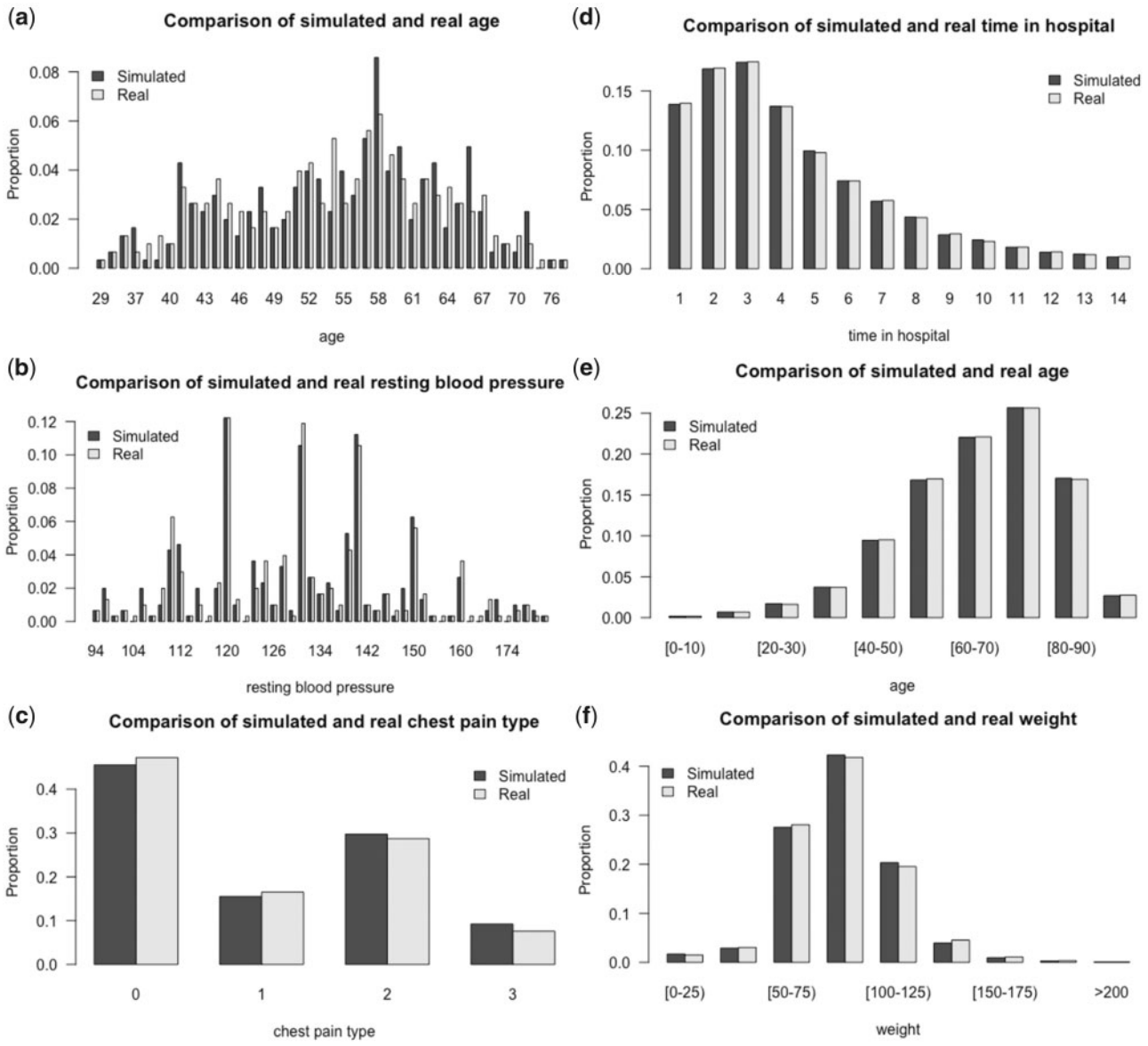

Figure 3. Histograms comparing real and simulated values for (a) age, (b) resting blood pressure, and (c) chest pain types from the heart disease dataset; (d) time in the hospital, (e) age, and (f) weight from the diabetes dataset.

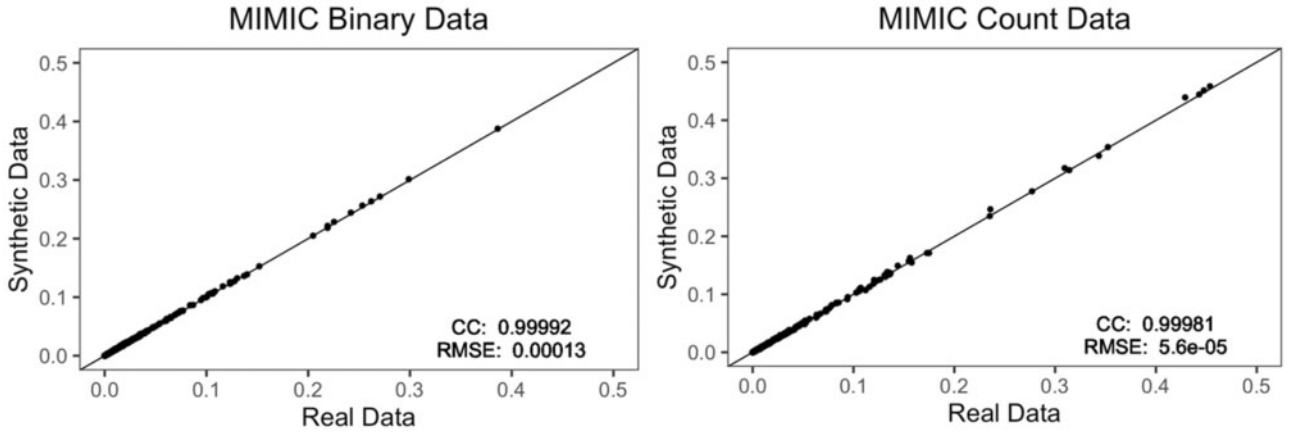

Figure 4. Comparison of dimension-wise means between real and synthetic data for binary and count versions of the MIMIC-III dataset.

Table 3. Retention of rare variables in synthetic data

Dataset	% of Diagnosis Codes Lost	# Rare Variables in Real Data (mean < 0.0001)	% of Rare Variables Retained in Synthetic Data
MIMIC-III binary	2.33%	129	82.9%
MIMIC-III count	1.91%	124	86.3%

Table 4. Comparison of relationships extracted via association rule mining

Dataset	Number of Variables	Number of Extracted Rules in Real Data	Number of Extracted Rules in Synthetic Data	Precision	Recall
Heart disease, UCI	2	166	167	0.9880	0.9940
	2–10	29,448	29,224	0.8637	0.8571
Diabetes, UCI	2	111	109	1.000	0.9819
	2–10	12,534	10,998	1.000	0.8775
MIMIC-III binary	2	231	189	0.9894	0.8095
	2–10	357	292	0.9692	0.7927
MIMIC-III count	2	231	136	0.9853	0.5801
	2–10	357	157	0.9682	0.4258

Abbreviation: UCI, University of California Irvine.

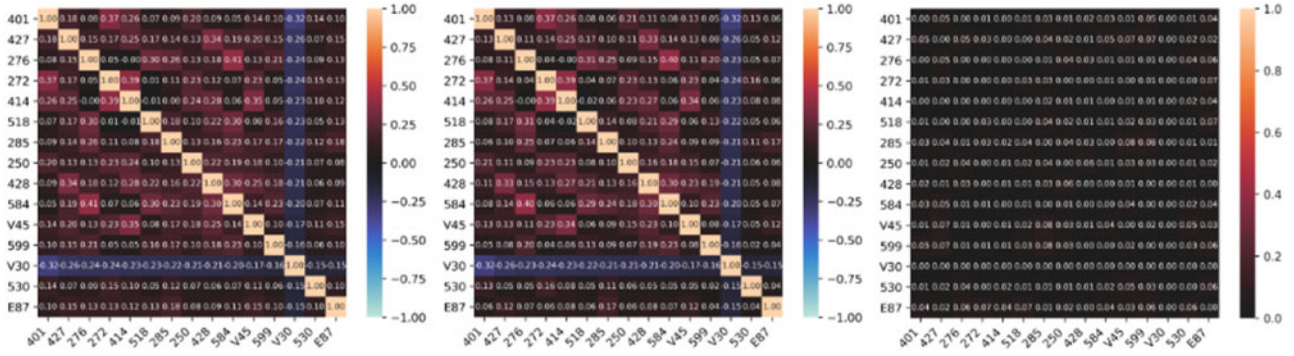

Figure 5. Correlation matrix for binary MIMIC data: real data (column 1), synthetic data (column 2), absolute difference between real and synthetic correlations (column 3).

Discriminator metric

We trained a neural network on a subset of labeled real and synthetic data and then measured its ability to correctly distinguish between the 2 in the testing set. With 10 iterations of neural network training, the mean in-sample and out-of-sample accuracy rates are shown (Table 5). The heart disease and diabetes datasets showed virtually indistinguishable real and synthetic datasets. The MIMIC-III count data had partially distinguishable real and synthetic data (out-of-sample accuracy 0.7057). Nonetheless, the relatively low in-sample accuracies (highest being 0.7377 for MIMIC count data) indicated the model did not train well in distinguishing between the real and synthetic datasets.

Disclosure risk

We verified no exact duplication of rows occurred in generating any of the synthetic datasets. We assessed disclosure risk via attribute disclosure and differential privacy.

Figure 7 shows the effect of varying the number of attributes known to the attacker and the number of nearest neighbors on the precision and sensitivity of predicting unknown positive variables. The sensitivity was less than 25% even with 256 variables revealed. An attacker who knows approximately 1% of the target patient's attributes estimates the target's unknown attributes with 16% precision and 17% sensitivity compared to 20% precision and 10% sensitivity observed in data generated by medGAN.

Comparison to MedBGAN

We tested the proposed method against medBGAN, the current state-of-the-art for synthetic data generation that has been applied to clinical data involving numerical data.[20] Table 6 shows that the Bayesian network outperformed medBGAN in similarity of univariate distributions as measured by the Kolmogorov–Smirnov test (100.0% vs 97.45% for binary data, 100.0% vs 89.70% for count data), and greater correlation between dimension-wise means existed in the proposed method (CC = 0.99992 vs. 0.9929, RMSE = 0.00013 vs 0.0041). In comparing the results of association rule mining, the proposed method had much greater precision (0.9894 vs 0.4379) but somewhat lower recall (0.8095 vs 0.9305). To determine which had a better tradeoff between precision and recall, the F1-scores (0.8905 vs 0.5955) indicated that the proposed method performs better in association rule mining. For the measure of dimension-wise prediction, the proposed method showed comparable correlation of F1-scores in the

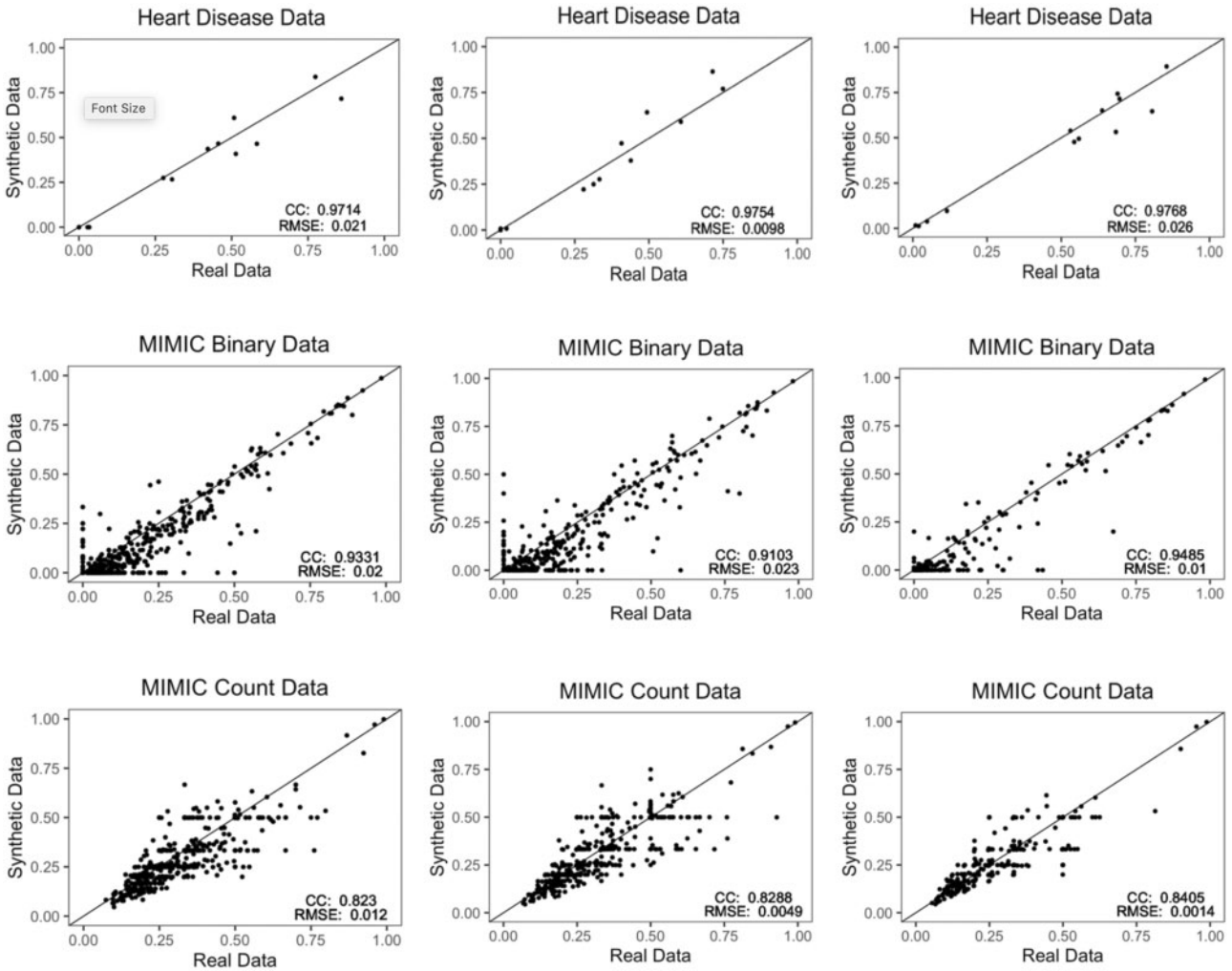

Figure 6. Correlation of F1-scores using models of logistic regression (first column), support vector machine (second column), and random forest (third column).

Table 5. Discriminator performance in distinguishing between real and synthetic data

Dataset	In-Sample Accuracy	Out-of-Sample Accuracy
Heart disease, UCI	0.5117	0.5039
Diabetes, UCI	0.5052	0.5052
MIMIC-III binary	0.6502	0.6099
MIMIC-III count	0.7377	0.7057

Abbreviation: UCI, University of California Irvine.

binary data and lower correlation of F1-scores in the count data. Lastly, Bayesian networks performed substantially better in preserving rare variables, losing only 2.33% of diagnosis codes, whereas medBGAN lost 17.30% of diagnosis codes.

DISCUSSION

In this study, we investigated the use of Bayesian networks to generate synthetic health data containing numerical and categorical variables. We evaluated the validity of the method in generating synthetic data that captures the key characteristics of the real data through comparison in graphical, statistical, and machine learning inference tests. We confirmed the method is not overfitting to the original data by verifying sufficiently low risk of disclosure.

Comparison of results to prior methods

We tested our method on the MIMIC count and binary datasets as a means of direct comparison to medBGAN, the predominant existing method for synthetic data generation of binary and count health data (Table 6). Our method offered improvements over medBGAN in all measures of univariate similarity, including mean (Figure 4) and cumulative probability distribution (Supplementary Table 1). Association rule mining showed that our method has much greater precision in preserving the relationships present in the original data, though

Table 6. Summary of results in comparison to medBGAN

Dataset	Performance Measure	medBGAN	Bayesian Network	Ideal Value
MIMIC binary data	Kolmogorov–Smirnov test	97.45% similarity	100.0% similarity	100% similarity
	Dimension-wise mean	CC: 0.9929	CC: 0.99992	CC: 1.0
		RMSE: 0.0041	RMSE: 0.00013	RMSE: 0.0
	Association rule mining (Length 2)	Precision: 0.4379	Precision: 0.9894	Precision: 1.0
		Recall: 0.9305	Recall: 0.8095	Recall: 1.0
	Logistic regression	CC: 0.9794	CC: 0.9331	CC: 1.0
		RMSE: 0.0421	RMSE: 0.0203	RMSE: 0.0
	Support vector machine	CC: 0.9540	CC: 0.9103	CC: 1.0
		RMSE: NA	RMSE: 0.0232	RMSE: 0.0
	Random forest	CC: 0.9733	CC: 0.9485	CC: 1.0
		RMSE: NA	RMSE: 0.0102	RMSE: 0.0
MIMIC count data	Kolmogorov–Smirnov test	89.70% similarity	100.0% similarity	100% similarity
	Dimension-wise mean	CC: 0.9935	CC: 0.99981	CC: 1.0
		RMSE: 0.0057	RMSE: 0.000056	RMSE: 0.0
	Logistic regression	CC: 0.9441	CC: 0.8230	CC: 1.0
		RMSE: 0.0709	RMSE: 0.0120	RMSE: 0.0
	Support vector machine	CC: 0.9589	CC: 0.8288	CC: 1.0
		RMSE: NA	RMSE: 0.00491	RMSE: 0.0
	Random forest	CC: 0.9593	CC: 0.8405	CC: 1.0
		RMSE: NA	RMSE: 0.00143	RMSE: 0.0
	Preserving rare variables	17.30% of diagnosis codes lost	2.33% of diagnosis codes lost	0% of diagnosis codes lost

Abbreviations: CC, correlation coefficient; NA, not applicable; RMSE, root mean square error.

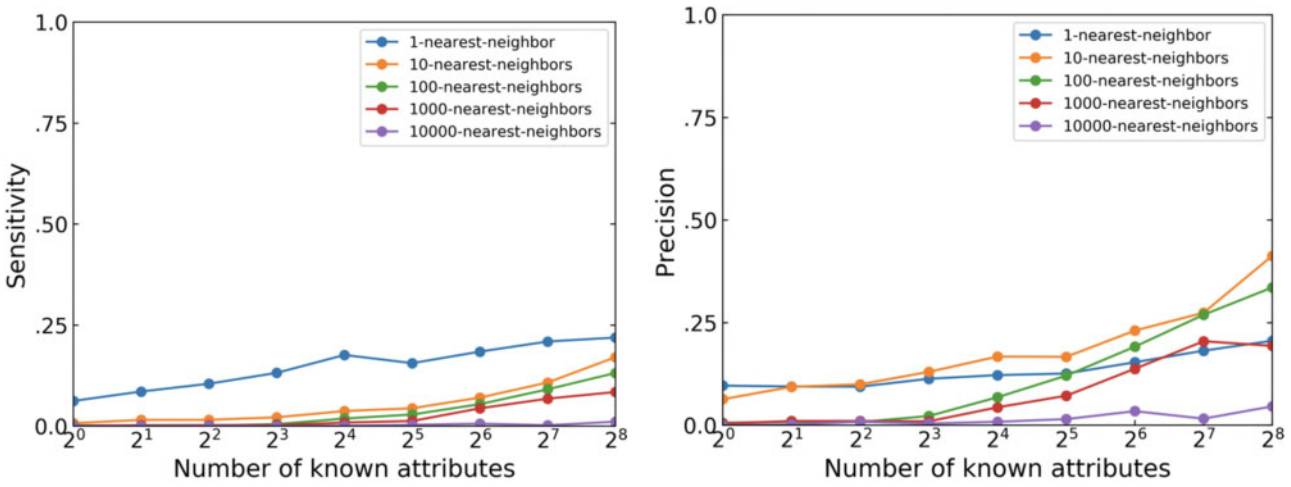

Figure 7. Effect of number of variables on precision and sensitivity, 943 total attributes.

medBGAN is slightly better in recall (Table 4). It was comparable to medBGAN in machine learning prediction tasks for the binary data, though medBGAN outperformed for the count data (Figure 6). Note, however, the medBGAN model lost 17% of the diagnosis codes—which were likely the most difficult to predict due to their infrequency—whereas our method only lost 2% of diagnosis codes (Table 3). Moreover, this measure was highly sensitive to the different parameters for training their models and calculating F1-scores. For example, we obtained much better correlation in our method using weighted F1-scores (see Supplementary Figure 4). Lastly, our method showed lower or comparable attribute disclosure when compared to medGAN (Figure 7). Our method did a significantly better job of capturing rare data attributes when applied to the MIMIC-III dataset whilst maintaining sufficiently low disclosure risk (Table 3).

Discriminator metric

With such a wide array of validation measures, it may be difficult to assess what matters most in assessment of synthetic data generation methods. A need exists to provide a generalizable standard for holistic dataset similarity. While generative adversarial networks have employed the idea of discriminating between real and synthetic data during model training,[15] here we used it as a metric of validation. In this study, the discriminator metric highlighted the diminished performance of the method in generating MIMIC count versus MIMIC binary data where all other metrics, besides association rule mining, failed to. Given it is not feasible to exhaustively compare the set of inferences that can be drawn from the real versus synthetic data, we believe future research in this area could use this discriminator metric as a holistic validation method for synthetic data generation.

Advantages of a Bayesian network

Bayesian networks are well-suited in their application to health data. They operate with great transparency, providing an easily interpretable graphical structure which is directly used for simulation. Moreover, this graphical structure allows for input of domain knowledge: the user can prevent certain edges from existing in the graph (blacklisting) and also require that certain edges exist in the graph (whitelisting).[29] Additionally, Bayesian networks are able to handle both categorical and numerical data, whereas medBGAN can only generate binary and count data.[20] While many machine learning methods require large amounts of data for training, Bayesian networks showed no decline in performance when used on small datasets, such as the heart disease dataset which comprised only 300 patients and 14 variables (Table 1). The time constraints of the method were remarkably low: to train the model and generate synthetic data, it took 42.26 minutes for the binary MIMIC data and 16.37 minutes for the count data whereas medBGAN took 1.88 hours on each.[20] Supplementary Table 2 summarizes the time and memory usage required to train and generate data for each dataset on a 2.6 GHz 6-Core Intel Core i7 laptop. Prior studies of synthetic data generation have noted the greater scalability and computational efficiency of Bayesian networks over other methods.[48] Further evaluation may be needed to determine the scalability of the proposed method to medical datasets on the scale of millions of records or several thousand variables.

Limitations

While the method to generate synthetic data from a real dataset was entirely automated, some limitations existed in preprocessing the data for effective network learning. We found that the model was best suited to capture continuous numeric variables following a Gaussian distribution. This may explain the diminished performance in preservation of multivariate relationships in the MIMIC count data, as seen in the lower recall of association rules (Table 4) and greater difference in correlation matrices (Supplementary Figure 5). Most continuous variables thus required discretization, such as through k-means clustering,[49] or z-score standardization.[50] The latter avoids the need for user input and parameter tuning. Furthermore, the structure of the Bayesian networks proposed in this study does not address the hierarchical notion of multiple encounters per patient; the method assumes that each patient corresponds independently to a singular row in the dataset. In the UCI diabetes dataset, visits from the same patient correspond to multiple rows, which violates this assumption. Although the generated data showed high validity nonetheless, some inconsistencies were present in the data itself (ie, the same patient had a different race across visits). We propose future work could use Dynamic Bayesian networks[51] to better capture the temporality of such data. Similarly, advancements upon Bayesian methods, such as mixture models, may enhance synthesis of health data.[10] Although the proposed method was not explicitly compared to methods of synthetic data generation applied outside the realm of medical data, studies have shown Bayesian networks are among the highest performing, alongside nonparametric Bayesian methods,[52,53] when compared to imputation methods[54] and GANs[55] for synthetic generation of categorical cancer registry data.[48]

Synthetic data can be useful in expediting development of tools for analysis and research in healthcare, but all final conclusions may need to be confirmed on the original dataset. We hope this method will allow healthcare organizations to efficiently disseminate synthetic health data to researchers with diminished concerns for patient confidentiality.

CONCLUSION

Bayesian networks can be applied to generate realistic synthetic healthcare data containing numerical and categorical variables. The proposed method outperforms medBGAN, the predominant existing method for synthesizing health data, and is characterized by low risk of disclosure. Overall, Bayesian networks represent a promising candidate for use in generating synthetic health data that preserve the features of the original data without compromising patient confidentiality.

FUNDING

This project was funded in part by Philips Research North America.

AUTHOR CONTRIBUTIONS

DK contributed significantly to the conception/design of the work; acquisition, analysis and interpretation of the data; and drafting and critically revising the article. MS and SP contributed to the conception/design of the work, data analysis and interpretation, and drafting the article. JL and PB contributed to conception/design of the work. AG and NM contributed to conception/design of the work and critical revision of the work. All authors gave final approval of the version to be published.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Information Association* online.

ACKNOWLEDGMENTS

We thank Dr. Peter Szolovits for his technical insights and Dr. Ralph Panos for his clinical feedback.

DATA AVAILABILITY

The data underlying this article are available in the Dryad Digital Repository at https://doi.org/10.5061/dryad.ttdz08kws.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Cowie MR, Blomster JI, Curtis LH, *et al*. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106 (1): 1–9.
2. Coppen R, van Veen EB, Groenewegen PP, *et al*. Will the trilogue on the EU Data Protection Regulation recognise the importance of health research? *Eur J Public Health* 2015; 25 (5): 757–8.
3. Huser V, Cimino JJ. Don't take your EHR to heaven, donate it to science: legal and research policies for EHR post mortem. *J Am Med Inform Assoc* 2014; 21 (1): 8–12.
4. Rothstein MA. Is deidentification sufficient to protect health privacy in research? *Am J Bioeth* 2010; 10 (9): 3–11.
5. Emam KE, Jonker E, Arbuckle L, *et al*. A systematic review of re-identification attacks on health data. *PLoS One* 2011; 6 (12): e28071.
6. Jayabalan M, Rana ME. Anonymizing healthcare records: a study of privacy preserving data publishing techniques. *Adv Sci Lett* 2018; 24 (3): 1694–7.
7. Surendra H, Mohan HS. A review of synthetic data generation methods for privacy preserving data publishing. *IJSTR* 2017; 6 (3): 95–101.
8. Rubin DB. Discussion: statistical disclosure limitation. *J Offic Stat*1993; 9 (2): 461–8.

9. Reiter JP. Using CART to generate partially synthetic public use microdata. *J Offic Stat* 2005; 21 (3): 441–62.
10. Hu J, Reiter JP, Wang Q. Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Anal* 2018; 13 (1): 183–200.
11. Raghunathan T, Reiter JP, Rubin D. Multiple imputation for statistical disclosure limitation. *J Offc Stat* 2003; 19: 1–16.
12. Caiola G, Reiter JP. Random forests for generating partially synthetic, categorical data. *Trans Data Priv* 2010; 3 (1): 27–42.
13. Drechsler J. Using support vector machines for generating synthetic datasets. In: Domingo-Ferrer J, Magkos E, eds. *Privacy in Statistical Databases*. Berlin, Germany: Springer; 2010: 148–61.
14. DeYoreo M, Reiter JP. Bayesian mixture modeling for multivariate conditional distributions. *J Stat Theory Pract* 2016; 14: 1–27.
15. Choi E, Biswal S, Malin B, *et al.* Generating multi-label discrete patient records using generative adversarial networks. In: *proceedings of the 2nd Machine Learning for Healthcare Conference*, Vol. 68; 2017.
16. Park Y, Ghosh J, Shankar M. Perturbed Gibbs samplers for generating large-scale privacy-safe synthetic health data. In: proceedings of the 2013 IEEE International Conference on Healthcare Informatics; 2013. doi: 10.1109/ICHI.2013.76.
17. Quick H, Holan SH, Wikle CK, Reiter JP. Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spat Stat* 2015; 14 (C): 439–51.
18. Walonoski J, Kramer M, Nichols J, *et al.* Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc* 2018; 25 (3): 230–8.
19. Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019; 19 (1): Article44.
20. Baowaly M, Lin C, Liu C, *et al.* Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc* 2019; 26 (3): 228–41.
21. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019; 6 (2): 94–8.
22. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018; 15 (11): e1002689.
23. Cao X, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018; 25 (10): 1419–28.
24. Padman R, Bai X, Airoldi EM. A new machine learning classifier for high dimensional healthcare data. *Stud Health Technol Inform* 2007; 129 (Pt 1): 664–8.
25. Cai X, Perez-Concha O, Coiera E, *et al.* Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J Am Med Inform Assoc* 2016; 23 (3): 553–61.
26. Shen Y, Zhang L, Zhang J, *et al.* CBN: constructing a clinical Bayesian network based on data from the electronic medical record. *J Biomed Inform* 2018; 88: 1–10.
27. Klann JG, Anand V, Downs SM. Patient-tailored prioritization for a pediatric care decision support system through machine learning. *J Am Med Inform Assoc* 2013; 20 (e2): e267–74.
28. Neapolitan R. *Learning Bayesian Networks*. Chicago, IL: Prentice Hall; 2003.
29. Scutari M. Learning Bayesian networks with the bnlearn R Package. *J Stat Softw* 2010; 35 (3): 1–22.
30. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach Learn* 1995; 20 (3): 197–243.
31. Young J, Graham P, Penny R. Using Bayesian networks to create synthetic data. *J Off Stat* 2009; 25: 549–67.
32. Zhang J, Cormode G, Procopiuc CM, *et al.* PrivBayes: private data release via Bayesian networks. *ACM Trans Database Syst* 2017; 42 (4): 1423–34.
33. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput Stat Data Anal* 2011; 55 (12): 3232–43.
34. Mclachlan S, Dube K, Gallagher T, *et al.* The ATEN framework for creating the realistic synthetic electronic health record. In: proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies; 2018.
35. Reiner-Benaim A, Almog R, Gorelik Y, *et al.* Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR Med Inform* 2020; 8 (2): e16492.
36. Pollard TJ, Johnson AEW. The MIMIC-III Clinical Database. 2016. https://mimic.physionet.org/Accessed April 18, 2020
37. Dua D, Graff C. UCI Machine Learning Repository Heart Disease data set. 1988. http://archive.ics.uci.edu/ml/datasets/Heart+Disease?spm=5176.100239.blogcont54260.8.TRNGoOAccessed April 18, 2020
38. Dua D, Graff C. UCI Machine Learning Repository Diabetes 130-US hospitals for years 1999-2008 data set. 2014. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 Accessed April 18, 2020
39. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, *et al.* International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am J Cardiol* 1989; 64 (5): 304–10.
40. Strack B, DeShazo JP, Gennings C, *et al.* Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Res Int* 2014; 2014: 1–11.
41. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn* 2006; 65 (1): 31–78.
42. Liu Z, Malone B, Yuan C. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinform* 2012; 13 (Suppl 15): S14.
43. Cormen T, Leiserson C, Rivest R, *et al. Introduction to Algorithms*, 3rd ed. Cambridge, MA: MIT Press; 2009: 549–52.
44. Scutari M. *Understanding Bayesian Networks with Examples in R*. University of Oxford, Department of Statistics; 2017. https://dipartimenti.unicatt.it/scienze-statistiche-23-25-1-17ScutariSlides.pdf Accessed May 1, 2020.
45. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records (EHRs): a survey. *ACM Comput Surv* 2018; 50 (6): 1–40.
46. Bilici E, Arvanitis T, Despotou G. Generation of realistic synthetic validation healthcare datasets using generative adversarial networks. *Stud Health Technol Inform* 2020; 272: 322–5.
47. Matwin S, Nin J, Sehatkar M, *et al.* A review of attribute disclosure control. In: Navarro-Arribas G, Torra V, eds. *Advanced Research in Data Privacy*. Cham, Switzerland: Springer; 2015: 41–61.
48. Goncalves A, Ray P, Soper B, *et al.* Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020; 20 (1): 108.
49. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability; 1967.
50. Kreyszig E. *Advanced Engineering Mathematics*. Jefferson City, MO: Wiley 1979: 880.
51. Dagum P, Galper A, Horvitz E. Dynamic network models for forecasting. In: proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence; 1992.
52. Gal Y, Chen Y, Ghahramani Z. Latent Gaussian processes for distribution estimation of multivariate categorical data. In: proceedings of the International Conference on Machine Learning; 2015.
53. Dunson DB, Xing C. Nonparametric Bayes modeling of multivariate categorical data. *J Am Stat Assoc* 2009; 104 (487): 1042–51.
54. Reiter JP, Drechsler J. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Stat Sin* 2010; 20 (1): 405–21.
55. Camino R, Hammerschmidt C, State R. Generating multi-categorical samples with generative adversarial networks. In: proceedings of the ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models; 2018.