This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

Evaluating a Longitudinal Synthetic Data Generator using Real World Data

Zhenchen Wang,
Puja Myles,
Anu Jain
CPRD, Medicines and Healthcare products Regulatory Agency
London, UK
zhenchen.wang@mhra.gov.uk

James L. Keidel,
Roberto Liddi,
Lucy Mackillop,
Carmelo Velardo
Sensyne Health
Oxford, UK
james.keidel@sensynehealth.com

Allan Tucker
Dept. of Computer Science
Intelligent Data Analysis Group
Brunel University London
London, UK
allan.tucker@brunel.ac.uk

***Abstract*—Synthetic data offer a number of advantages over using ground truth data when working with private and personal information about individuals. Firstly, the risk of identifying individuals is reduced considerably, which enables the sharing of data for analysis amongst more organisations. Secondly, the fine tuning of synthetic datapoints to suit particular modelling and analyses could help to build more suitable models that can avoid biases found in the original ground truth data.**

In this paper we explore how a probabilistic synthetic data generator can be used to model data with high enough fidelity that it can be used to develop and validate state-of-the-art machine learning models. In particular, we use a Bayesian network model trained on gestational diabetes data, generated from a mobile health app collected from a number of health trusts in the UK. These data are used to train and test an established machine learning model developed by Sensyne Health using real-world data, and the resulting performance is compared to performance on ground truth data. In addition, a clinical validation is undertaken to explore if human experts can differentiate real patients from synthetic ones.

We demonstrate that the Bayesian network synthetic data generator is able to mimic the ground truth closely enough to make it difficult for a human expert to distinguish between the two. We show that the data generator captures the interactions between features and the multivariate distributions close enough to enable classifiers to be inferred that imitate the key performance characteristics of models inferred from ground truth data. What is more, we demonstrate that the discovered misclassifications found when testing using the synthetic data, are as informative as when testing using ground truth data.

***Index Terms*—Synthetic Data, Bayesian Networks, Machine Learning, Diabetes**

I. Introduction

The increasing interest in synthetic patient data has been driven by a number of factors. These include concerns around patient privacy which may hinder data sharing, simulation of patient sub-groups or characteristics which may be missing in the underlying ground truth or 'real' data and boosting of sample sizes where the prevalence of relevant features or outcomes is rare. The work presented in this paper is motivated by the question of whether synthetic data could be used to validate machine learning algorithms for clinical decision making that are trained using real patient data and vice versa. This paper adds to the evidence base in achieving this goal based on experiments with real world data collected from a medical app developed by Sensyne Health.

There is now a growing interest in the development of approaches to generate fully synthetic data. For example, generative models such as Generative Adversarial Networks [9], or Bayesian Networks [8] can be inferred from data and synthetic data samples can be generated from the resulting models. Previously we have explored a synthetic data generation framework which uses Bayesian network analysis to learn complex clinical relationships and distributions from ground truth data and then generated high-fidelity synthetic data from them [6], [7]. Our previous work has focused on cross-sectional data and was assessed on machine learning models in-house. This paper reports on an extension of our methodology to generate synthetic data from a gestational diabetes monitoring app, *GDm-Health*™, including time-stamped blood glucose data.

Here we present two approaches to generating synthetic time-stamped data, one of which presents efficiencies in processing with a view to facilitate scalability at pace. We include the following experiments to evaluate the performance of our synthetic (SYN) data generation methods: univariate and multivariate comparisons between ground truth (GT) and SYN data, including features which were not present in the version of the GT data used to generate SYN data; training a machine learning algorithm on the GT data and testing on the SYN data; training a Machine Learning algorithm on the SYN data and testing in prospectively collected real data from the GDm-Health app; a clinical validation to explore if human experts can differentiate real patients from synthetic ones. These experiments also enable us to explore the trade-offs between efficiency and fidelity.

II. Dataset

GDm-Health is software that allows clinicians to remotely monitor glycaemic control in pregnant women affected by Gestational diabetes mellitus (GDM). It consists of a mobile app for patient use and a web interface for clinicians. Typically, the mobile app is connected via Bluetooth to the patient's blood glucose (BG) monitor, so that readings are automatically

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

uploaded for review by clinicians. In some cases, patients use non-bluetooth BG monitors and manually input the readings.

In addition to BG reading history, the *GDm-Health* dataset includes demographic information about the patient, prescription information (e.g., insulin, metformin), how the patient's condition is being managed (medication/diet only/both) and metrics of clinician-patient interaction. The use of *GDm-Health* can be advised at any time during the pregnancy, although generally occurs around the 24th week. The system is generally used until the end of the pregnancy period. Depending on clinician assessment of the patient, clinicians recommend a set number of BG readings to be taken each week. When the patient takes a reading, they are asked to supply information about whether the reading is pre- or post-prandial, and which meal the reading relates to. Because normal BG values fluctuate greatly during the day as a function of mealtimes and carbohydrate intake, this information is essential for understanding the significance of the reading.

The data used as the basis for synthetic data generation were extracted from Sensyne's database in August 2019. This dataset contains 2660 patients and 439,567 BG readings. However, after filtering out patients unsuitable for predictive modelling (e.g., those that have not been diagnosed with GDM, those that have not actually used the app and therefore have no BG history) there were 1109 patients who had not been prescribed medication and 471 that had prescriptions. The patients' demographic data and BG reading history formed the basis of the datasets used to train models that predict whether a patient will be prescribed medication. For detailed information about the generation of the features please see [1]. The models are trained on a set of features based on aggregate statistics computed over 7 days of data.

The variables used to define a patient entity include

Variable name	Details
patient-id	UUID format
age	Age in months
edd	Estimated Delivery Date
parity	A number of births of fetuses of 24weeks or older regardless of the outcome
gravidity	A number representing the number of pregnancies regardless of the outcome
induced	A boolean variable indicating if the labor was induced
complications	A selection of complications at birth
dayreadings	How many readings per day were prescribed
weekreadings	how many readings per week were prescribed
diabetes	Type of diabetes
expected-babies	Number of babies expected during the current pregnancy
height	Height in cm
weight	Weight in kg

III. Synthetic Data Generation Methods

A. Probabilistic Graphical Models

Probabilistic Graphical Models are generative models that capture the underlying joint distribution of multiple interacting variables [4]. They do this efficiently by making conditional independence assumptions about the variables which are represented in a directed graph. By encoding local probability distributions at each node in the graph, this family of models can be used to perform inference about new data, or to generate samples of new data. Bayesian Networks (BNs) are a form of probabilistic graphical model that can be inferred from data (both structure and parameters). Furthermore, the inference allows data to be sampled from the model under different conditions. For example, data can be sampled but only for people with high-blood pressure, or only people over a specific age. In this way, synthetic data can be generated that suits a specific cohort. Temporal information can be incorporated into these models in a number of ways: extending into the time domain using dynamic BNs [2], or by including explicit temporal nodes within the structure that capture different temporal behaviours such as trends. Figure 1 shows an example BN with four nodes where each probability distribution is conditioned upon the parents of the node.

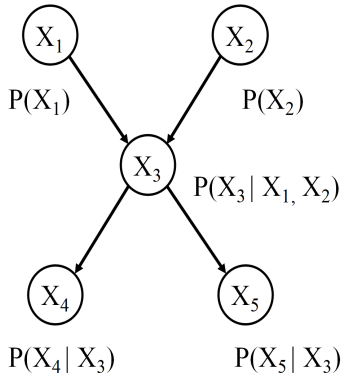

Fig. 1. Bayesian Network with Four Nodes and associated local distributions

B. Incorporating Trend Information

For these experiments we used a standard BN approach by inferring both structure and parameters from the ground truth data. A BN is learnt from the data described in the previous section. Synthetic data is then sampled from this BN using a standard logic sampling approach [3]. We use the *bnlearn* package in the statistical package *R* for implementing our experiments [5]. As previously stated, we can bias the sampling of data to patients with particular characteristics or symptoms. For example, we can generate samples for patients living in a certain region, or those greater than a threshold age. This could be used to compensate for known biases within some ground truth datasets. For the longitudinal data (the blood glucose readings), we repeatedly sampled for each patient and use Euclidean distance to identify samples that are close to the original ground truth time-series. These form the synthetic time-series. The synthetic instance must meet two criteria:

1) the same trend to previous sequential value as to GT trend, i.e. up/down/level
2) the closest to the ground truth instance by Euclidean distance

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

C. MARK & HERMES

We have explored two approaches to generating the synthetic data based on practical constraints. In the BG reading data processing, we found that $t+1$ trend comparison requires us to build a new BN for each sequence and this repetitive task of learning and fitting can be an overhead during production. The first approach which yielded a synthetic dataset we call the 'MARK' version, used $t+1$ trend comparison so that a new BN is built to represent each sequence in the series. In the second approach, which seeks to make efficiency gains in processing, we used $t+2$ trend comparison so that a new BN is built for every two sequences in the series. The resulting synthetic dataset is termed the 'HERMES' version.

In HERMES, we also filled the BG value, i.e. the value at $t+1$. Here we are using a jitter function $f(v) = v + z/50$, to introduce minor noise to the ground truth value, where v is the ground truth value at $t+1$, and z is the difference between adjacent values at t and $t+2$. Despite the jitter function in which the associated overhead is trivial, the efficiency HERMES gained is half of the overhead as MARK during BN learning and fitting and this results in the BG readings being generated faster than MARK in practice.

IV. Experiments

In this paper we wish to explore how BNs can be used to model time-series data that displays complex diurnal variations using as an exemplar, blood glucose levels in gestational diabetes. We investigate whether the synthetic data generated using the approaches outlined in the Methods can be used to train and validate machine learning models with enough accuracy to be considered as good as (or better than) working directly with ground truth data. This means exploring not just accuracy as a metric but also the potential underlying biases in a model by carrying out a full sensitivity analysis.

A. Receiver Operator Characteristic and Precision / Recall Analysis

Synthetic data of the same sample size as the ground truth data were generated using the methods described in the previous section and the two approaches- MARK and HERMES. After comparing the univariate distributions of each feature for MARK and HERMES with those from the ground truth (from now on *GDM-1*), we explore the behaviour of one machine learning classifier selected from the work of [1] trained and validated on this synthetic data. We assess its performance using standard sensitivity analysis with re-sampling. This allows us to calculate the sensitivity, specificity, precision and recall of the classifier. We then compare these performance statistics with those on the ground truth data (*GDM-1*) as well as a further extended ground truth dataset collected from the same digital-health application (during the period August 2019-October 2020). This last dataset (from now on *GDM-2*) includes more NHS Trusts and a considerably larger number of participants (7,733). This enables us to see how close the model and performance statistics of synthetic data are to a model trained and tested on a more substantial and fully independent ground truth dataset.

B. Multidimensional Scaling Visualisations

Whilst the sensitivity analysis will tell us how a machine learning classifier compares when trained on real or synthetic data it does not tell us how the individual test data points compare and in particular how false positives (FPs) and false negatives (FNs) compare. We want to explore in detail the general decision boundaries of the classifiers when trained on synthetic and ground truth data and to do this we performed multidimensional scaling analysis to see how the misclassifications compared when observed on their respective regions of the input feature space. Ideally, we want to see a similar pattern of FPs and FNs for synthetic and ground truth models in terms of where they occur in the input feature space.

To this end, we ran 1,000 train-test iterations of the selected model using either GDM-1, MARK, or HERMES as the training set and GDM-2 as the test set. During the training phase of each iteration, we took a random sample of patients to generate measurable variation in the model outputs. We then tested the trained model on the entire GDM-2 dataset. To assess how similar the classification of these test data were between models trained on synthetic data and models trained on real-world data, we used a Spearman correlation to obtain the correlation coefficient between the vectors of average classifications.

C. Clinical Validation of Synthetic Datasets

In order to further validate the fidelity of the synthetic data to ensure it is indistinguishable to ground truth data we carried out an analysis whereby two clinical experts (one of whom is based within the UK Medicines and Healthcare Products Regulatory Agency, and provided a generalist medical perspective, while the second is a consultant obstetric physician based within Sensyne Health) were asked to inspect 30 patient records and to identify the ground truth from the synthetic. We randomly selected 10 ground truth and 20 synthetic records (10 from HERMES and and 10 from MARK) for inspection. We asked two clinicians to identify which data they thought were from real patients and which were synthetic.

V. Results

A. Analysis of feature distributions

After clinical inspection, we explored some simple distributional comparisons between features within the datasets, both Ground Truth (GDM-1) and Synthetic (MARK). Figure 2 shows these distributions the features used in the models tested here. There is substantial overlap between the synthetic and ground truth distributions, with no significant differences at the critical alpha of 0.008 observed in the distributions of the four continuous features as measured by the Kolmogorov-Smirnov test (Feature 1, $p = 0.25$; Feature 2, $p = 0.83$; Feature 4, $p = 0.051$; Feature 5, $p = 0.61$), and no significant difference in the two binary features (Feature 3 and the Target)

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

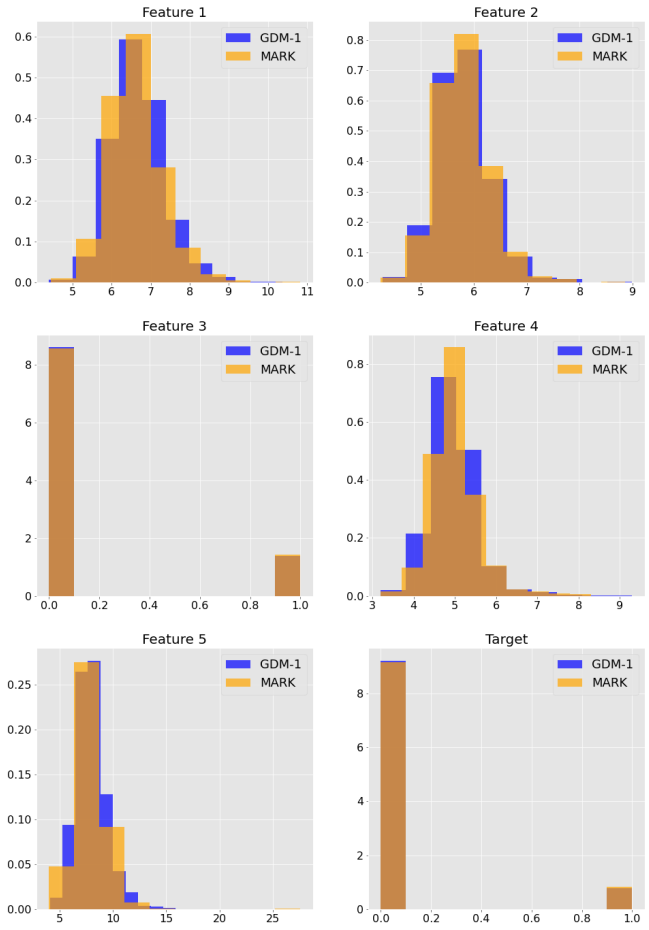

Fig. 2. Distribution Comparison of Ground Truth and Synthetic Data

as measured by a z-test for proportions (Feature 3, $p = 0.73$; Target, $p = 0.5$). For HERMES (which involved efficiency savings) the similarity between the synthetic and ground truth was less pronounced. Significant differences were observed in Features 1, 3, 4 and 5 (all $p < 0.001$).

B. Machine learning model comparison

We now turn to the behaviour of the synthetic datasets compared to the ground truth when carrying out training and testing of machine learning classifiers. These models are designed to predict medical intervention (specifically, prescription of medication to aid BG control) in the data. We trained and tested 100 models on each of GDM-1, MARK and HERMES, selecting a different random 20% of the data for testing and using the rest for training on each iteration. We show the resulting average ROC curves and PR curves in Figures 3 and 4, respectively. In 3 the average ROC curve for the ground truth (*GDM-1*) has the highest area underneath but this is closely followed by the synthetic dataset (MARK). The overlap is considerable and demonstrates no significant difference in using synthetic data to validate the models as compared to ground truth data (2-sample t-test of AUC values

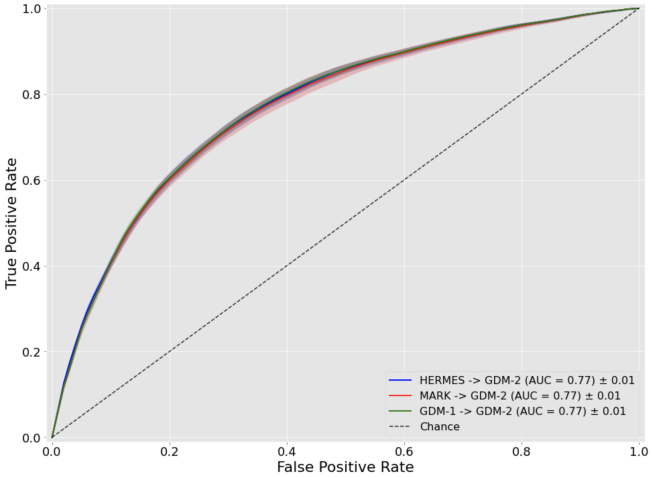

Fig. 3. ROC curves for Ground Truth and Synthetic Data Compared

$t(198) = 0.66$, $p = .51$). HERMES shows less convincing results with significantly lower AUC values than GDM-1 ($t(198) = 11.78, p < .001$) and MARK ($t(198) = 12.5, p < .001$). Interestingly, however, the average AUC when running 100 models trained on a sample (80%) of GDM-1, MARK or HERMES and testing on GDM-2 was highly similar across the three datasets (GDM-1: 0.773, MARK: 0.770, HERMES: 0.772), though the figures for GDM-1 and HERMES were significantly different statistically from MARK (GDM-1 vs. MARK $t(198) = 3.21$, $p = .001$; HERMES vs. MARK $t(198) = 2.27$, $p = .02$). Though the average AUCs were very similar, more pronounced differences were observed in the average percent correct for the three models when the logistic regression output was thresholded at 0.5 (GDM: 82% correct, MARK: 82% correct, HERMES: 79% correct).

We also explore the use of Precision Recall (PR) curves. This is especially important due to the the imbalanced nature of the data. Figure 4 shows a sample of PR curves for models that are trained on GDM-1, MARK and HERMES and tested on GDM-2. Notice that the variance in the curves for GDM is higher due to the differing training and testing samples as opposed to the curve for MARK which does not change.

C. MDS

Sensitivity Analysis helps us to capture the overall classification performance as well as the tradeoffs between Sensitivity / Specificity, and Precision / Recall. However, to better understand the detailed types of mis-classifications we explore a dimensionality reduction approach to see if the types of mis-classification are similar between the models trained and validated on the ground truth and the synthetic data. In particular, we use multidimensional scaling with Euclidean distance as the distance metric. Figure 5 shows the class allocations for the ground truth data (a) along with the associated mis-classifications (b), as well as the class allocations for the synthetic data (c) and their associated mis-classification (d).

Notice that the overall shape of the data is similar in the ground truth (a) and synthetic data (c) and that the overlap of

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

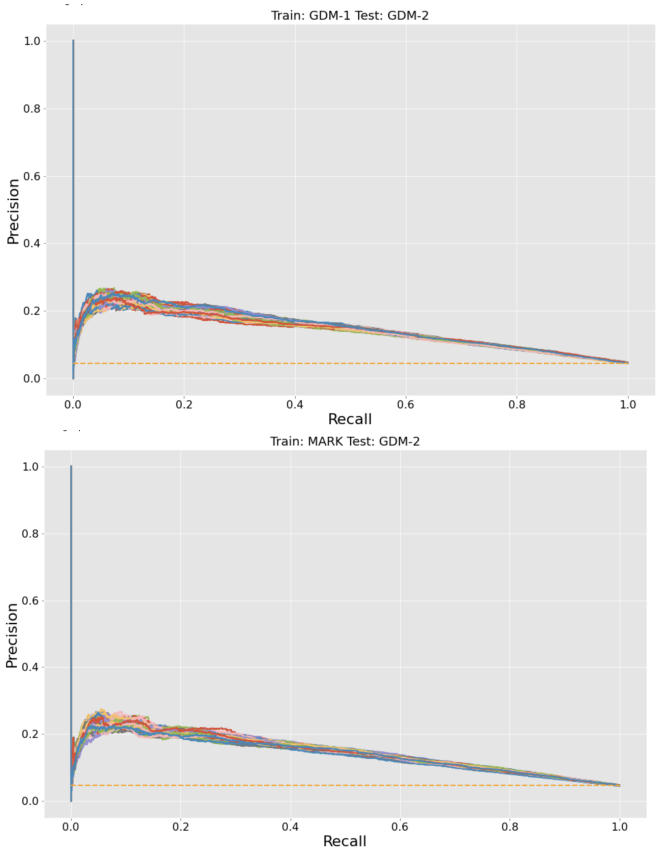

Fig. 4. PR curves for Ground Truth and Synthetic Data Compared

classes are also similar though the synthetic data appears to have slightly more densely clustered positive cases of patients taking medication. Note that the dataspace has been flipped along the x-axis as a result of the scaling procedure. Looking at (b) and (d) we can see how the true negatives, true positives, false negatives and false positives are distributed over the dataspace. Again, we see similar shapes of the data with each class of classification being somewhat similar, the only notable difference being the true positives being slightly more densely clustered as seen in (c).

To quantify the similarity of model response patterns at the individual item level, we trained 100 models on each of GDM-1, MARK and HERMES and tested them on GDM-2. We then calculated the phi coefficients between the individual vectors of binary responses from the models trained on GDM-1 and MARK, and GDM-1 and HERMES. We then converted these values to be normally distributed via the Fisher z-transform and carried out a 2-sample t-test, which showed that the response similarities between models trained on GDM-1 and MARK were significantly higher than those between GDM-1 and HERMES ($t(198) = 8.10, p < .001$).

D. Clinical Validation

Table II shows the results of the clinicians identification of synthetic patients compared to the real. It can be seen that of the 30 synthetic patient records presented to the two

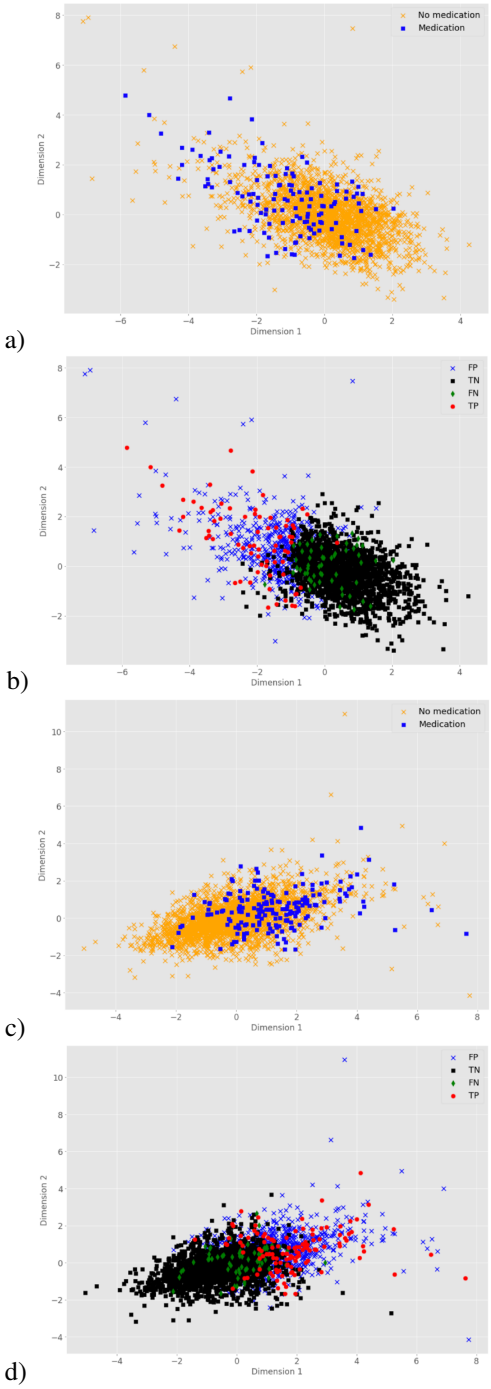

Fig. 5. Multidimensional Scale Plots - Class allocation and Classification errors for (a,b) Ground Truth Data (GDM-1) and (c,d) Synthetic Data (*Mark*)

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/CBMS52027.2021.00074 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)

experts, only 6 (clinician A) and 7 (clinician B) were correctly identified as synthetic, whilst 3 and 8 were correctly identified as real. 14 and 13 were incorrectly identified as ground truth and 7 and 2 were incorrectly identified as synthetic. These results indicate that it was not easy for the experts to distinguish synthetic from ground truth. The Table also includes a breakdown of the two synthetic datasets indicating that *Hermes* was slightly more difficult to distinguish.

TABLE I
CONFUSION MATRIX OF CLINICAL VALIDATION FOR IDENTIFYING SYNTHETIC FROM REAL RECORDS - CLINICIAN A

Total = 30	Predicted Synth	Predicted GT	Total
Actual Synthetic	6	14	20
Actual GT	7	3	10
Breakdown			
HERMES (actual synthetic)	2	8	10
MARK (actual synthetic)	4	6	10
Ground Truth (actual GT)	7	3	10

TABLE II
CONFUSION MATRIX OF CLINICAL VALIDATION FOR IDENTIFYING SYNTHETIC FROM REAL RECORDS - CLINICIAN B

Total = 30	Predicted Synth	Predicted GT	Total
Actual Synthetic	7	13	20
Actual GT	2	8	10
Breakdown			
HERMES (actual synthetic)	3	7	10
MARK (actual synthetic)	3	7	10
Ground Truth (actual GT)	3	7	10

VI. CONCLUSIONS AND FUTURE WORK

We have demonstrated experimentally that our synthetic data generation approach is able to handle complex time-series data like blood glucose measurements which exhibit diurnal variation in response to meals, dietary intervention, and medications. Both clinical and statistical validation tests demonstrated that these synthetic datasets can capture underlying ground truth data characteristics with a high degree of fidelity. Furthermore, this high fidelity is maintained even when comparing derived features which were not included in, but only derived from the original ground truth dataset for the purpose of model training. The high fidelity is maintained also when comparing to real world data collected prospectively. These results give us confidence that our synthetic data could be used to train and validate future models. What is more, in cases where ground truth data are limited, one could use synthetic data to boost limited datasets where there are undersampled cases.

We also examined trade-offs between efficiency and fidelity by comparing two approaches to synthetic data generation for time-series data, MARK (which considers each sequence in the time series) and HERMES (which considers every second sequence in the time series and thus, is computationally more efficient). We observed slightly lower fidelity with HERMES but none of our tests suggested a clinically or statistically meaningful loss of fidelity.

In this paper we have focused on replicating the characteristics of the ground truth data in the synthetic data. However, as stated earlier, BN approaches offer the opportunity to bias the sampling of data to patients with specific characteristics. This conditional generation of synthetic data could be used to compensate for known biases in ground truth data or to examine the effects of biases which are of concern to policy makers, on the performance of machine learning algorithms. Future work will focus on such conditional generation as well as exploring other approaches to time-series modelling such as dynamic Bayesian networks [2], the inclusion of latent variables to improve fits of the distributions and to handle unmeasured effects [7], the use of Spatial Bayesian Networks [10] to model geospatial health effects regionally, and the exploration of data drift with respect to model fit over time.

VII. ACKNOWLEDGEMNTS

The work presented in this paper was funded by NHSX using the synthetic data generation and evaluation framework developed under a grant awarded to the Medicines and Healthcare products Regulatory Agency (MHRA) by The Department for Business, Energy and Industrial Strategy (BEIS) and managed by Innovate UK.

REFERENCES

[1] Velardo C, Clifton D, Hamblin S, Khan R, Tarassenko L, Mackillop L, Towards a multivariate prediction model of pharmacological treatment for women with gestational diabetes mellitus, Journal of Medical Internet Research, 17/01/2021:21435, http://doi.org/10.2196/21435

[2] Proceedings of the Fifth Workshop on Data Science for Social Good (SoGood 2020), Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020), Ghent, Belgium, September 14–18, 2020, Springer. (in press)

[3] Max Henrion, Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling, Machine Intelligence and Pattern Recognition, Vol 5, 1988, Pages 149-163, https://doi.org/10.1016/B978-0-444-70396-5.50019-4.

[4] Spirtes, P., Glymour, C., and Scheines, R. (2001). Causation, Prediction, and Search, 2nd edn. Cambridge, MA: MIT Press.

[5] M Scutari, Learning Bayesian Networks with the bnlearn R Package, Journal of Statistical Software 35 (3), 1-22, 2010

[6] Wang, Z, Myles, P, Tucker, A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2020; 1– 33. https://doi.org/10.1111/coin.12427

[7] Tucker, A., Wang, Z., Rotalinti, Y., Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. npj Digit. Med. 3, 147 (2020). https://doi.org/10.1038/s41746-020-00353-9

[8] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. ACM Trans. Database Syst. 42, 4, Article 25 (November 2017), 41 pages. DOI:https://doi.org/10.1145/3134428

[9] J. Yoon, L. N. Drumright and M. van der Schaar, "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2378-2388, Aug. 2020, doi: 10.1109/JBHI.2020.2980262.

[10] Trifonova, N. Kenny, A. Maxwell, D. Duplisea, D. Fernandes, J. Tucker, A. Spatio-temporal Bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology, Ecological Informatics, 30 (2015), pp. 142-158