Practical Lessons from Generating Synthetic Healthcare Data with Bayesian Networks

Juan de Benedetti[1], Namir Oues[1], Zhenchen Wang[1],
Puja Myles[1], and Allan Tucker[2]

[1] Medicine and Health Regulatory Authority, UK
[2] Intelligent Data Analysis Group, Brunel University London, UK
allan.tucker@brunel.ac.uk
http://www.ida-research.net

Abstract. Healthcare data holds huge societal and monetary value. It contains information about how disease manifests within populations over time, and therefore could be used to improve public health dramatically. To the growing AI in health industry, this data offers huge potential in generating markets for new technologies in healthcare. However, primary care data is extremely sensitive. It contains data on individuals that is of a highly personal nature. As a result, many countries are reluctant to release this resource. This paper explores some key issues in the use of synthetic data as a substitute for real primary care data: Handling the complexities of real world data to transparently capture realistic distributions and relationships, modelling time, and minimising the matching of real patients to synthetic datapoints. We show that if the correct modelling approaches are used, then transparency and trust can be ensured in the underlying distributions and relationships of the resulting synthetic datasets. What is more, these datasets offer a strong level of privacy through lower risks of identifying real patients.

Keywords: Synthetic Data · HealthCare Data · Bayesian Networks.

1 Introduction

Health care data encodes vast amounts of individual patients' visits over years of their life. It represents a detailed if noisy and uneven record of an entire population including cases of many different types of disease. If this data were to be freely available it would clearly enrich society with respect to our knowledge of disease and population health. However, there are convincing reasons to protect this data. People are generally very wary of enabling their personal data, including their primary care information, from being made available without any protection. The General Data Protection Regulations, implemented in 2018 [4], aims to protect individuals from their personal data being made public (or being unknowingly released to companies or institutions). As a result, there is a demand for the generation of synthetic data: data that mirrors many of the characteristics of real Ground Truth (GT) data with similar distributions and relationships but made up of purely simulated patients. These datasets would offer the ability to train and validate many of the state-of-the-art machine learning models that are emerging, which in turn should lead to better detection and managing of many different diseases.

Previously, there have been a number of key approaches to working with synthetic data. The concept of k-anonymisation [10] works with the idea of measuring how likely it is to identify an individual from a small population, e-differential privacy [9] explores how aggregates of data can be released without identifying individuals from multiple requests of samples, generative adversarial networks have been used to build highly parameterised models from large datasets [11], [14] whilst PrivBayes highlighted the importance of transparency of the underlying model as well as the concept of adding noise to ensure no individual can be re-identified [7]. We agree that underlying models must be transparent so that there is confidence and trust in the generated data and as a result, we focus on the use of graphical model approaches based on our earlier work [1]. If a GAN is used to generate data where knowledge of the underlying dependencies are not clear, then there are risks that biases, incorrect dependencies or even prejudices can be introduced [15].

In this paper we explore three fundamental questions: Firstly, can probabilistic graphical models be used to capture key distributions and relationships to generate realistic synthetic data? Secondly, can they be extended to longitudinal data such as is common in primary care settings? Finally, does the generated synthetic data protect against the identification of real patients and their sensitive features? In the next section we explore the methods, datasets and results of our experiments that explore these questions, before concluding.

2 Methods and Results

2.1 Datasets

MIMIC: For the first part of this paper we will explore the potential of Bayesian Networks (BNs) for modelling and generating synthetic data on the MIMIC III dataset [2]. MIMIC III is a publicly available dataset that records details of the stay of a patient. It contains general information about the subjects such as the age, religion, ethnicity, type of healthcare insurance. It also contains clinical information, such as the diagnoses which are represented by ICD-9 codes among the stay in the hospital. The dataset has a total number of 47,764 observations with 36,243 different subjects. All numerical data was discretised into 5 states using a frequency based approach ensuring a similar number of each state for all features.

CPRD: In order to demonstrate the ability of probabilistic graphical models to deal with temporal patient data we use the Clinical Practice Research Datalink (CPRD Aurum Database) [3]. CPRD primary care data cover 21% of the UK population and include over over 17,400 clinical event types across patients with 25% of the patient data tracing back at least 20 years. We will focus on two key temporal features related to blood pressure, namely the Systolic

Blood Pressure (SBP) and Diastolic Blood Pressure (DBP). This will be used to see if the BN modelling methods used on the MIMIC dataset can be extended to generate high-fidelity synthetic data reflecting the temporal characteristics.

2.2 Modelling MIMIC data with Bayesian Networks

A Bayesian Network (BN) encodes the joint distribution of a dataset using a combination of a graphical structure that represents conditional independence between features and local conditional probability distributions [5] (see Figure 1a for an example). They facilitate the integration of expert knowledge and data, and can handle missing data naturally. Inference can be used to extract posterior probabilities over sets of features given some observations. This means that they can be used for classification (Figure 1b) and prediction. They are also generative models and can be used to generate samples of data based on the underlying distributions and independencies. What is more, they can be inferred from data using a number of different approaches including score-and-search methods [12] or contraint based methods [5]. An extension of the BN is the Dynamic Bayesian Network (DBN) which models time-series (Figure 1c) and the Hidden Markov Model which encodes an unmeasured latent process [6] (Figure 1d).

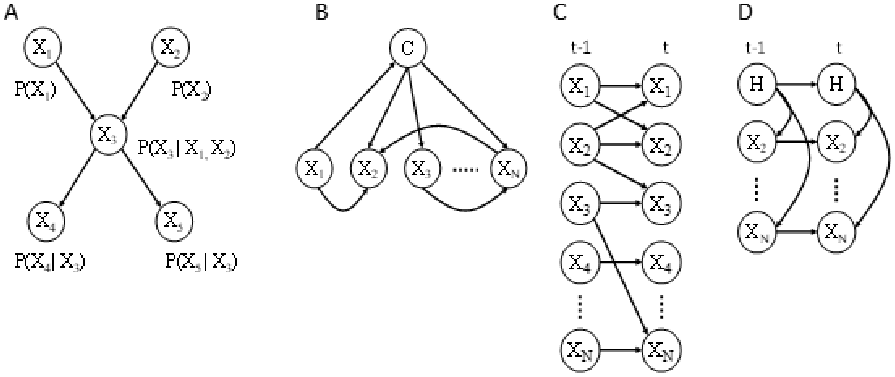


Fig. 1. Example A-Bayesian Network, B- Naive Bayes Classifier, C - Dynamic Bayesian Network and D- Hidden Markov Model

In order to test the BN framework on the MIMIC data, we used a well known BN learning algorithm that can deal with missing data known as Structural Expectation Maximization (SEM) to infer the structure and parameters of the model [13]. A key advantage of the BN is that it models the data in a transparent way where relationships between variables can be hard-coded or removed to influence learning.

Figure 2a shows the resulting structure. Two relationships that were known to exist (between blood disease and infections, and between age and circulatory conditions) were manually added. We then used this parameterised model to generate data synthetic data and compare the correlations (Figure 2b and c) and distributions (Figure 3) to the GT data. It can be seen that both the correlations and distributions are extremely similar to the GT data. We applied Kullbaeck Liebler tests and found no significant difference between all distributions. We wanted to see how this modelling technique could be extended to temporal data that is common in many health datasets by using the CPRD data.

2.3 Modelling Time

We exploited a natural extension of the BN known as the Dynamic Bayesian Network (DBN) which allows model structures over discrete time slices. See Figure 4a for an example DBN that includes a hidden variable. In fact, the Hidden Markov Model (HMM) can be seen as a special case of DBN with fixed structure and a single hidden variable that models an underlying and unobserved process.

We used a hidden variable with 4 hidden states to capture the dynamics of the relationship between SBP and DBP. Figure 4b shows the resulting state transition diagram that has been inferred from the data (again using the SEM algorithm). The plots in Figure 4c show a number of different characteristic comparisons between the original time-series ground-truth (GT) and the generated synthetic time-series. Firstly, in the top row can be seen a simple comparison of values for each patient for DBP, for SBP, and for the difference between DBP and SBP. All of these have a tight correlation. On the bottom row can be seen comparisons of temporal characteristics, namely the Auto-Correlation (ACF) per patient for DBP and SBP and also the Cross Correlation (CCF) between the two variables per patient. The ACFs show slightly skewed predictions where the synthetic data is often slightly lower than the ground truth, whereas the CCFs show very tight prediction values. These results imply that a realistic time-series of synthetic patient data can be generated using DBNs.

2.4 Risks of Matching Real Patients to Synthetic Data

Synthetic data offers an ideal way to allow the sharing of data that captures many of the characteristics of real patient data but without any of the privacy concerns. However, there is still a risk that synthetic data can be used to match individuals to simlar synthetic data and infer personal information about them. For example, if someone has access to some limited ground truth data about an individual and that information results in the individual being matched to similar outlying synthetic datapoints, then other more personal information may be inferred.

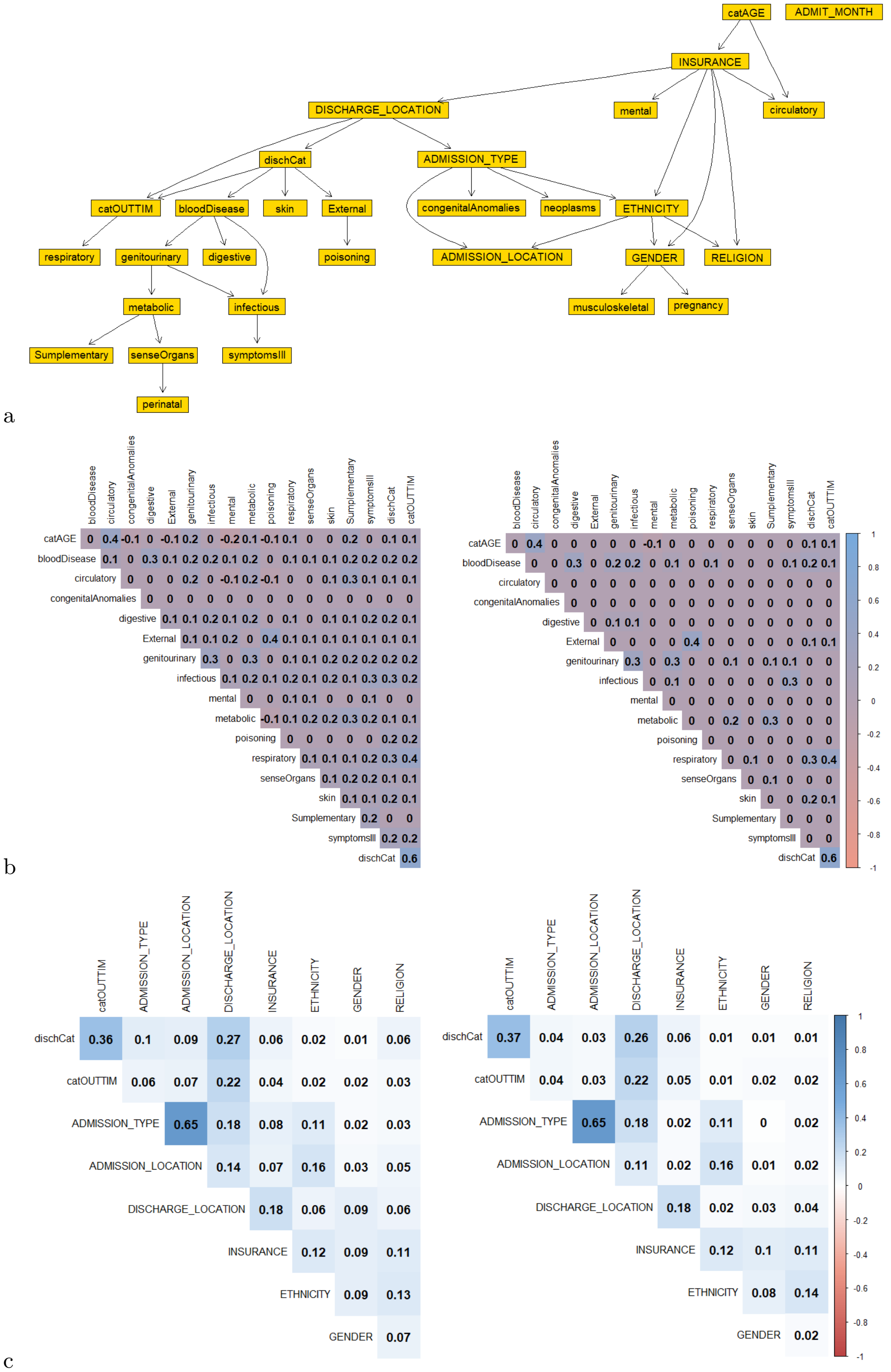


Fig. 2. a) the MIMIC BN and b/c) correlation matrices for ground truth (left) compared to synthetic data (right)

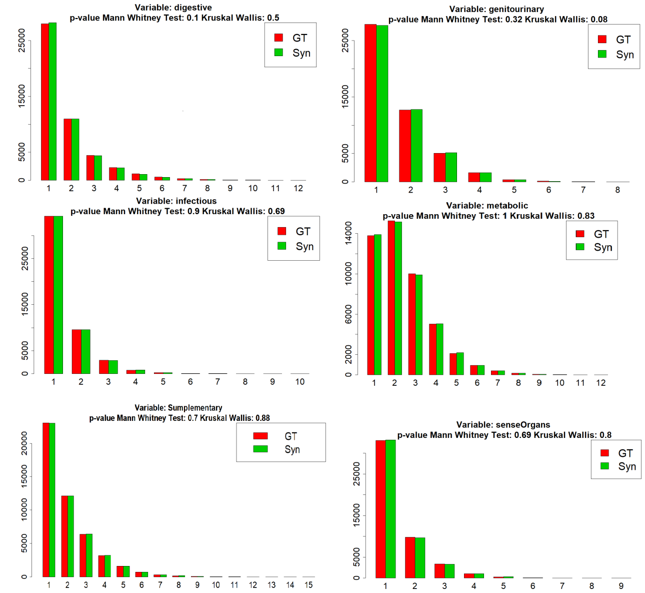


Fig. 3. Synthetic Data Distributions comparing frequencies for GT in red and synthetic in green

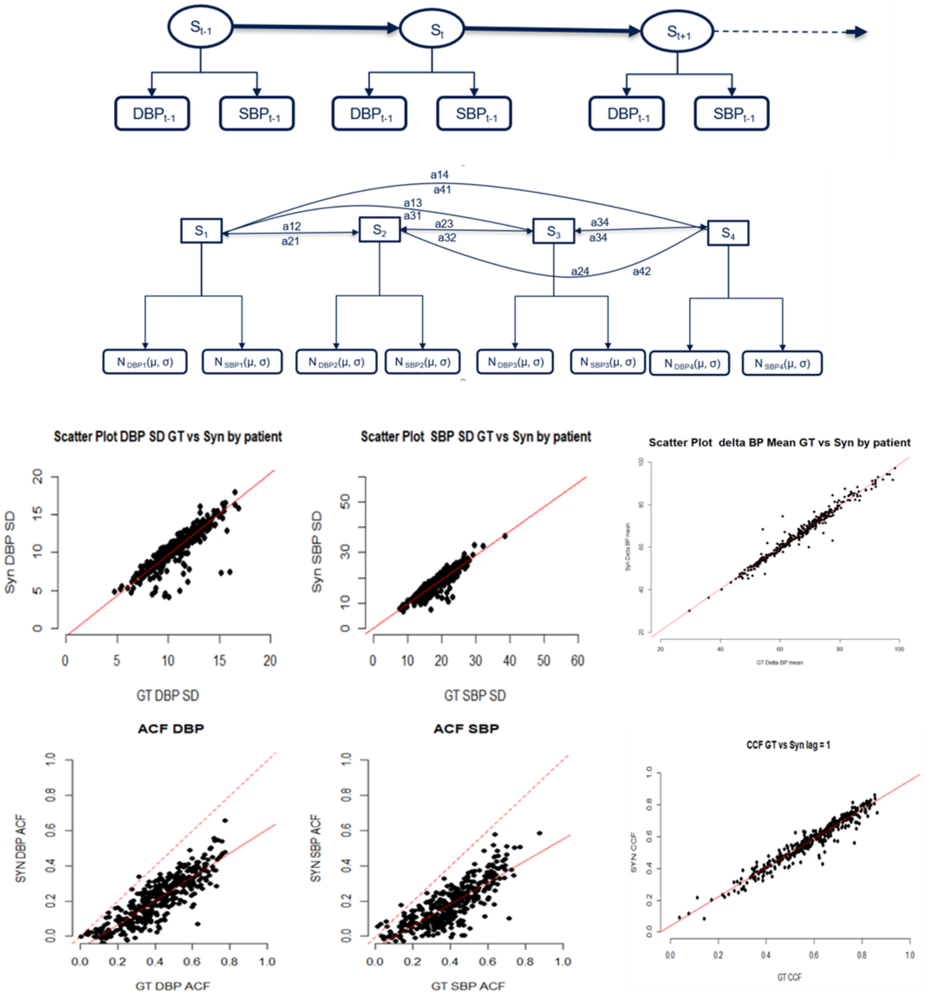


Fig. 4. DBN structure / state transitions and Resulting Data Characteristics

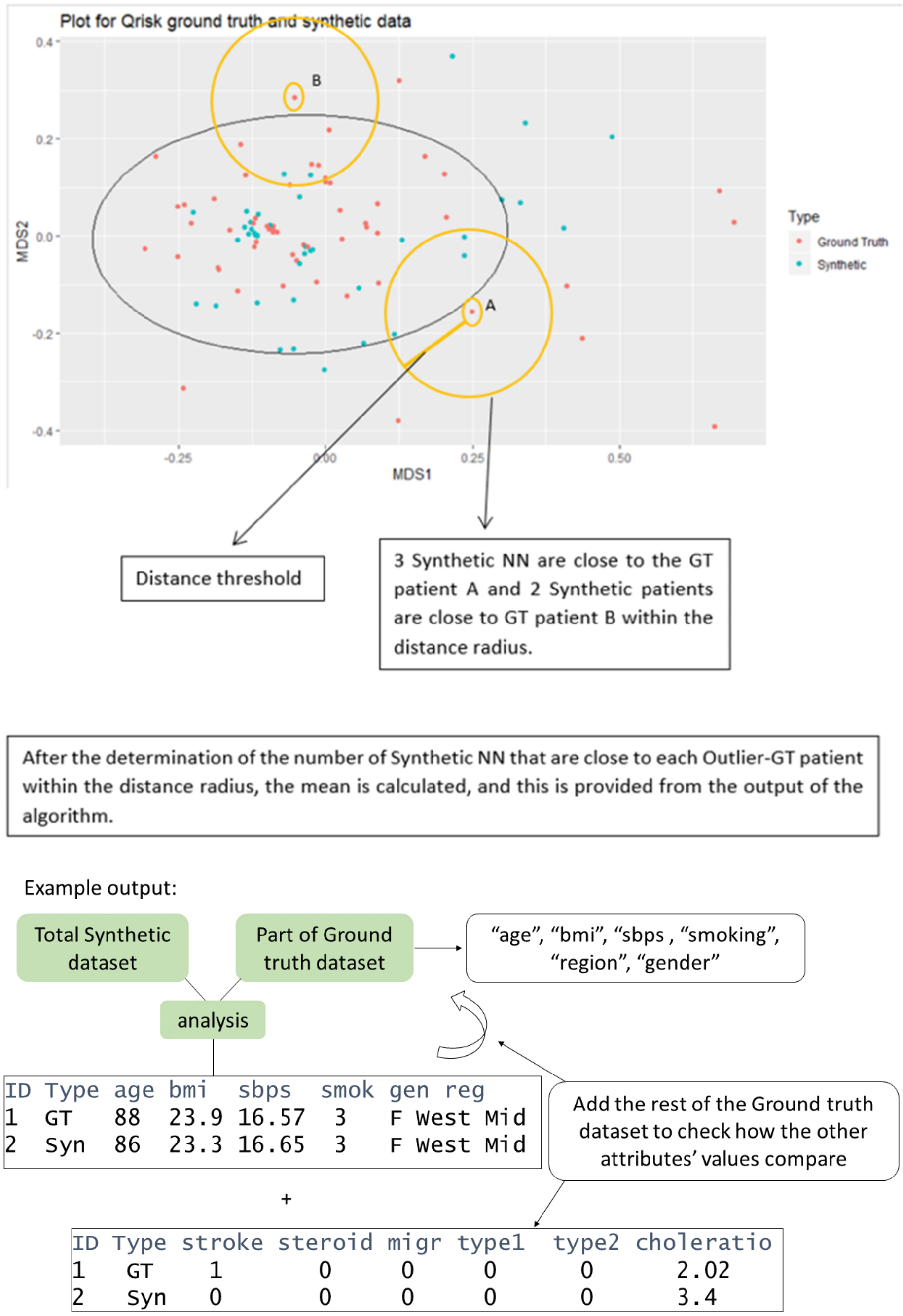
We explored this by looking at outliers in the Ground Truth (GT) data - those GT patients who have a small number of ”nearest neighbours” which are significantly separate from the rest of the population of data. We did this by repeatedly sampling data from a mix of GT data (red in top of Figure 5) and synthetic data (blue). Synthetic data is generated using the BN methodology described earlier on CPRD data. Outlying ground truth datapoints (see A and B in top of Figure 5) are then identified and these are used to explore the nearest neighbours that are from the synthetic data.

First we calculated the number of outlying ground truth datapoints that contain a single significant nearest neighbour from the synthetic dataset. This would mean that many of the characteristics of the synthetic datapoint could also be characteristics of the ground truth and risks the inferring of personal information. For example, in Figure 5 (bottom) we can see how 6 attributes that are known about a real individual can be used to match them to a synthetic datapoint, and therefore infer more attributes that are *mostly* the same.

By running 100 repeated samples to calculate these risks we found the results in Table 1 for 4 different combinations of starting ground truth features. 1. Firstly, notice how the number of GT outliers (first column) only slowly decreases as more GT features are made available. This implies that adding more knowledge of an individual to the attacker does not increase risk greatly. Secondly, notice that as the number of these available GT features increase, the number of single synthetic nearest neighbours decrease (column 3 representing the number of GT outliers with exactly one synthetic nearest neighbour and column 4 represents this as a proportion of all GT datapoints). Remember a very low number here may enable one to infer new personal information about the GT patient by exploring other features in the synthetic datapoint. The observed fall in number is expected as an attacker can identify more precisely which synthetic data points match our ground truth patients. In our worst case, When we have 12 features known about a GT patient (in the bottom row) then we can identify a single nearest synthetic neighbour in only 43.5 out of 6180 cases on average, which is a real but small risk (0.7%). What is more, in many of these cases, the synthetic datapoint is not identical to the GT patient in semantically substantial ways (such as the example in Figure 5 where the nearest synthetic datapoint is not a stroke sufferer unlike the GT patient).

3 Conclusions

This paper has explored some key issues when attempting to use synthetic data by learning models from sensitive healthcare data. It has carried out an empirical analysis on two real datasets using a probabilistic graphical modelling approach in the form of Bayesian networks for transparently capturing the key characteristics of data and emulating them in generated samples with very positive results. It has also extended this approach using dynamic Bayesian networks to model longitudinal data and has been successful in capturing the key temporal characteristics of blood pressure data. Finally, a set of simulations have been



Distance threshold

3 Synthetic NN are close to the GT patient A and 2 Synthetic patients are close to GT patient B within the distance radius.

After the determination of the number of Synthetic NN that are close to each Outlier-GT patient within the distance radius, the mean is calculated, and this is provided from the output of the algorithm.

Example output:

Total Synthetic dataset

Part of Ground truth dataset

“age”, “bmi”, “sbps , “smoking”, “region”, “gender”

Add the rest of the Ground truth dataset to check how the other attributes’ values compare

```
ID Type age bmi  sbps  smok gen reg
1  GT   88  23.9 16.57  3   F West Mid
2  Syn  86  23.3 16.65  3   F West Mid
```

```
ID Type stroke steroid migr type1  type2 choleratio
1   GT    1       0      0    0     0      2.02
2   Syn   0       0      0    0     0      3.4
```

Fig. 5. Simulating attacking Synthetic Data with limited Ground Truth Info

Table 1. Statistics for matching similar individuals from synthetic data

GT Attributes	Num Of GT Outliers	Num Of GT Patients in 10000 Sample	Num of *Single Synth* NN	Proportion of GT Outliers with Single Synth NN
age, smoking, region, gender, ethnicity, ckidney	396.3	6248	113.8	1.82%
age, smoking, region, gender, ethnicity, bmi, sbps, ckidney	377	6250	77.7	1.24%
age, smoking, region, gender, ethnicity, bmi, choleratio, sbp, sbps, ckidney	363.1	6289	48.22	0.76%
age, smoking, region, gender, ethnicity, bmi, sbps, ckidney, sle, atyantip, type1, streroid	363.4	6180	43.5	0.70%

carried out to extract risks of matching ground truth data to similar synthetic data using nearest neighbour analysis and this has shown that whilst the risk is real, it is remote and the ability to infer information on sensitive features is extremely difficult.

Whilst this paper has made some very positive findings there are still a number of issues that need to be explored. For example, the data that is used to train any model may be biased and it is important that transparent models are inspected carefully to check for this. This may be easy on relatively small models such as explored here but for those with many 100s of features this will become far more tricky. We have looked at only two datasets and considerably more experimentation is needed to ascertain proper statistics with confidence bounds for the risks involved in matching real patients to synthetic data.

References

1. Wang, Z. Myles, P. Tucker, A. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility Patient Privacy, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems, pp 126-131, 2019
2. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
3. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019 Mar 11. pii: dyz034.
4. https://gdpr-info.eu/
5. Spirtes P., Glymour C., Scheines R.(1993): Causation, Prediction and Search, Lecture Notes in Statistics 81, Springer-Verlag.

6. Rabiner, R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. of the IEEE, Vol.77, No.2,pp.257–286, 1989
7. Zhang, J., Cormode, G., Procopiuc, CM., Srivastava, D., Xiao, X. PrivBayes: Private Data Release via Bayesian Networks, SIGMOD'14,June 22–27, 2014, Snowbird, UT, USA
8. Patki, N., Wedge, R., Veeramachaneni, K. The Synthetic Data Vault, IEEE 3rd International Conference on Data Science and Advanced Analytics (DSAA), Volume: 1, Pages: 399-410, 2016, DOI Bookmark:10.1109/DSAA.2016.49
9. Snoke J., Slavkovi , A. pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity, arXiv:1805.09392v1, 2018
10. Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,10 (5), 2002; 571-588
11. Abay, N., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L. (2018). Privacy Preserving Synthetic Data Release Using Deep Learning. 510-526. 10.1007/978-3-030-10925-7, 31.
12. G. F. Cooper and E. Herskovits. A Bayesian methodfor the induction of probabilistic networks fromdata.Machine Learning, 9:309–347, 1992.
13. Friedman N (1997). Learning Belief Networks in the Presence of Missing Values and Hidden Variables. Proceedings of the 14th International Conference on Machine Learning, 125–133.
14. Xu, L. et al. Modeling tabular data using conditional GAN, 33rd Conference on Neural Information Processing Systems 2019.
15. Jia, S. Lansdall-Welfare, T. and Cristianini, N. Right for the Right Reason: Training Agnostic Networks, Advances in Intelligent Data Analysis XVII 17th International Symposium, IDA 2018.