MRC-GAN Deliverable 1.1: Generative Model for Causality Learning and Synthetic Data

Abstract This report describes the first stage of the work in the MRC-GAN project. We have improved the core methods to learn causal structures from real-world clinical data based on generative adversarial learning. The trained generative adversarial model also generates a set of synthetic health records which hold the same distribution as the real data. Our model is non-parametric and hence is not restricted by any pre-fixed model equations. We can handle mixed continuous and discrete data variables and utilise measurements over a time series. The model can be trained from data with missing measurements. A privacy preserving mechanism has also been made available to the training process to protect the training data. Our initial evaluation of the model on the simulation benchmarks achieved good results in comparisons with the state-of-the-art continuous optimisation based causality learning models.

The report also covers the work on the preparation and pre-processing of the real world clinical data. We have selected the data by carefully defining the inclusion and exclusion criteria based on clinical knowledge, and have identified a set of target variables of interest. Our initial experiments showed an encouraging trend of convergence from the synthetic data towards the real data.

The report is organised as follows: Section 1 gives the general background of the project. Section 2 provides an overview of the generative adversarial causality learning approach; Section 3 gives more details on the model design and implementation; and Section 4 shows our evaluation results so far.

1. Project Background

Project context Health data contains important knowledge that enables clinical research to assess treatment effect in real world settings. However, there are significant limitations: they are typically imbalanced across different population, diseases and interventions; they contain bias, noise and missing measurements; the process of removing patient identifiable information may take significant time and effort, which also faces the risk of deleting valuable information from the original data.

The MRC-GAN project is designed to investigate an alternative approach to support clinical research through the use of synthetic data. We will study the feasibility of creating synthetic health data with the help of the latest AI technology, namely the generative AI, to generate synthetic data that preserve the same value for research as real data. To answer the clinical questions about treatment effect, our clinical trial emulation will run a “virtual trial” on the synthetic data.

The goal of MRC-GAN is to study the feasibility of this new approach through a specific use case in the context of Type 2 diabetes mellites (T2DM). Through training the AI model with the SCI Diabetes data on Safe Haven, we aim to create their synthetic version and then we will carry out a virtual trial to assess the effect of a target medicine. We will assess this new approach by comparing the outcomes from the trial emulation with the real ones.

Benefit and public interests The potential benefit of the new approach with synthetic health data include the following aspects:

- Quality of the data: This approach can generate synthetic data to address the problems that lie within the real data, including bias, data imbalance, noise and missing measurements.
- Research agenda: Trial emulation can be tailored to create virtual populations to address target clinical questions via clinical trial emulations, which would otherwise be impossible to address in real-world trials. For example, the clinical questions that are associated with underrepresented populations of children, older adults, and patients with multi-morbidities and polypharmacy – these people are commonly excluded in clinical trials. In fact, Randomised Controlled Trials (RCTs) are far from being able to answer all clinical questions. In many situations conducting RCTs with real patients is logistically challenging or unethical due to their potentially harmful nature. This leaves a significant knowledge gap. For example, a major drawback of the current clinical guidelines is that most of them only address single diseases with very few recommendations for multi-morbidity management despite the high prevalence.
- Privacy protection: Compared with anonymised real data (which contains reduced information about real patients), this new approach can generate synthetic data in unlimited volume while containing no

identifiable information about real individuals. Hence, this is in a much better position to overcome legal barriers in data protection and sharing.

Overall, this research will assess the potentials of a new way to provide real-world evidence to support future clinical research with better quality and privacy protection. This will open doors for further research in this direction, which could ultimately bring a landscape change to revolutionise future biomedical and health research by broadening its research agenda, liberating its restrictions, saving cost and time. Research in this direction will speed up new timelines for treatment discovery, address increasingly complex healthcare landscape in elderly population and multi-morbidity, and potentially transform regulatory and policy making process.

Note, we do not claim that this single research will provide all the solutions and answers to this synthetic data approach and yield impact. Rather, it is a feasibility study to bring first set of evidence to assess whether this can be achieved by leveraging the latest AI technology – see more details below in the Methodology.

Methodology This research is a feasibility study of the above mentioned synthetic data based approach for clinical trial emulation. To test the feasibility, the primary research questions include:

- Can we generate synthetic data that preserve the same value for research as real-world health data?
- Can we perform virtual clinical trial emulations by discovering correct causal relations from the synthetic data?

To answer the first research question on synthetic data generation, we will leverage the latest AI advance in creating synthetic data with generative models. The idea is to train the AI model with real data from Safe Haven, namely the SCI Diabetes. The output, which is the synthetic data, is expected to be statistically equivalent to the real data without information loss (e.g. it will contain all the variables at in the real dataset). However, the privacy preserving mechanism lying within the AI model will make sure that the synthetic data contain no information about any real patient. In other words, no real patient can be identified from the synthetic data. The work will be based on an existing AI model from the team that has been tested on other datasets. The quality of the synthetic data will be assessed with established metrics.

To answer the second question on virtual clinical trial emulations, we will focus on a specific T2DM use case within the scope of this feasibility study. The idea is to run a confirmatory study. We will emulate an established clinical trial, namely, the LEAD program (Liraglutide Effect and Action in Diabetes) so that we can check the results. The trial targets the effect of Liraglutide, a GLP-1 receptor agonist. We will use the synthetic data to carry out the trial emulation.

SCI Diabetes in Safe Haven is a good dataset to use in this study. This is an inclusive national dataset of individuals with diabetes containing a broad range of longitudinal demographic, phenotypic, biochemical and screening data. There are approximately 300K individuals with diabetes. Over 3K individuals with MODY (Maturity-onset diabetes of the young) are recorded with certainty (genetic information) along with records of individuals with negative genetic test results.

2. Generative models and causality learning

To run trial emulations with either observational or synthetic data, we need to build a simulation model that captures causal relations between multiple variables. This is done through causality learning from data. Within the context of this research project, we experiment with the use of generative AI models for causality learning.

The use of generative AI is well supported by the recent track record of generative AI models such as the generative adversarial network (GAN). GANs have achieved incredible performance by using neural networks to learn how to replicate real world data distributions. They are also very effective at inferring missing information from data – all these make them an attractive alternative to address issues in real-word data such as data imbalance, noisy and missing measurements.

Inferring and understanding causal relations from observational data is an important research topic nowadays. Bayesian Network plays an important role in describing causality [1,2]. Causality learning amounts to the discovery of Directed Acyclic Graphs (DAGs) from data as Bayesian Network structure can be represented with DAGs. In recent years we have witnessed a series of score function based approaches for causality learning [3-8] using the NOTEARS framework originally proposed in [7]. These methods have fundamentally transformed causal structure learning from combinatorial search into a solvable continuous optimization problem through the use of acyclicity constraints. However, these methods have some limitations. In particular, most of the current methods have only involved reconstruction loss in the causal inference from data, which is equivalent to the estimation of maximum likelihood. Recent research has shown that likelihood is not an adequate training objective for generating realistic sample since it encourages the model to capture all data modes by distributing the probability mass across the entire training data space, which leads to a mean seeking mode with blurry results. It is well recognised that the adversarial loss that corresponds to the minimization of reverse Kullback-Leibler divergence should be used to synthesize highly realistic data [9].

By taking the generative modelling approach, our method involves the use of adversarial loss as part of the objective function for causality learning in a generative modelling framework. We view causality learning as a process to construct a causal graph model (CGM) that generates equivalent data distribution to the real training data. The generative framework simultaneously learns causal structures while improving its data generation capability by minimizing the gap between the data distributions measured by the Wasserstein metric. Our initial experiments have already demonstrated a certain level of success on benchmarks [3].

This deliverable report focuses on synthetic data generation. More details about causality learning is to be provided in the next deliverable D2.1.

3. Model design and implementation

Problem statement: Given data observations, we learn a DAG (CGM) to match the underlying joint data distribution and yield a synthetic data distribution. We use a weighted adjacency matrix to represent the DAG, where a non-zero entry indicates the existence of a weighted directed edge between two corresponding nodes.

Assumptions: We make the following basic assumptions for causality learning:

- Markov: Given its parents in the DAG, a node is independent of all its non-descendants
- Strict Causal Edges: In a directed graph, every parent is a direct cause of all its children.
- Faithfulness: the converse of the Markov assumption and is known as the faithfulness assumption. This assumption allows us to learn causal graph from data distributions.
- Causal sufficiency: no unobserved confounders of any of the variables in the graph[1].
- Model assumptions: our experiments assume a variety of models (e.g. additive noise models, generalised linear models) based on their causal identifiability

[1] However, we will perform sensitivity analysis to gauge the impact of potential confounders

3.1 Basic model

To learn models from the observations, we follow the NOTEAR framework that was recently proposed by [7] to learn nonparametric DAGs. This involves a nonlinear and nonparametric structural equation model (SEM).

Without loss of generality, we use neural networks to approximate and learn the functional relations between the variables. More specifically, each variable is modelled with a fully connected neural network. Given the data observations, we train all the neural networks through optimisation. The loss function involves generative adversarial loss. In the context of generative adversarial training, these neural networks are called generators (because they can predict values for the variables).

Similar to [6,8], in our model the adjacency matrix that represents the DAG is defined implicitly through the weights of these neural networks. This allows us to add the acyclic constraint to remove loops in the learned causal structure.

The generative adversarial training also involves a discriminator, which measures the distance of the distributions between the (synthetic) data and the (real) training data. The training aims to minimise the difference between the true data and the synthetic data by discovering the right causal structure (DAG). Under the abovementioned assumptions, we can only achieve the global minimum if a true causal structure is discovered.

3.2 Mixed continuous and discrete variables

We extend the basic model to cover both discrete and continuous data variables. Until now, most of the existing approaches assume that the variables of the dataset are drawn from continuous distributions. However, there are many cases where causal structure can be inferred between variables of both data types, particularly within practical settings such as healthcare.

We consider the following discrete (categorical) data types: **Nominal data**, where variables have two or more categories, but do not possess intrinsic order; **Ordinal data,** where variables represent discrete and ordered units (e.g., low, medium, high).

We treat ordinal variables as continuous as there is a natural order between the data values. For nominal variables, we use one-hot encoding. An input nominal variable is encoded with one-hot encoding. To predict a nominal variable, the neural network learns the probabilities of different categories with the softmax function, and the cross-entropy loss is used to calculate its difference from the training (real) data.

3.3 Model training with missing data

Missing data may affect the model identifiability due to the uncertainty from the missing information. This is because the missing measurements in the observations can be completed by different imputation approaches, which may lead to a variety of data distributions underlined by different causal graphs. Such uncertainty is more likely to occur in the situations where there is more missing measurements within the data.

We can naturally build data imputation into the model training process, which is similar to the existing work in using GAN for data imputation [10] but with a different goal for causality learning. We repeatedly apply a variety of alternative imputation methods to the data and the model learns causal structures from the imputed data. The model reports all the candidate causal graphs under different imputation schemes.

3.4 Time series data

Temporal information has not been taken into account in most of the related works, as their underlying models only infer causal relations between contemporaneous variables (i.e. variables that exist at or occur in the same period of time). In our model, the basic architecture of the generator can be easily extended to handle time-series data by imposing the *temporal causal constraint*, which uses the temporal information to enforce a nature causal order. Besides, following Chen and Hsiao [11], two additional reasonable assumptions are 1) The *time-invariant causal structure constraint*, which states that the causal structure between variables at time points t and s is the same as the causal structure between variables at time points $t + c$ and $s + c$, and 2) The *time-finite causal influence constrain*t, which states that a causal influence is limited within certain time steps τ (i.e. a variable do not have causal effect on any other variables beyond τ time steps).

Under the assumptions of the temporal causal constraint, the time-invariant causal structure constraint and the time-finite causal influence constraint, we update the generators to their temporal version by adding the variable from the previous τ steps to the input of each generator.

3.5 Privacy preserving

With the generative model we explore a synthetic data based approach for clinical research with a better protection on data privacy. However, generative models (particularly GANs) tend to learn a data generating function whose distribution concentrates around the training data. To a certain extent, this leads to the memorisation of the training data.

To ensure that the synthetic patients generated from the model do not contain information about the real data, we invoke the principle of differential privacy with DPSGD [12] into our training process. Intuitively, differential privacy seeks to limit the amount of information that the parameters can learn from a given dataset, thereby preventing the retrieval of any information from the data by a motivated attacker. Compared with other privacy methods that attempt to privatise by 'noisifying' the model implicitly, DPSGD instead 'noisifies' the gradients/parameter updates.

Differential privacy places a theoretical upper bound on the difference in outputs observed from a given mechanism, to make sure that changes in output from a model is bounded so that the addition/removal of a data point does not affect performance (thus untraceable), meaning that the outputs are constrained to lie within a fixed 'distance' from one another. Within the context of this project, this means that our DAG model and generated data correspond to the mechanism and outputs of interest, respectively, such that we seek to constrain the parameters of the DAG to be differentially private. The tightness of the privacy bound is controlled by the privacy budget hyper-parameter ϵ, which creates a trade-off between the level of privacy necessary for protection and the quality of the samples generated by the DAG.

4. Experimental Results

4.1 Experiments overview

As the generative model learns causality structure and generates synthetic data simultaneously, we evaluate the model outputs in both of these two aspects.

To evaluate the outcome from causality learning, we need to compare the model inferred causality with the causal ground truth. To this end, we use both a simulation based approach and a real data based approach. The simulation based approach uses a set of mathematical equations to generate the synthetic causal graphs and training data, which allow us to evaluate if the trained model can discover the underlying ground truth causal graphs. The real data based approach involves a real-world diabetes datasets (namely SCI diabetes), together with the data that contain causal relations from the established clinical trials, including LEAD5[13] and LEADERs[14].

Meanwhile, we assess the characteristics of the generated synthetic data by comparing features with the original training data to ensure the consistency of variance and inter-variable relationships.

4.2 Datasets and benchmarks

Here is a brief description of the simulation benchmarks and the real data.

- Simulation benchmark data: The synthesis of the simulation benchmark data was performed in two steps: 1) ground truth graph generation; 2) sample generation from the ground truth graph. In Step (1), we generated an Erdos-Renyi directed acyclic graph with an expected node degree of 3. In Step (2), data samples were generated based on the following non-linear equation.

$$X = 2\sin\left(A^T(X + 0.5) + A^T(X + 0.5) + Z\right) \quad (1)$$

  Where Z represents random noise, and A is the weighted adjacency matrix to represent the ground truth graph(DAG). This equation was widely used in other related works, including DAG-GNN, DAG-NoCurl and DAG-WGAN [3-8]. This allows us to make a comparison against all other models.
- Real data: In this project we will conduct the virtual clinical trials emulation (LEAD-5 and LEADER) with real-world observational health data (electronic health records). This mainly involves the SCI Diabetes[15], which is an inclusive national dataset of individuals with diabetes containing a broad range of longitudinal

demographic, phenotypic, biochemical and screening data. There are approximately 300K individuals with diabetes (and about 480,000 legacy individuals). Over 3K individuals with MODY (Maturity-onset diabetes of the young) are recorded with certainty (genetic information) along with records of individuals with negative genetic test results. Table 1 provides an overview of the observational data that are involved in this study.

Table 1: Overview of the Real-world Datasets

GPLES	Local Enhanced Service reported data from GP surgeries covering a range of long- term health conditions managed in primary care.
Pharmacy	Drug data including their prescription and dispense dataset
SCI_Diabetes	A fully integrated shared electronic patient record to support treatment of NHSScotland patients with Diabetes.
SCI_Store	SCI Store is a data repository which retains patient information at a health board level, accepts various clinical laboratory reports, and includes patient episode tracking.
SMR00	An SMR00 is generated for outpatients receiving care in the specialties listed when they attend different types of clinics.
SMR01	An SMR01 is generated for patients receiving care in General / Acute specialties when they are admitted as inpatients under various circumstances.

Based on the criteria for both LEADER and LEAD-5 (to include patients appropriate for either), our pre-processing selects the data by applying the following *inclusion criteria* 1) T2DM; 2) bA1c at least 53mmol/mol (7%) at any time; 3) Age under 80 & over 18 at first data point; and the *exclusion criteria* 1) T1DM; 2) BMI >45; 3) Continuous renal replacement therapy; 4) End stage liver disease; 5) Previous or awaiting solid organ transplant; 6) Malignant neoplasm

After the pre-processing and data selection, the datasets include 78 demographics variables (e.g. ability to self care, BMI, age, alcohol status, blood pressure); 362 laboratory variables (e.g. biochemistry measurement such as glucose); 123 drugs (e.g. aspirin, liraglutide), and other specialist medical records.

4.3 Improved results on synthetic benchmarks

The experiments were conducted with three datasets, each of which has 5000 samples. The graph sizes used in the experiments were 10. They were all generated with Equation (1).

For the evaluation, we measure:

- SHD (structural hamming distance) between the output (i.e. discovered causal graph) of our model and the ground truth causal graph
- True Positive Rate (TPR) – The percentage of the edges that are correctly discovered.
- False Positive Rate (FPR) – The percentage of the edge that are discovered with wrong direction.
- False Discovery Rate (FDR) - The percentage of the discovered edges that are not in the ground truth.

The results from the three datasets are provided in Table 2. In comparison, typically the related work achieved SHD between 2 to 60.

Table 2: Results of causality learning on simulation benchmarks

Dataset	Threshold	TPR	FPR	FDR	SHD
1	0.1	1.0	0.615	0.457	16
	0.2	1.0	0.077	0.095	2
	0.3	0.94	0.038	0.053	2
2	0.1	1.0	0.633	0.559	19
	0.2	1.0	0.167	0.25	5
	0.3	1.0	0.033	0.06	1

3	0.1	1.0	0.516	0.533	16
	0.2	1.0	0.259	0.363	8
	0.3	1.0	0.065	0.123	2

To reduce the number of false discoveries, we apply thresholding to the estimated adjacency matrix by following [7]. We apply a range of thresholds between 0.1 to 0.3. All the thresholds can achieve a good TPR. However, 0.3 produces good results in terms of FPR, FDR and SHD.

4.4 Initial experiments and results on real-world datasets

The experiments on the real data have so far targeted selected drugs. The purpose of the experiments is to set up a workflow for identifying causal relations from the data.

So far, we have received clinical data containing 33674 patients (we expect this number to triple in the next stage of the project).

According to the advice from the clinical experts in T2DM, we select the following variables from the data in the experiments:

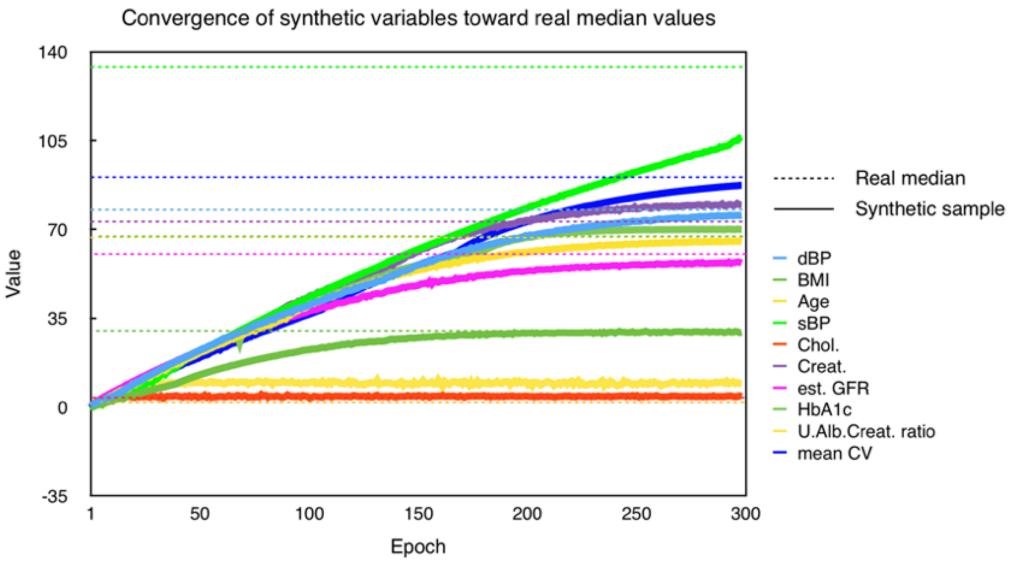
- Biochemistry: Cholesterol, Creatinine, GFR, HbA1c, U_Alb_Creat_Ratio and Mean_Cell_Volume
- Drugs: Alogliptin, Alogliptin_metformin, Biphasic_insulin_aspart, Biphasic_insulin_lispro, Biphasic_isophane_insulin,canaglifozin, Dapagliflozin, Dulaglutide, Empagliflozin, Exenatide, Gliclazide, Glimepiride, Glucagon, Ins_degludec_liraglutide, Insulin_aspartr, Insulin_deglude, Insulin_detemir, Insulin_glargine, Isophane_insulin, Linagliptin,
- Demographics: DiastolicBP, BMI, Diagnosis, Age, SystolicBP

Appendix A shows the distributions of these variables within the datasets.

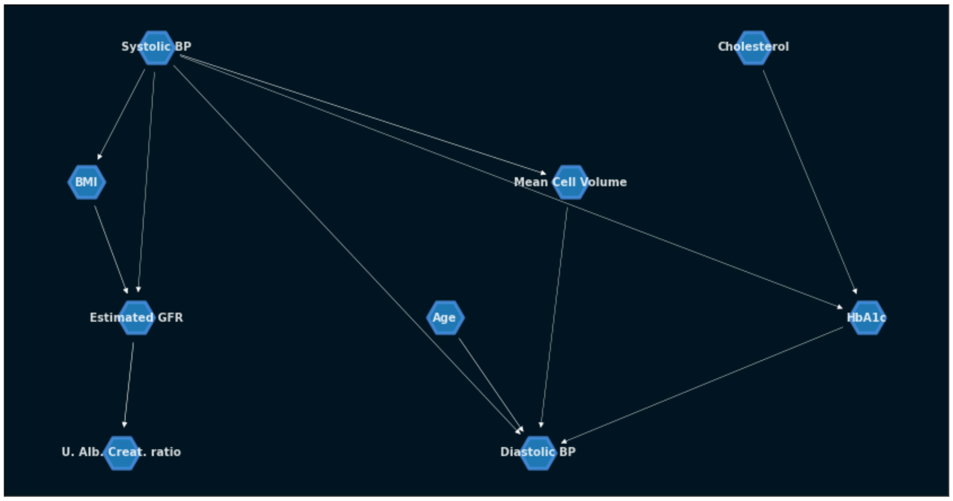
At this stage, we only tested one drug, namely Gliclazide. Of the original 33674 patients, 6583 patients received Gliclazide in their medical history. After the removal of missing data, we were left with 1581 patients 500 of which were used for training, and the remaining 1081 for testing. We involved 10 biochemistry variables and demographics variables.

The training took 2 days to complete 300 epochs. We compared the synthetic data after the training with the median values of the real data – see Figure 1.a, which shows the trend of convergence of the synthetic variables towards the real data. Figure 1.b is the visualization of the learned causal graph that shows causal relations between the variables. Figure 1.c gives a snapshot of the synthetic patients generated by the model. We also performed correlation analysis to compare the correlations between the real and synthetic patients in Figure 1.d.

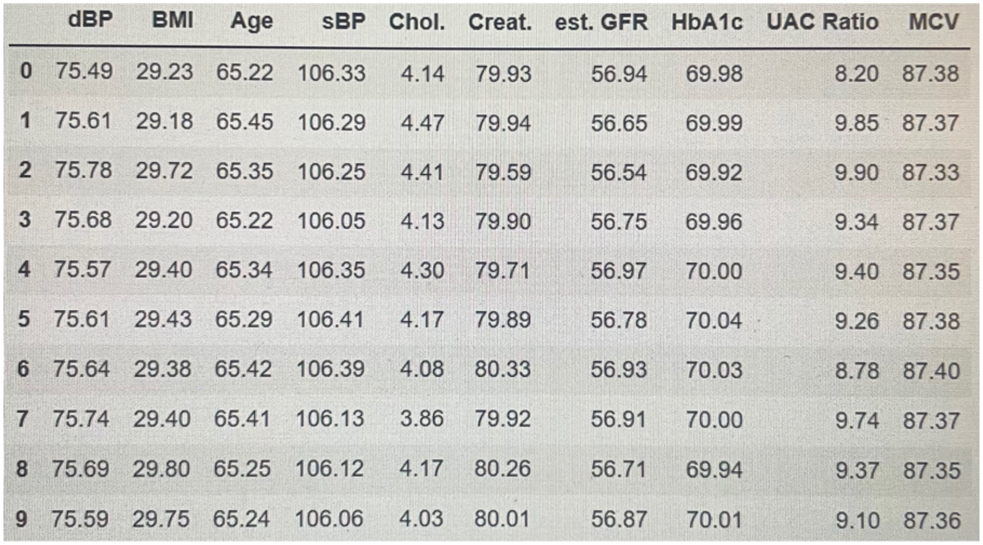
The evaluation is only at its initial stage. Figure 1.a shows an encouraging trend that the synthetic data from the model converges towards the real data. The causal graph in Figure 1.b and the snapshot of the synthetic cohort in Figure 1.c also look reasonable. However, the results in Figure 1.d show that so far the synthetic data haven't learned useful correlations between the data variables from the real data. This is unsurprising due to poor sample diversity in the synthetic patients.



Convergence of synthetic and real data

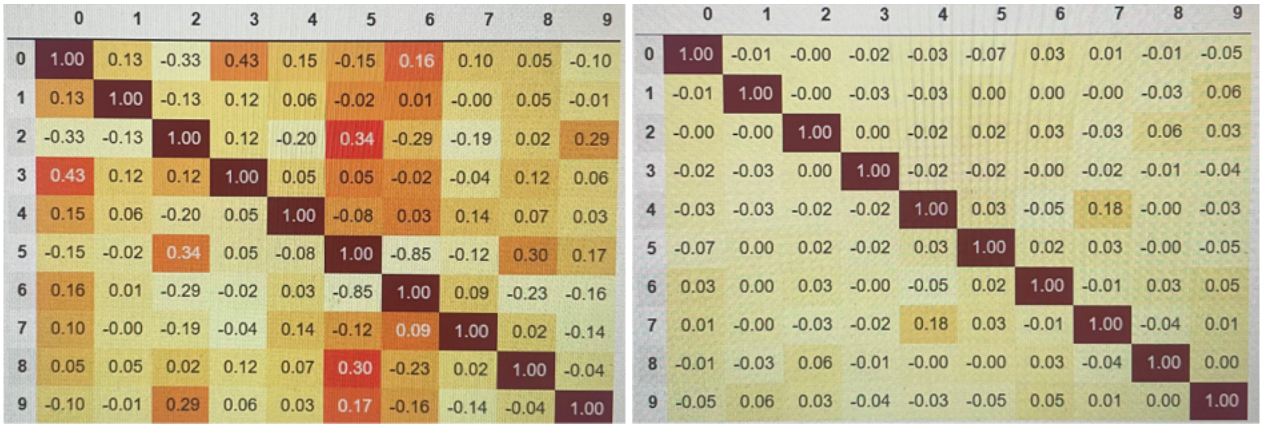


(a) Visualization of the learned causal graph



	dBP	BMI	Age	sBP	Chol.	Creat.	est. GFR	HbA1c	UAC Ratio	MCV
0	75.49	29.23	65.22	106.33	4.14	79.93	56.94	69.98	8.20	87.38
1	75.61	29.18	65.45	106.29	4.47	79.94	56.65	69.99	9.85	87.37
2	75.78	29.72	65.35	106.25	4.41	79.59	56.54	69.92	9.90	87.33
3	75.68	29.20	65.22	106.05	4.13	79.90	56.75	69.96	9.34	87.37
4	75.57	29.40	65.34	106.35	4.30	79.71	56.97	70.00	9.40	87.35
5	75.61	29.43	65.29	106.41	4.17	79.89	56.78	70.04	9.26	87.38
6	75.64	29.38	65.42	106.39	4.08	80.33	56.93	70.03	8.78	87.40
7	75.74	29.40	65.41	106.13	3.86	79.92	56.91	70.00	9.74	87.37
8	75.69	29.80	65.25	106.12	4.17	80.26	56.71	69.94	9.37	87.35
9	75.59	29.75	65.24	106.06	4.03	80.01	56.87	70.01	9.10	87.36

(c) A snapshot of the synthetic population from the model



	0	1	2	3	4	5	6	7	8	9
0	1.00	0.13	-0.33	0.43	0.15	-0.15	0.16	0.10	0.05	-0.10
1	0.13	1.00	-0.13	0.12	0.06	-0.02	0.01	-0.00	0.05	-0.01
2	-0.33	-0.13	1.00	0.12	-0.20	0.34	-0.29	-0.19	0.02	0.29
3	0.43	0.12	0.12	1.00	0.05	0.05	-0.02	-0.04	0.12	0.06
4	0.15	0.06	-0.20	0.05	1.00	-0.08	0.03	0.14	0.07	0.03
5	-0.15	-0.02	0.34	0.05	-0.08	1.00	-0.85	-0.12	0.30	0.17
6	0.16	0.01	-0.29	-0.02	0.03	-0.85	1.00	0.09	-0.23	-0.16
7	0.10	-0.00	-0.19	-0.04	0.14	-0.12	0.09	1.00	0.02	-0.14
8	0.05	0.05	0.02	0.12	0.07	0.30	-0.23	0.02	1.00	-0.04
9	-0.10	-0.01	0.29	0.06	0.03	0.17	-0.16	-0.14	-0.04	1.00

	0	1	2	3	4	5	6	7	8	9
0	1.00	-0.01	-0.00	-0.02	-0.03	-0.07	0.03	0.01	-0.01	-0.05
1	-0.01	1.00	-0.00	-0.03	-0.03	0.00	0.00	-0.00	-0.03	0.06
2	-0.00	-0.00	1.00	0.00	-0.02	0.02	0.03	-0.03	0.06	0.03
3	-0.02	-0.03	0.00	1.00	-0.02	-0.02	-0.00	-0.02	-0.01	-0.04
4	-0.03	-0.03	-0.02	-0.02	1.00	0.03	-0.05	0.18	-0.00	-0.03
5	-0.07	0.00	0.02	-0.02	0.03	1.00	0.02	0.03	-0.00	-0.05
6	0.03	0.00	0.03	-0.00	-0.05	0.02	1.00	-0.01	0.03	0.05
7	0.01	-0.00	-0.03	-0.02	0.18	0.03	-0.01	1.00	-0.04	0.01
8	-0.01	-0.03	0.06	-0.01	-0.00	-0.00	0.03	-0.04	1.00	0.00
9	-0.05	0.06	0.03	-0.04	-0.03	-0.05	0.05	0.01	0.00	1.00

(d) Correlation matrices for 1000 real (left) and 1000 fake (right) patients

Figure 1: Results of the synthetic data from the initially trained model

This stage of the evaluation was restricted by limited access to GPUs. All the training experiments were carried out with CPUs and this took considerably longer time than the use of GPUs. This is due to the setting in the current computational environment on the Safe Haven platform, which is expected to be updated to accommodate our needs for GPUs.

In the next step, the evaluation will use Memorization-informed Fréchet Inception Distance (MiFID)[16], which is a standard GAN evaluation metric to measure the quality and diversity of the generated data (It penalises the replication of single data records from training data). We will also assess the quality of the synthetic data by measuring their effectiveness to train machine learning models. We will use both the synthetic and original data to train the same machine learning models and compare the performance of the trained machine learning models with established metrics (e.g. confusion matrix).

5. Conclusions and future steps

In summary, the work since the beginning of the project covers:

- Update the original generative adversarial causal learning mode in [3] to a nonparametric structural equation model to support the modelling of a range of causal relations within the dataset without being restricted by a prefixed causal equation. We also extend the model to support mixed continuous and discrete variables. Privacy preserving has also been implemented to add noise into the gradients of the model parameters during training for the privacy protection of the training data. We have also considered how to learn causality from data with missing measurements, and how to utilise time series measurements. We have conducted an evaluation using the simulation benchmarks and achieved good results in causal structure learning measured by SHD.
- We have been given access to the real world observational clinical data in the Safe Haven platform. The data are from linked clinical datasets. After a careful consideration by the clinical experts, we have designed and applied inclusion and exclusion criteria to select the first patient cohort including 33674 patients. We have carried out pre-processing to the data and identified the variables of interest including biochemistry, drugs and demographics variables.
- Our initial experiment on 500 patients on the training shows that the model has started to identify causal relations from the training data, and the initial results are encouraging. However, we are limited by the access to GPUs and hence so far we are not able to carry out training on larger scale samples with sufficient iterations.

Looking ahead, the next step will involve:

- Experiment with larger sample size – we have made arrangements with the data provider and we will have access to about 100,000 patients in total. We have also made arrangement with the Safe Haven team about the GPU access issue. By solving these two issues, we will be carry out experiments with larger samples to gain more valuable insights into the model performance.

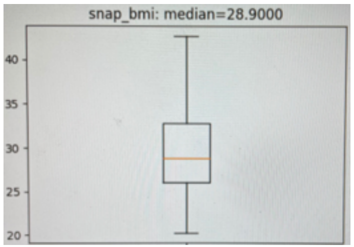
- We will also add increased regularisation terms to improve sample diversity in the training. This will help address mode collapse problem, which happens when the generative model only produces synthetic data with limited diversity. This is one of the most common issues in GANs[17].
- We will integrate the advanced mechanism for training with differential privacy guarantee and training with missing data measurements.
- We will investigate automated hyper-parameter tuning to identify optimal values for a set of hyper parameters of the model, involving methods such as hyper-gradients and Bayesian optimisation.
- We will evaluate synthetic population at a much larger scale with additional metrics including MiFID, machine learning classifiers etc.

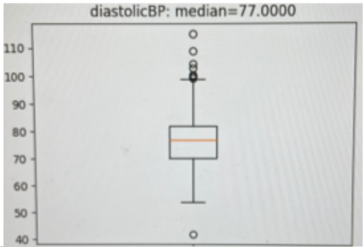
References

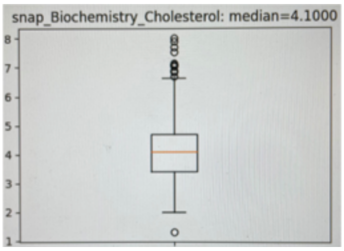
1. Greenland, S., Pearl, J., Robins, J.M., 1999. Causal diagrams for epidemio-logic research. Epidemiology , 37–48.
2. Hern´an, M.A., Robins, J.M., 2006. Instruments for causal inference: an epidemiologist's dream? Epidemiology , 360–372
3. Petkov, H., Hanley, C., Dong, F., 2022. Dag-wgan: Causal structure learning with wasserstein generative adversarial networks. ArXiv abs/2204.00387.
4. Yu, Y., Chen, J., Gao, T., Yu, M., 2019. Dag-gnn: Dag structure learning with graph neural networks. International Conference on Machine Learning
5. Yue Yu, Tian Gao, N.Y., Ji, Q., 2021. Dags with no curl: An efficient dag structure learning approach. Proceedings of the 38th International
Conference on Machine Learning
6. Lachapelle, S., Brouillard, P., Deleu, T., Lacoste-Julien, S., 2020. Gradient-based neural dag learning. ArXiv abs/1906.02226
7. Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P., 2018. Dags with no tears: Continuous optimization for structure learning. Conference on Neural Information Processing Systems .
8. Zheng, X., Dan, C., Aragam, B., Ravikumar, P., Xing, E.P., 2020. Learning sparse nonparametric dags. ArXiv abs/1909.13189
9. Ferenc Huszar, How (not) to train your generative model: scheduled sampling, likelihood, adversary? Iclr 2016? ICLR 2016
10. S. Li, B. Jiang, B. Marlin, MisGAN: Learning from Incomplete Data with Generative Adversarial Networks, https://arxiv.org/abs/1902.09599
11. Chen, P. and Hsiao, C. (2007). Learning causal relations in multivariate time sereis data. Economics: The Open-Access, Open-Accessment E-Journal, 1, 2007- 11.
12. Martín Abadi, Andy Chu,Ian Goodfellow,H Brendan McMahan,Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 308–318.
13. D. Russell-Jones&A. Vaag&O. Schmitz&B. K. Sethi&N. Lalic&S. Antic&M. Zdravkovic&G. M. Ravn&R. Simó&, Liraglutide vs insulin glargine and placebo in combinationwith metformin and sulfonylurea therapy in type 2 diabetesmellitus (LEAD-5 met+SU): a randomised controlled trial, Diabetologia (2009) 52:2046–2055DOI 10.1007/s00125-009-1472-y
14. Marso, SP., Daniels, GH., Brown-Frandsen, K., et al; (2016) Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes, New England Journal of Medicine 375:311-322
15. NHS Greater Glasgow and Clyde & NHS Tayside Health Board area data https://www.nhsggc.org.uk/about-us/professional-support-sites/glasgow-safe-haven
16. Bai, C. et al , On Training Sample Memorization: Lessons from Benchmarking Generative Modeling with a Large-scale Competition, KDD '21, August 14–18, 2021, Virtual Event,
17. Wiatrak et al (2020), Stabilizing Generative Adversarial Networks: A Survey, https://arxiv.org/abs/1910.00927

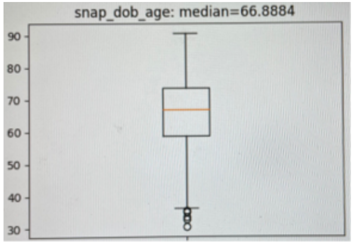
Appendix A: Snapshot data variable distributions

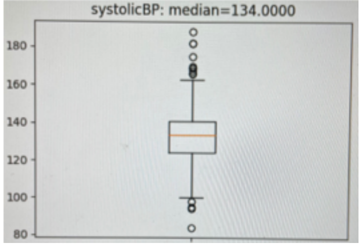
- Snapshot biochemistry: Cholesterol, Creatinine, GFR, HbA1c, U_Alb_Creat_Ratio and Mean_Cell_Volume
- Drugs: Alogliptin, Alogliptin_metformin, Biphasic_insulin_aspart, Biphasic_insulin_lispro, Biphasic_isophane_insulin,canaglifozin, Dapagliflozin, Dulaglutide, Empagliflozin, Exenatide, Gliclazide, Glimepiride, Glucagon, Ins_degludec_liraglutide, Insulin_aspartr, Insulin_deglude, Insulin_detemir, Insulin_glargine, Isophane_insulin, Linagliptin,
- Demographics: DiastolicBP, BMI, Diagnosis, Age, SystolicBP

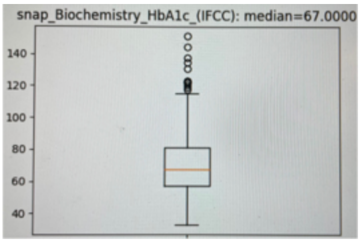

snap_bmi: median=28.9000

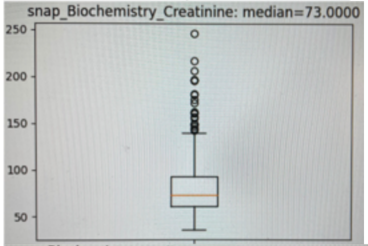

diastolicBP: median=77.0000

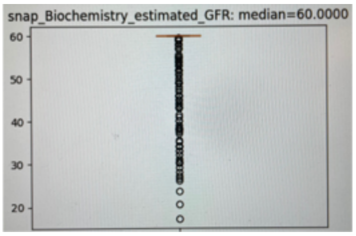

snap_Biochemistry_Cholesterol: median=4.1000

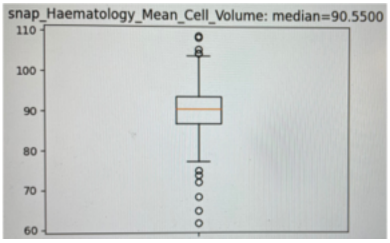

snap_dob_age: median=66.8884

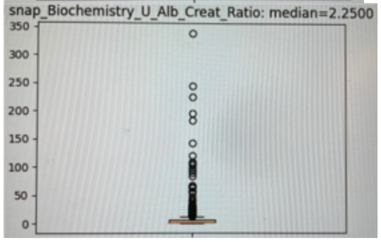

systolicBP: median=134.0000


snap_Biochemistry_HbA1c_(IFCC): median=67.0000


snap_Biochemistry_Creatinine: median=73.0000


snap_Biochemistry_estimated_GFR: median=60.0000


snap_Haematology_Mean_Cell_Volume: median=90.5500


snap_Biochemistry_U_Alb_Creat_Ratio: median=2.2500