MRC-GAN Deliverable 1.2: Technical validation results for causality learning

Abstract This report describes the second stage of the work in the MRC-GAN project, which focuses on the evaluation of the causality learning model using the real-world dataset. We performed data and variable selection by carefully defining the inclusion and exclusion criteria based on clinical knowledge, and identified a set of target variables of interest. We have performed a series of experiments on the SCI Diabetes dataset. The training curves illustrate the evolution of the generated synthetic data and their convergence during training. We have also evaluated the quality of the independent statistical structure of each synthetic feature in comparison to their real counterparts. To give more detailed insight into the properties of the synthetic data, we compute the marginal distributions across each feature, as well as the correlations between the features. We have also performed regression analysis on both real and synthetic data to evaluate the efficacy of the generated features for predicting relevant clinical outcomes with random forest and compared the saliency of the features used in these predictions to discern whether the random forest model trained on the synthetic data makes use of sensible clinical features. A key change to our methodology was to assign the drugs to their respective classes. Expressing each drug class as separate nodes in the graph, we observed that the network-like structure presented in the previous report, where all features have some connection to other features, has been replaced by strong one-way connections from a salient set of features and drug classes to the remaining features.

The report is organised as follows: Section 1 gives the general background of the project. Section 2-3 presents the model; and Section 4 shows our evaluation experimental set up, training process and the outcomes measured in terms of the synthetic data and causal relationships.

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 in the real dataset). However, the privacy preserving mechanism 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. More details can be found in deliverable D1.1 This deliverable is focused on reporting the performance of the model on the SCI Diabetes dataset[4]

3. Model design and implementation

To learn models from the observations, we follow the NOTEAR framework that was recently proposed by [1] to learn nonparametric DAGs. This involves a nonlinear and nonparametric structural equation model (SEM). We present the details of the model and its implementation in Deliverable D1.1.

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[2] and LEADERs[3].

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

Deliverable D1.1 reports the simulation datasets and their results. This deliverable reports the results on the real data.

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 NHS Scotland 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) HbA1c 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 Experiments with SCI Diabetes

In this section, we provide the results to a series of experiments on the SCI Diabetes dataset. For brevity, we include salient results that we believe best represent the current progress of the work: we defer additional results and experimental findings to the Appendix.

4.3.1. Experimental setup

Dataset. The dataset consists of an array of clinical variables, drug prescriptions, and demographics variables for 56,476 unique patients with T2DM. For our analysis, we considered 10 features and extracted the median values of features with multiple measurements collected over the span of a single year. We also considered each drug as a separate discrete variable from which to infer causal effects with the features. To reduce the number of possible drug combinations, and to account for the confounding effects imposed by other prescribed drugs, we decided to categorise all drugs into their corresponding class. This allowed us to model the 'global' causal effects of each drug, under the assumption that all drugs within each class cause similar effects to a given patients' features: we provide an overview of all features and drug classes in Table 2. To mimic the inclusion criteria for the LEAD-5 trial, we screened the original dataset of all patients who were prescribed a metformin-based drug, which are commonly used in the initial phases of management. After screening all metformin patients, in addition to removing missing data, the final dataset comprised 8012 patients, of which we assigned 5,000 and 3,012 to training and testing sets, respectively. For model training, we performed a 90/10 split of the training set into training and validation subsets.

Table 2: Overview of the data features using for model training. Drugs were organised into their corresponding class, where the proportion of patients on each specific drug class is provided. Patients were included in training only if they received a metformin-based drug.

Feature	Drug class (proportion)
Diastolic blood pressure (DBP)	Nil (41.7%)
Body mass index (BMI)	Bd insulin (4.6%)
Age	DPP-4 (14.3%)
Systolic blood pressure (SBP)	Basal insulin (3.8%)
Cholesterol	Prandial insulin (1.1%)
Creatinine	SGLT-2 (29.5%)
Estimated GFR	GLP-1 (14.3%)
HbA1c	SU (26.4%)
UAC ratio	TZD (1.9%)
Mean cell volume	Metformin (100%)

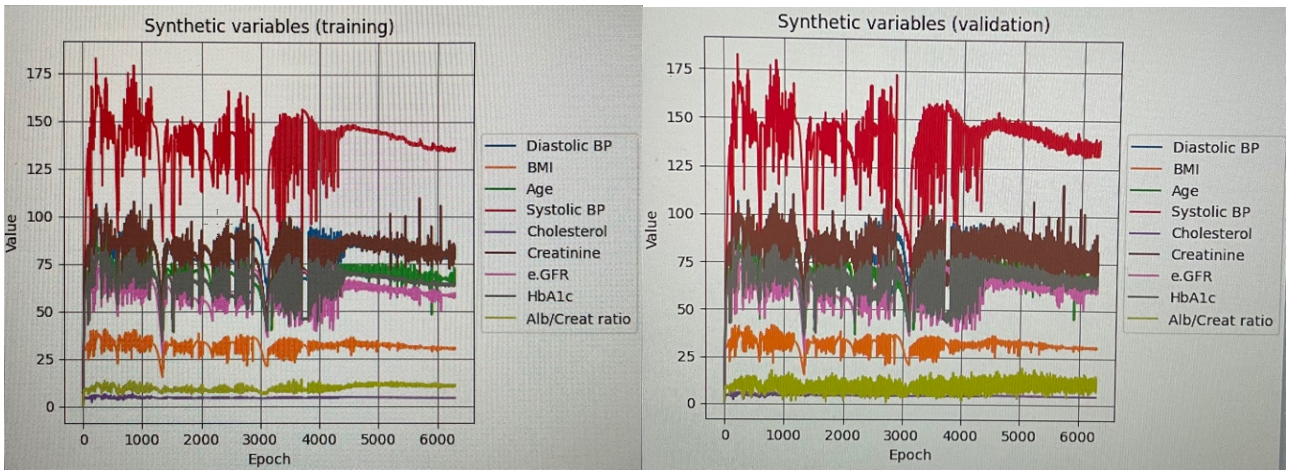
Model and training parameters. To simultaneously learn causal structure and a function for realistic data generation, we implemented a Wasserstein GAN with a gradient penalty loss (WGAN-GP). This architecture makes use of a discriminator network whose role is to act as a 'critic' to inform a separate generator network of how realistic its generated patients are. To help mitigate the risk of mode collapse, we further implemented a PAC-based discriminator [7] with a PAC value of 10. Both discriminator and generator networks were optimised with Adam using initial learning rates of 3e-3, which were gradually decayed to 2e-4 according to the Langrangian parameters: we refer the reader to Appendix A.2 for a comprehensive list of our hyper-parameter settings. We monitored the training behaviour of the networks using the GP and Wasserstein loss terms, in addition to the maximum mean discrepancy (MMD) and mean squared error (MSE) between the real and synthetic data to measure closeness-of-fit. For interpretability, we averaged over all features and drug probabilities produced by the

generator throughout training to monitor sample diversity and to identify possible mode collapsing amongst the drug combinations.

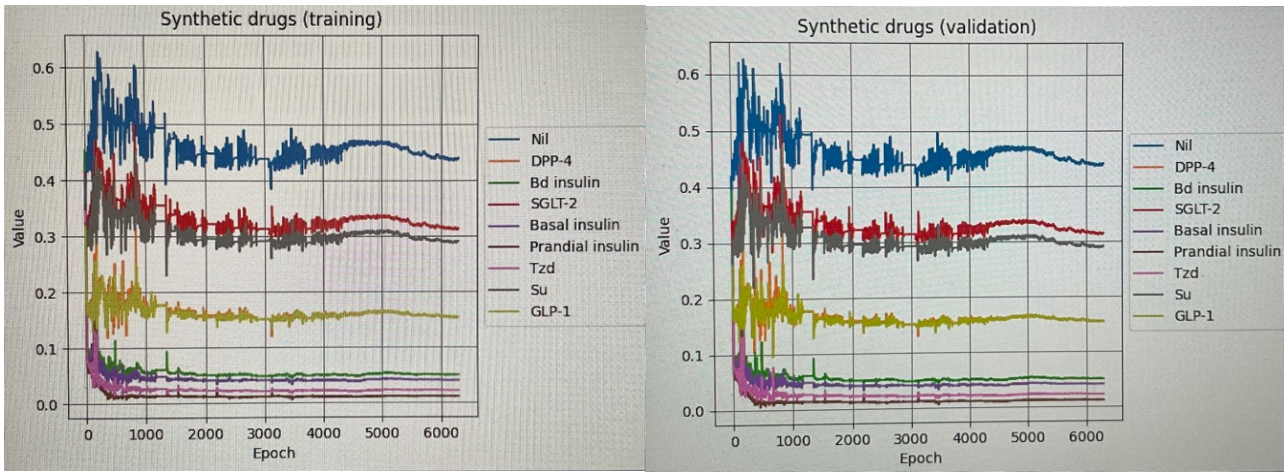
Evaluation. For evaluation, we generated 3,012 synthetic patients to match the test set size. Patient features were generated by fixing the model weights and predicting each child feature from its parent node based on the learned topological ordering. The weights selected for synthetic data generation were those that showed sensible causal relationships between the drug classes and clinical features (see Figure 4). In this case we used the weights at the 12th Lagrangian iteration, which revealed edges of relevance to the LEAD-5 clinical trial (e.g., GLP-1 and SU to blood pressure). To illustrate the effects that further training has on these connections, we provide results for the weights at the final Langrangian iteration in Appendix B.

4.3.2. Training curves

In Figure 1, we show the evolution of the synthetic features and drug combinations to determine whether the model converges toward clinically meaningful values: more detailed training results (including the loss curves) are available in Appendix A.1. From Fig. 1a, we observe that the model appears to have improved feature diversity (compared with Deliverable 1.1's results) but tends to converge toward a fixed probability for each drug class (Fig. 1b). Closer inspection reveals that the converged values correspond to the proportion of each drug class in the dataset (see Table 2), which implies that the model only generates patients from the dominant drug classes instead of predicting realistic drug combinations. In this case, we observe that the model assigns a probability of less than 0.5 for all drug classes, meaning that, after binary thresholding, the synthetic cohort will only represent patients on the metformin drug class.



(a)



(b)

Figure 1: *Evolution of the generated synthetic data during training.* (a) Clinical features of the synthetic patients. (b) Probabilities for the presence of each drug class. Each data point in all curves represents the average over all mini-batch estimates across a given epoch.

4.3.3. Synthetic data distributions

In the previous section, we showed that the model appears to generate patients whose features converge within reasonable margins as training progresses (Fig. 1a, right), but that these patients most likely correspond to those that only receive metformin-based treatment. Here we select a set of weights from this training process (iteration 12, approx. epoch 3500) and study the statistical structure of each generated synthetic feature in more detail.

Boxplot analysis. We begin by examining the summary statistics of each feature in both the real and synthetic datasets using boxplot analysis. The results are illustrated in Figure 2.

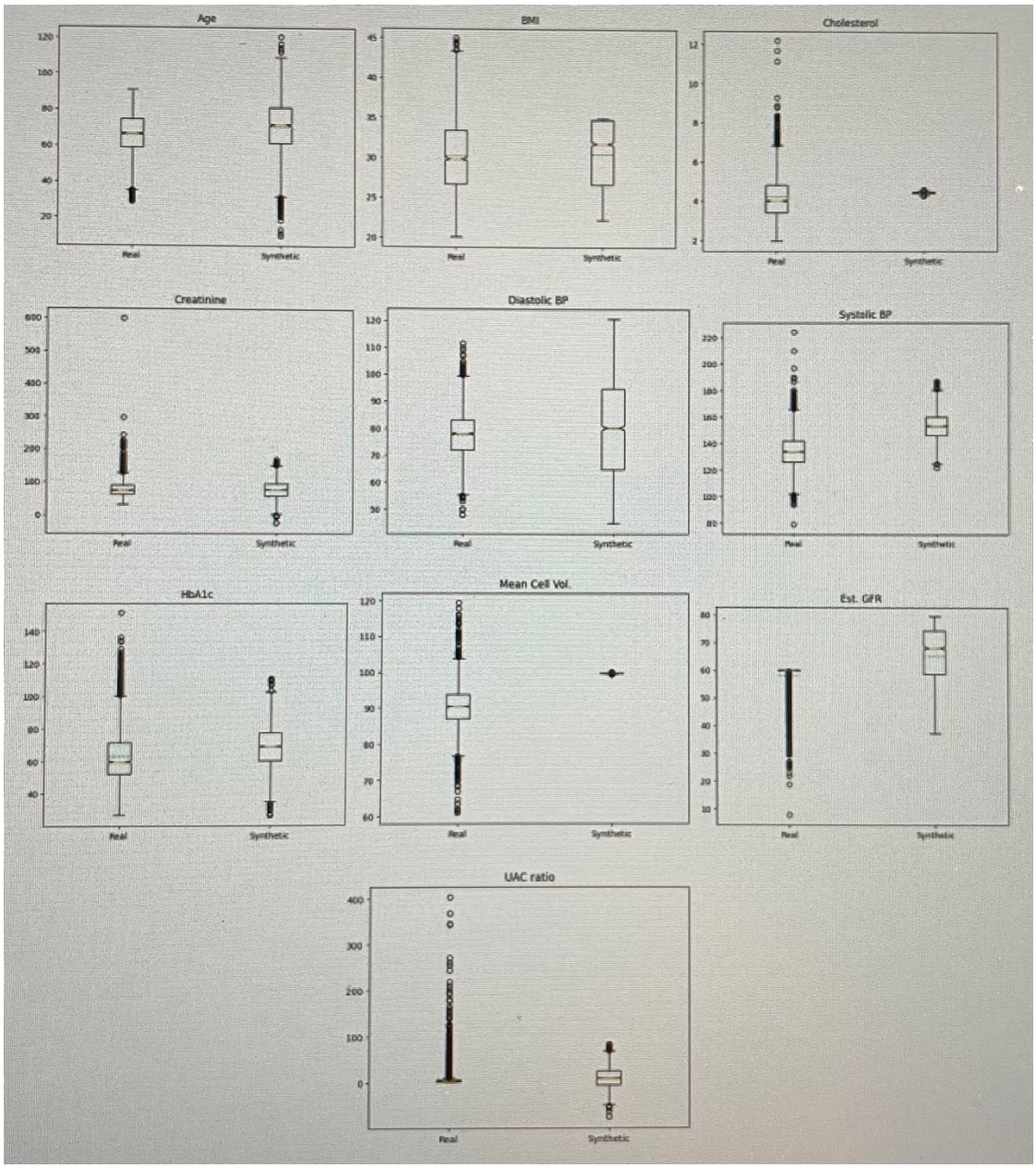
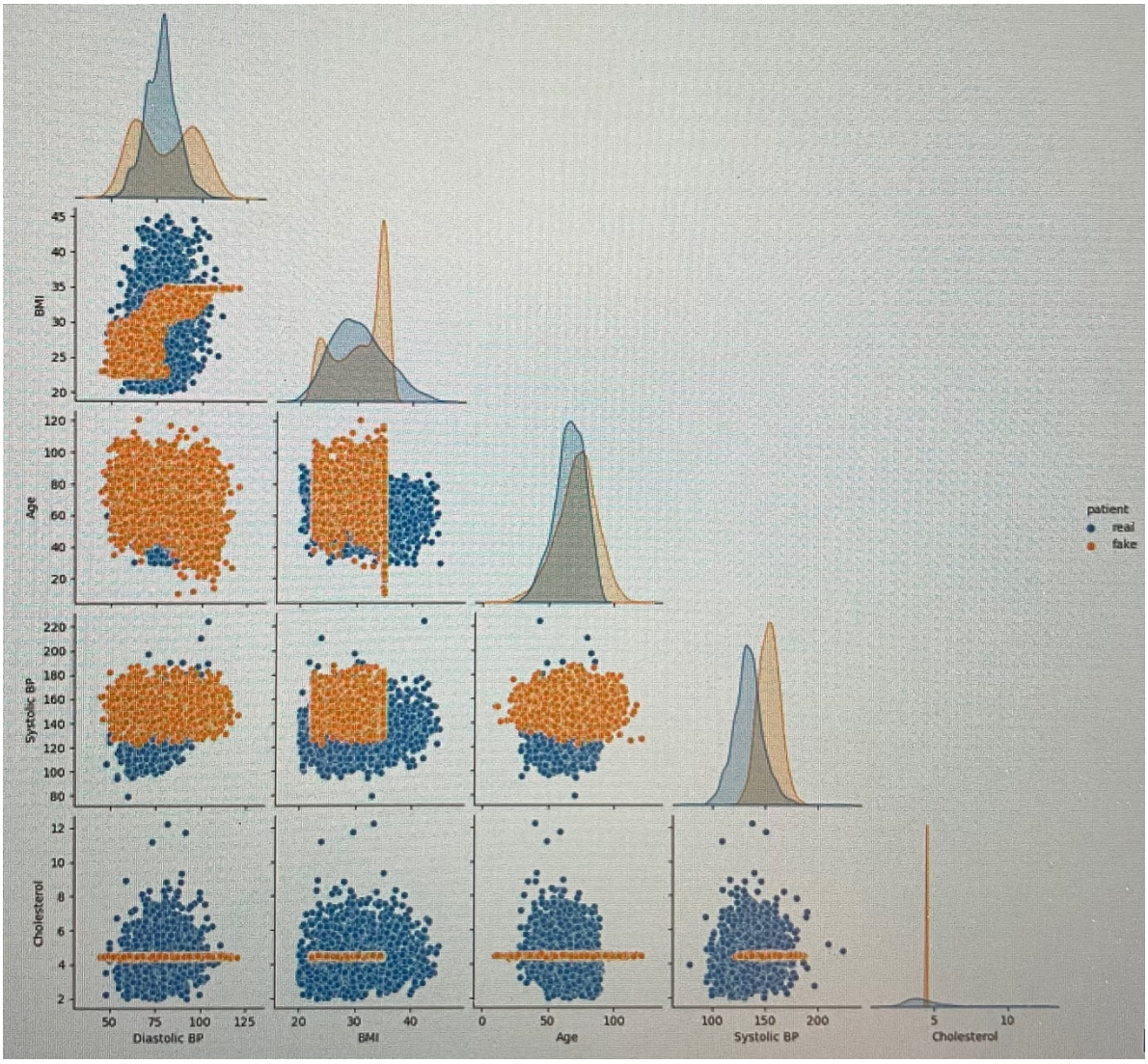


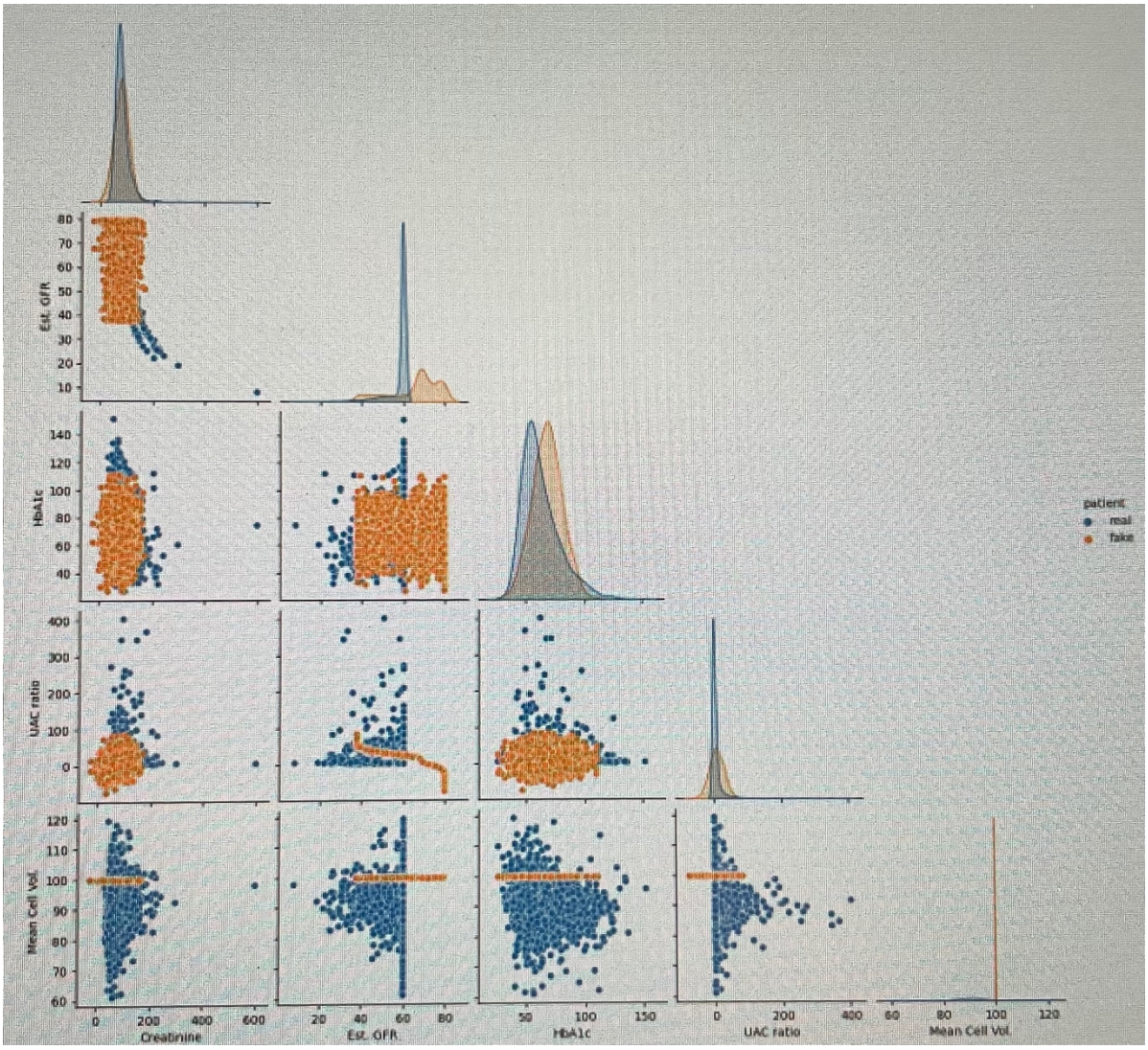
Figure 2: *Boxplot analysis for comparing the summary statistics of the real and synthetic data.* The orange and green lines in each boxplot represent the median and mean values, respectively.

From the boxplots above, we observe that the expected value of each synthetic feature lies reasonably close to their real counterparts. We also observe that most features exhibit a representative interquartile range, but that the model struggles to produce samples that sufficiently capture the real outlier structure. Overall, the results suggest that the model does not succumb completely to mode collapse, but nevertheless evidence the need to further improve sample diversity in some of the features (such as cholesterol and mean cell volume).

Marginal distributions. For deeper insight into the properties of the synthetic data, we compute the marginal distributions across each feature. This approach is particularly well-suited to our problem where parametric assumptions about the data do not apply, enabling us to verify whether the model generates data with overlapping support. To this end, we used kernel density estimation (KDE) to approximate the underlying probability density function (PDF) responsible for generating each feature. In the plots that follow in Figure 3, the resulting KDE estimate of each marginal distribution is shown on the diagonal axis, together with the corresponding joint distributions on the lower triangle.



(a)



(b)

Figure 3: *Marginal distributions over each feature from the real (blue) and synthetic (orange) datasets.* For readability, the first five features from Table 2 are presented in (a), leaving the last five features to (b).

Inspecting the marginal distributions above reveals that the generator successfully learns a PDF for each feature that has overlapping support with the corresponding true PDF, particularly with creatinine, systolic blood pressure, HbA1c, and age. In addition, the model seems to understand that features such as diastolic blood pressure do not have parametric structure (i.e., non-Gaussian), and that a strong correlation exists between creatinine and UAC ratio. However, an important joint distribution that appears to have been poorly modelled is between diastolic and systolic blood pressure. In the real case, the joint distribution shows a positive correlation between the two features, but in the synthetic case, the joint distribution shows no correlation structure.

4.3.4. Causal relationships

In this section, we examine the causal structure learned by the generator. The graph illustrated in Figure 4 depicts the adjacency matrix inferred from the generator weights.

Feature edges. When we introduce the drug classes as additional nodes in the graph, we observe that the network-like structure presented in the previous deliverables report, where most features have some connection to the others, has been replaced by strong one-way connections from a salient set of features. For example, we would expect edges between age and features such as HbA1c and BMI, and edges between systolic and diastolic blood pressure. What we observe instead are direct one-way connections from UAC ratio and cholesterol to the remaining features. Interestingly, based on discussions with our clinical experts, the edges from UAC ratio could imply deeper insights into a key mechanism of T2DM involving endothelial cell functioning. However, for effective virtual trial emulation, it is not desirable for the model to disregard salient connections between other relevant features (e.g., between systolic and diastolic blood pressure). We may therefore need to regularise model training to ensure that the interdependencies between the features are not displaced in favour of the drug connections. We believe this could be addressed with additional terms within the cost function (BCE loss with drugs forced model to learn edges between drugs: similar idea for features to prevent their edges disappearing)

Drug edges. In addition to capturing interdependencies between the features, we would also expect the model to capture structure between the drug classes and particular sets of features. Encouragingly, the graph illustrates that the model indeed learns sensible connections of this nature. Specifically, connections between the GLP-1 class (which includes drugs such as liraglutide and semaglutide) and systolic blood pressure are expected, as demonstrated by the LEAD-5 trial. However, we should not expect either the 'DPP-4' nor the 'nil' class to possess such a connection. A possible explanation for the spurious 'nil' edge is that this class contains statin-based placebo's, which are used to regulate blood pressure: future efforts will ensure that such effects are disentangled from the 'nil' class.

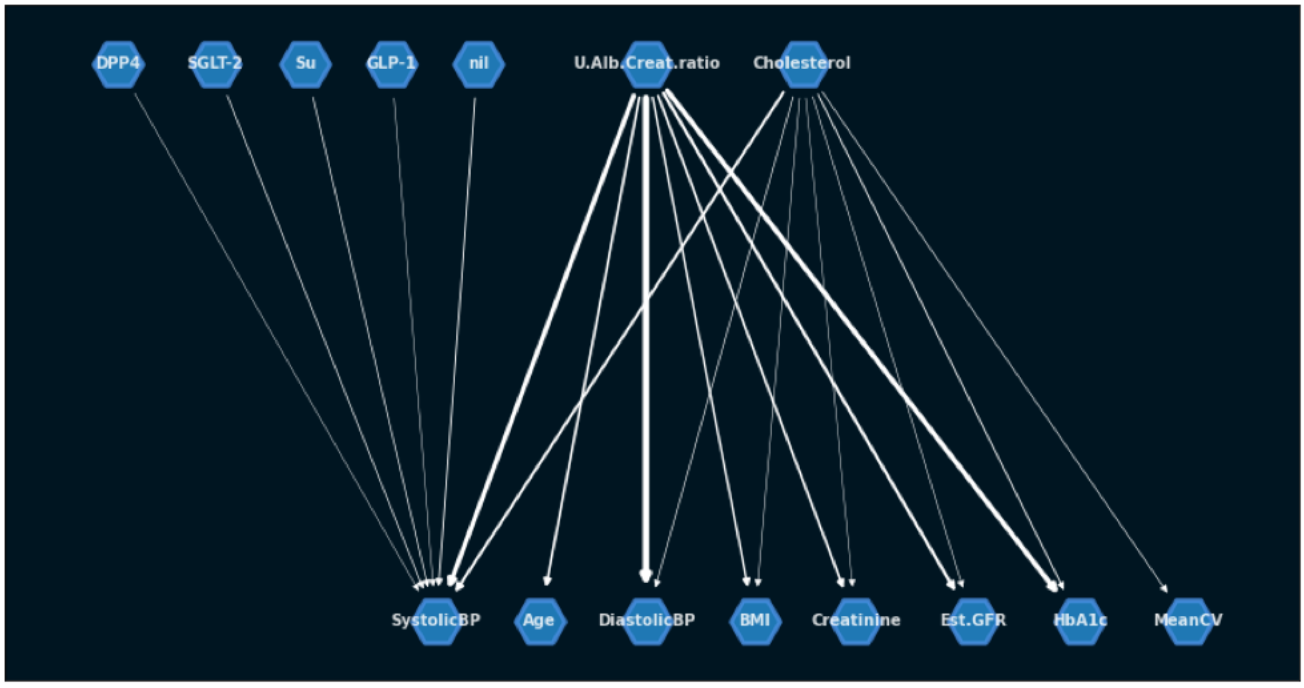


Figure 4 : *Causal graph learned by the generator.*

4.3.5. Correlation matrices

To give further insight into the quality of the synthetic data, we compute the correlation matrices for both real and synthetic datasets.

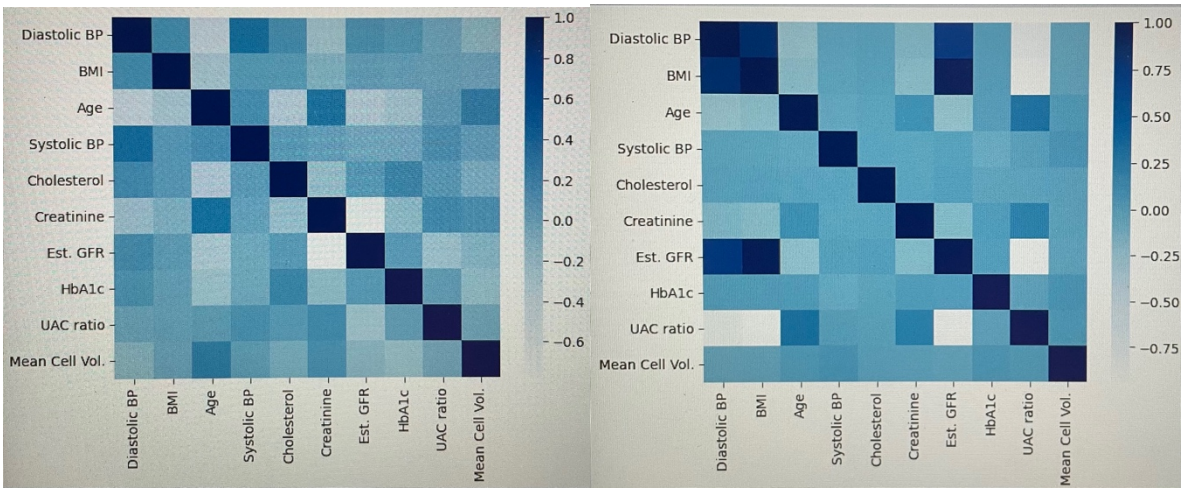


Figure 5: *Correlation structure between features in the real (left) and synthetic (right) datasets.*

From the matrices illustrated in Figure 5, we can observe that, overall, the synthetic data shares much of the correlation structure with the real data, but also neglects some important relationships in addition to introducing new ones. For example, both the real and synthetic data show a positive correlation between BMI and diastolic blood pressure, but in the synthetic data, we observe a strong negative correlation between UAC and diastolic blood pressure, and a positive correlation between est. GFR and BMI, both of which do not exist in the real data. Arguably, the creatinine feature has been modelled most successfully, which we also observed in the marginal distribution plot (Fig. 3b). On the other hand, the synthetic data does not capture the expected relationship between systolic and diastolic blood pressure. Thus, although we see a vast improvement to the correlation structure over the previous deliverables report, there are clear gaps that should be filled. We believe that with more data and wider hyper-parameter sweeps we can encourage the model to, at the very least, partially recover this information.

4.3.6. Machine learning regression analysis

Lastly, we perform regression analysis on both the real and synthetic data to evaluate the efficacy of the generated features for predicting clinical outcomes. We implemented random forest (RF) regression to predict the clinical outcomes, given the remaining patient features. We trained separate RF models on the real and synthetic datasets and assigned a separate set of real data for evaluating the performance of both trained models. For each RF, we used 1000 decision trees with a max depth (number of splits) of 5 per decision tree.

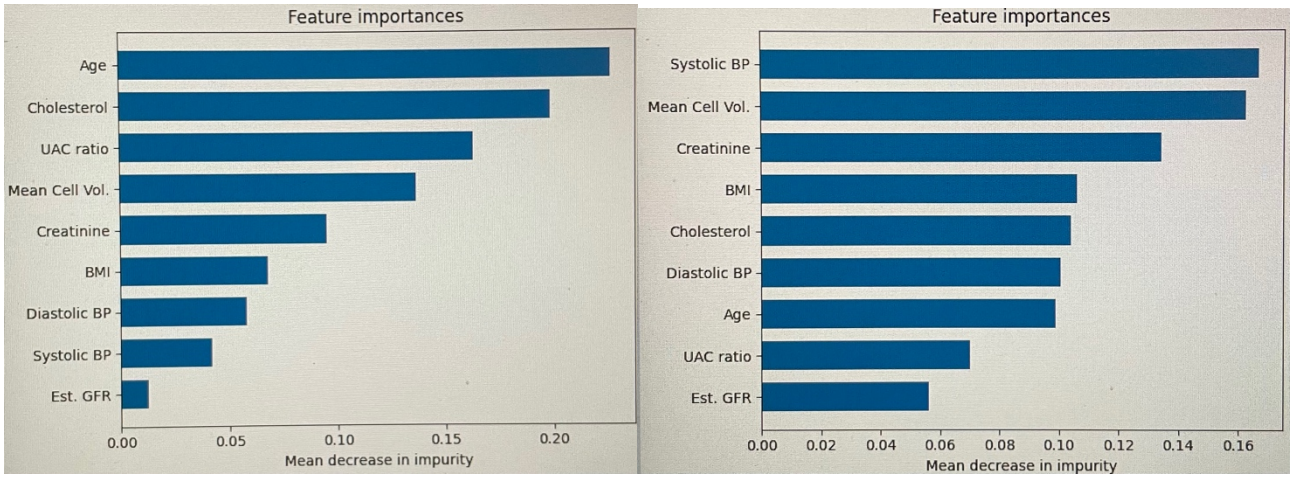
Regression performance. To quantify the efficacy of the synthetic data for predicting clinical outcomes, we compared the performance of both the real and synthetic RF models using the mean squared error (MSE) and R2 metrics. The regression results are presented in Table 3, from which we can observe that HbA1c, BMI, cholesterol, and UAC ratio yield similar MSE in the real and synthetic RF models. However, the synthetic RF model struggles to predict the remaining features, particularly with systolic blood pressure and creatinine.

Table 3: Performance comparison between random forest regression models trained with real and synthetic datasets.

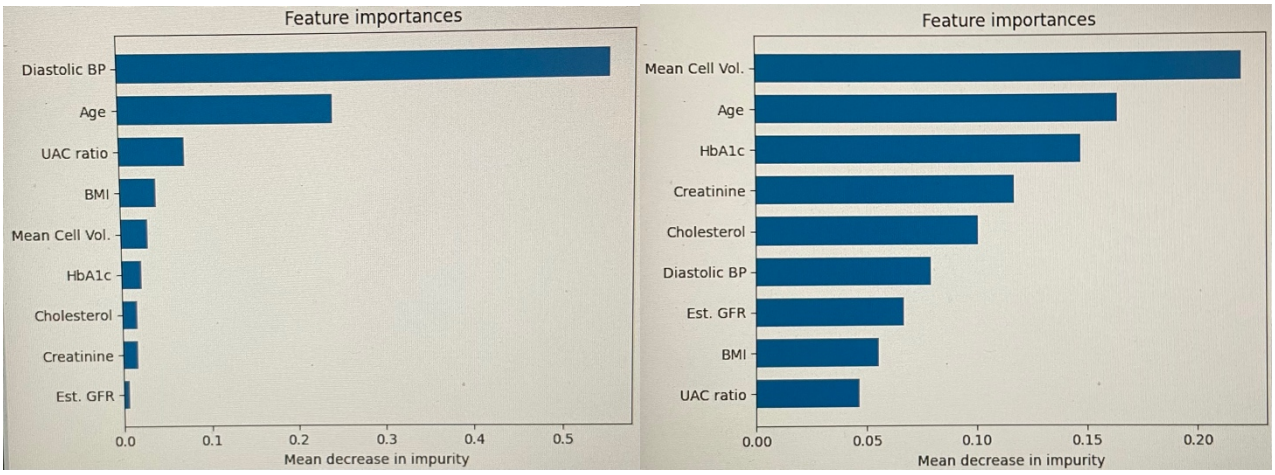
Outcome	Mean squared error (MSE)		R2	
	Real	**Synthetic**	**Real**	**Synthetic**
HbA1c	233.75	282.83	0.05	-0.14
Systolic blood pressure	118.05	513.89	0.28	-2.09
BMI	21.87	29.71	0.07	-0.25
Diastolic blood pressure	44.87	173.15	0.34	-1.54

Creatinine	226.74	981.93	0.61	-0.66
Cholesterol	1.18	1.40	0.12	-0.03
UAC ratio	370.33	535.03	-0.41	-1.04

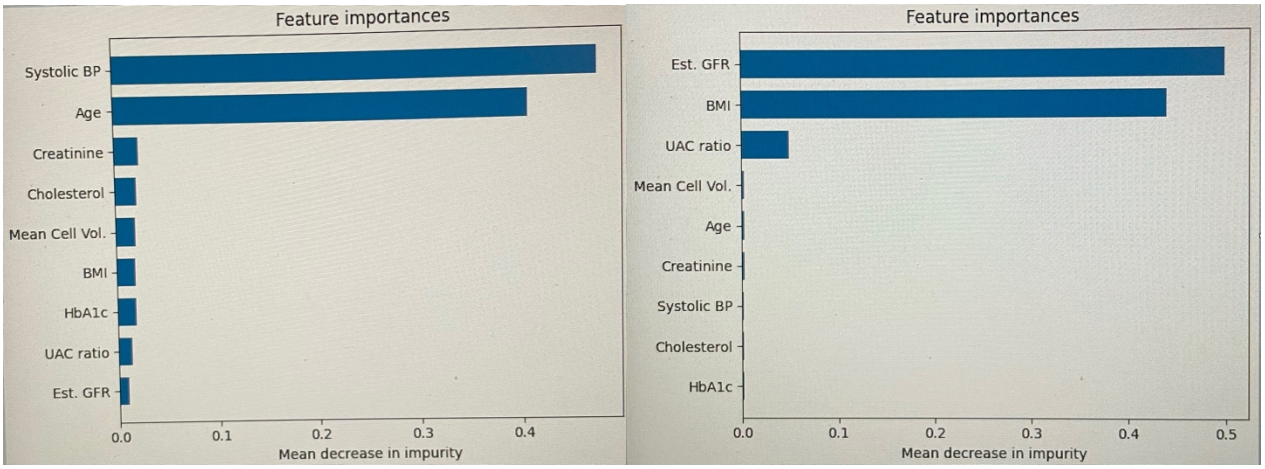
Feature importance. To provide a clinical interpretation of the regression results, we then compared the saliency of the features used in the RF predictions to discern whether the synthetic model makes use of clinically sensible features. We obtained this information by computing the impurity decrease in each tree for both RF models and plotted the features in descending order of importance. We can observe in Figure 6 that, although the features used by the real and synthetic RF models are not precisely matched, the synthetic model generally depends upon sensible clinical features when making predictions. Strangely, despite showing the poorest regression results, the creatinine feature appears to depend upon the same predictors in both the real and synthetic datasets.



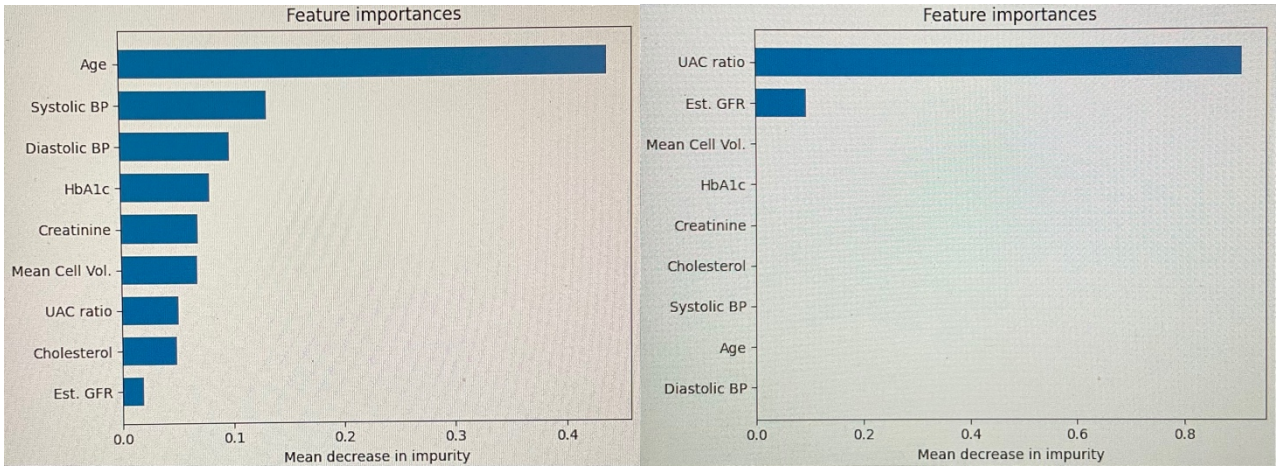
(a) HbA1c



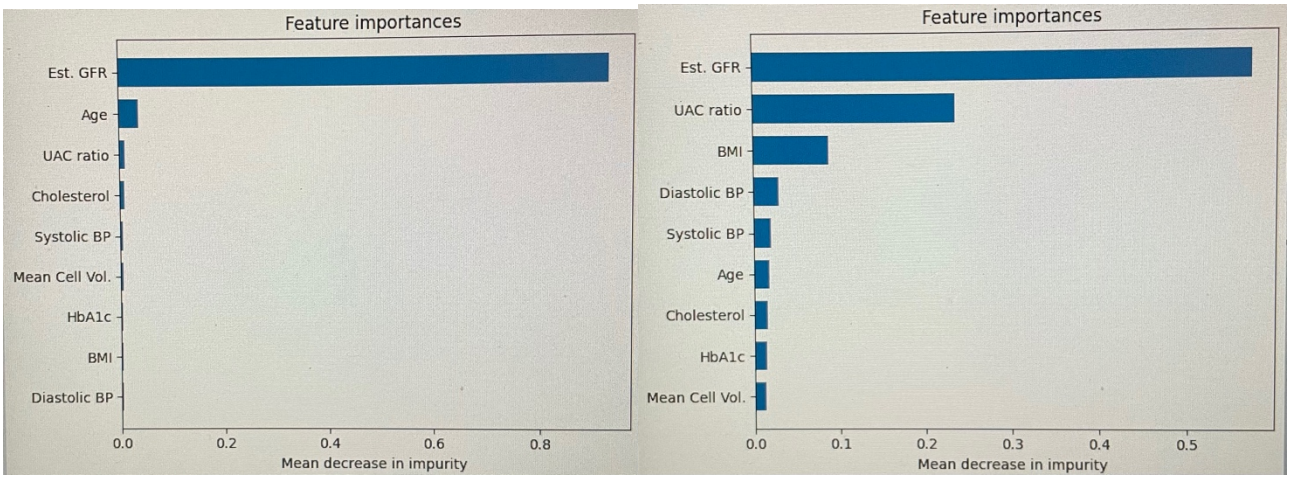
(b) Systolic blood pressure



(c) Diastolic blood pressure



(d) BMI



(e) Creatinine

Figure 6: Most important features (ranked in descending order) used for predicting outcomes with random forest (RF) regression. RF models were trained with real (left) and synthetic (right) datasets to predict outcomes (a) to (e).

4.4. Conclusions and future steps

In summary, we have completed the following work since the last deliverable:

- **Regularisation**. We established dropout as an effective approach to help mitigate the early signs of mode collapse during training. By using dropout in all layers of the discriminator, this appeared to help prevent the discriminator from overfitting too quickly and causing the loss to converge to zero. Furthermore, we also found that the original MMD loss function contributed to the overfitting problem. As a result, we removed two sources of mode collapse, which gave rise to improved synthetic data distributions.
- **Data processing.** In the previous deliverables report, we encoded the drug as a single binary variable. In this deliverable, we presented a new approach that instead separates the single variable into multiple nodes to represent the inherent class that each drug belongs to. With this approach, we showed that we can disentangle the effects of each drug by instead considering the global effects of each independent drug class. This led to the discovery of relevant edges between drug classes (e.g., from GLP-1 to systolic blood pressure, which relates to LEAD-5). However, our data still represents a single 'snapshot' in time, which limits the amount of causal information we can infer about the drugs.
- **Results.** We investigated the quality of the generative model using three approaches: the statistical properties of the synthetic data; the causal structure between the synthetic features; and the efficacy of the synthetic features for machine learning prediction tasks. Overall, our current model yields promising results in all aspects, though there are clear routes for improvement. For example, we observed that the model converged on the dominant drug classes, instead of generating realistic drug combinations. This forced the model to learn causal structure and synthetic features for patients with limited characteristics, ignoring patients that were less populated in the training dataset.

The following steps will be taken at the next stage of the project:

- **Data preparation**. The next phase will remove the need to isolate patients based on presence of metformin. We also intend to disentangle underlying placebo drugs from the nil class, which will help to disentangle spurious connections.
- **Larger patient population.** At the moment we have approx. 56,000 patients, which reduces to around 14,000 after data cleaning. To facilitate repeatability and reproducibility of experiments within a reasonable time frame, we have elected to train on 5000 patients from this filtered cohort. The next step will implement a more efficient model that reduces training times, meaning we can include a greater population into training without sacrificing repeatability. Further, with this model, we have the option to training with missing data which will increase the population size considerably.
- **Time-series data.** Incorporating time-series data will enable our models to capture more representative causal structure from the data. Currently, we have a simple pre- and post-treatment data series that will allow us to draw closer comparisons to the LEAD-5 trial with propensity score matching.
- **Model architecture**. We have developed a novel generative architecture for causality learning that makes use of the latest research in differentiable DAG sampling [6]. In contrast to the model studied in this report, which requires patient data at the input layer, this creates a way to guarantee privatised synthetic data with a DAG.
- **Results for privacy preservation**. With the new model architecture, we will integrate advanced mechanisms for differential privacy into training to privatise the generator.
- **Automated hyper-parameter optimisation.** 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.

- **Further evaluation.** We will evaluate synthetic population with additional metrics including Memorization Informed Fréchet Inception Distance (MiFID)[5] and causality estimation using propensity score matching. With the latter approach, we can begin to draw direct comparisons with the LEAD-5 results.

References

1. 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 .
2. 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
3. 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
4. NHS Greater Glasgow and Clyde & NHS Tayside Health Board area data https://www.nhsggc.org.uk/about-us/professional-support-sites/glasgow-safe-haven
5. 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,
6. Charpentier, B., Kibler S., and Gunnemann S.: Differentiable DAG sampling. ICLR 2022.
7. Lin, Z., Khetan, A., Fanti, G., and Oh, S.: PacGAN: the power of two samples in generative adversarial networks. NIPS 2018.

Appendix

We provide supplementary material to support the discussions in the main text.

A. Further details

A.1. Training results

In this section, we include detailed training results for the model from the main text.

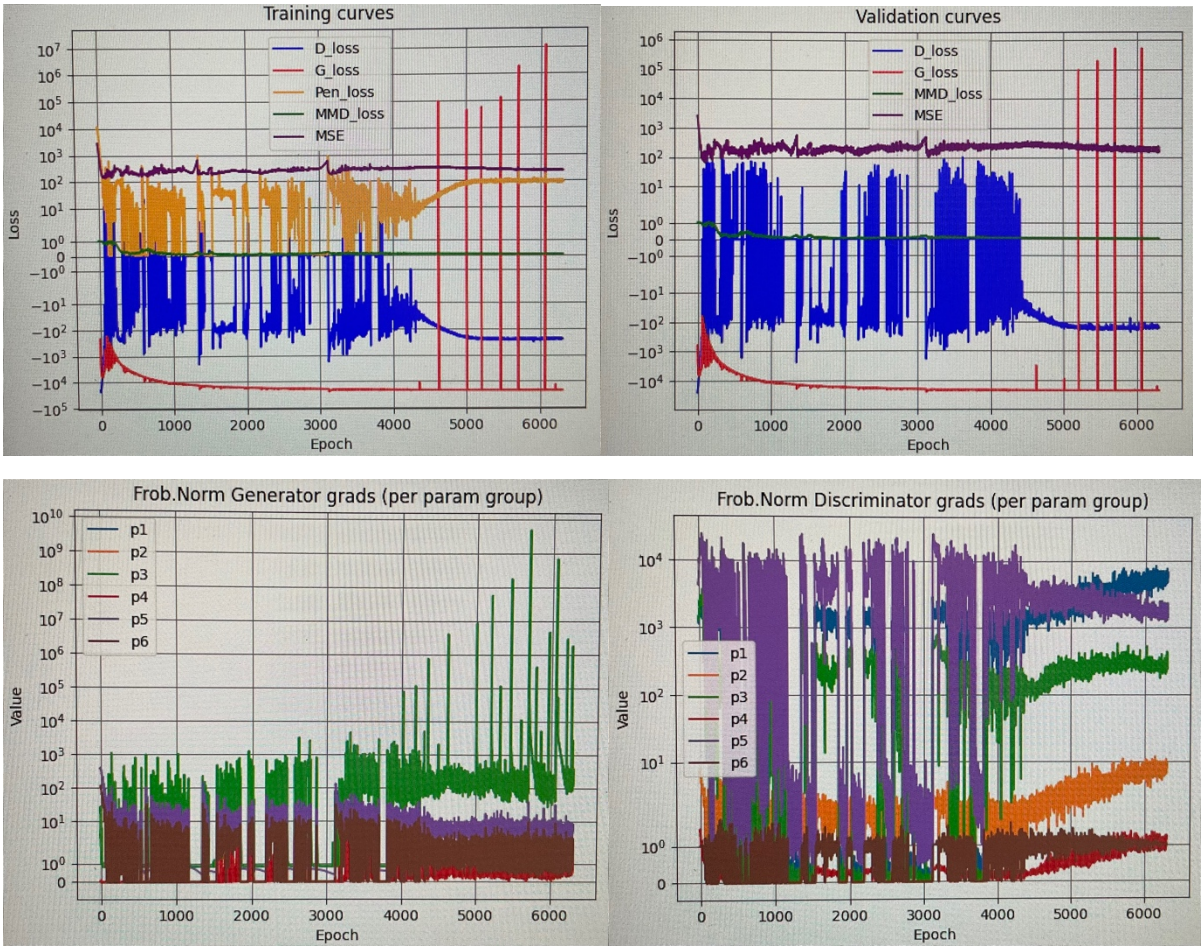


Figure A: *Training results for the main model, including the loss curves (top) and gradient magnitudes (bottom).* For the gradients, we computed the Frobenius norm over all gradients in each parameter group.

A.2. Hyper-parameters

Table A: List of model and optimisation hyper-parameters. **Note**: the (+1) in the generator network structure indicates the concatenation of the 19-dim activation vector with the 1-dim latent noise variable (z), which is mapped to a 1-dim output feature. This is repeated for each of the 19 synthetic variables (hence x19).

Hyper-parameter	**Architecture**	
	Generator	**Discriminator**
Learning rate	3e-3	3e-3
Batch size	100	100
First moment (β_1)	0.9	0.5
Second moment (β_2)	0.999	0.9
Dropout (all layers)	0.0	0.2
Weight decay	0.0	1e-6
PAC	n/a	10
c: Coefficient for absolute value of h(A)	1.0	n/a
Lambda: DAG constraint for h(A)	0.0	n/a
Latent dim (z)	1	

Network structure (MLP)	FC1: 19-361 FC2: 19(+1) -1 (x19)	n/a 190-256-256-10
Updates per mini-batch	1	1
Epochs	300	300

B. Additional results

As described in the main text, the results presented so far were produced using the generator weights after the 12th Lagrangian iteration in training (approx. epoch 3,500). With these weights, we found that relevant connections from the drug classes to the outcomes of interest were recovered by the model, which is desirable for drawing comparisons with the LEAD-5 trial (e.g., edges between GLP-1 and blood pressure). In this section we repeat our analysis but instead using the generator weights at the final Lagrangian iteration (approx. epoch 6,500). As we will show, both synthetic data and causal structure change later in training, and that that longer training times do not necessarily correspond with better results. We believe this highlights the need for a possible early stopping mechanism, which is common in GAN training to avoid making inferences with an overfit model.

The results for the final generator weights are provided in the following series of Figures.

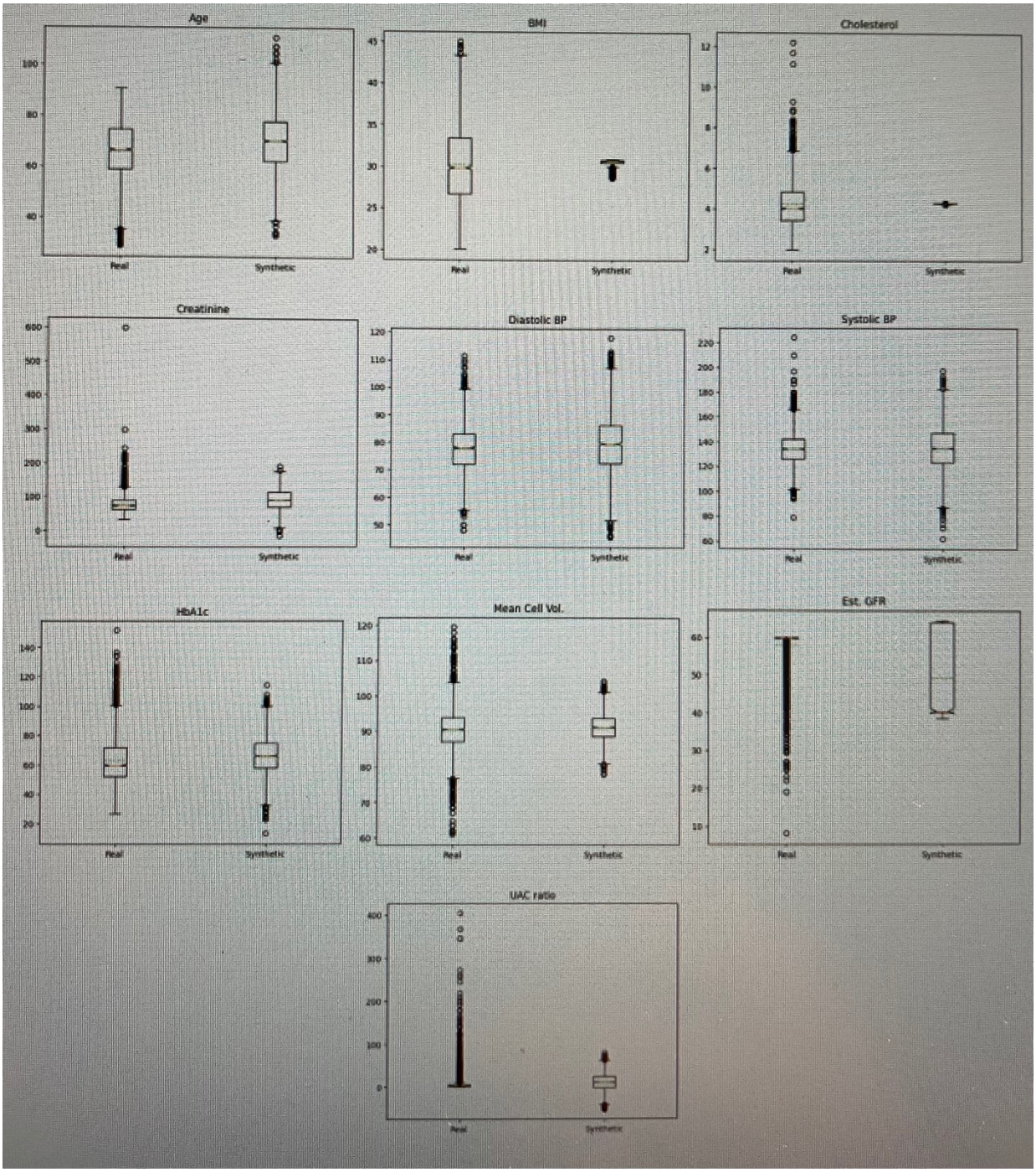
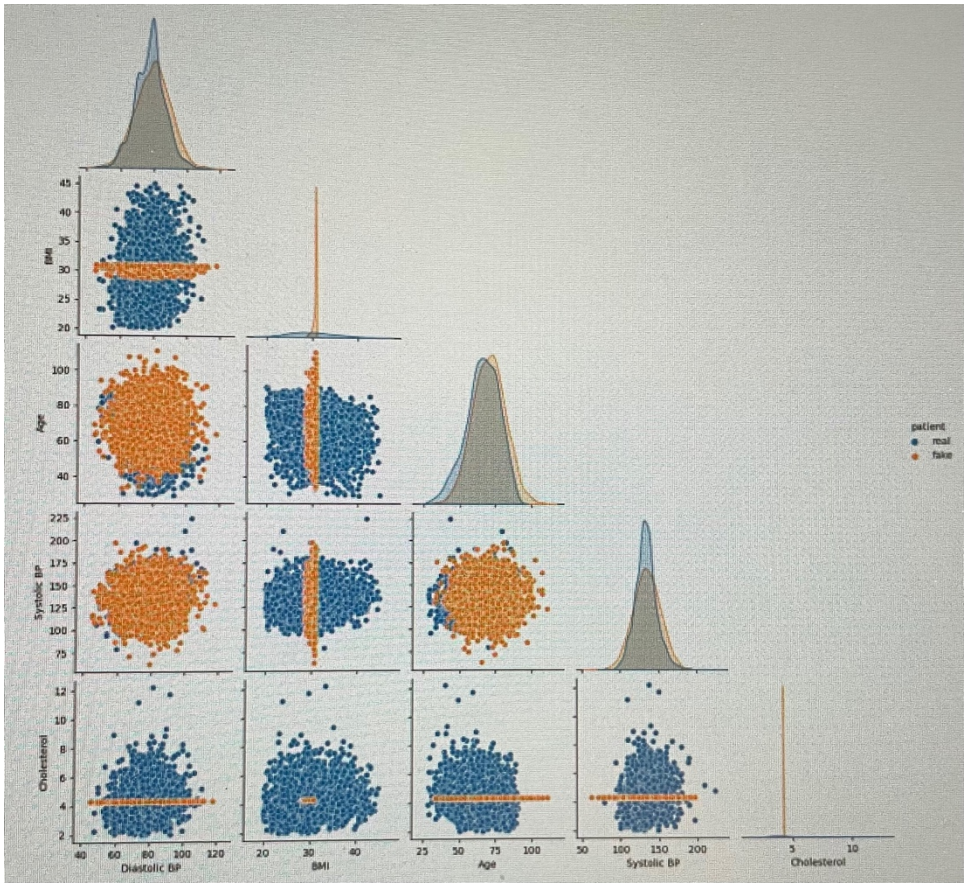
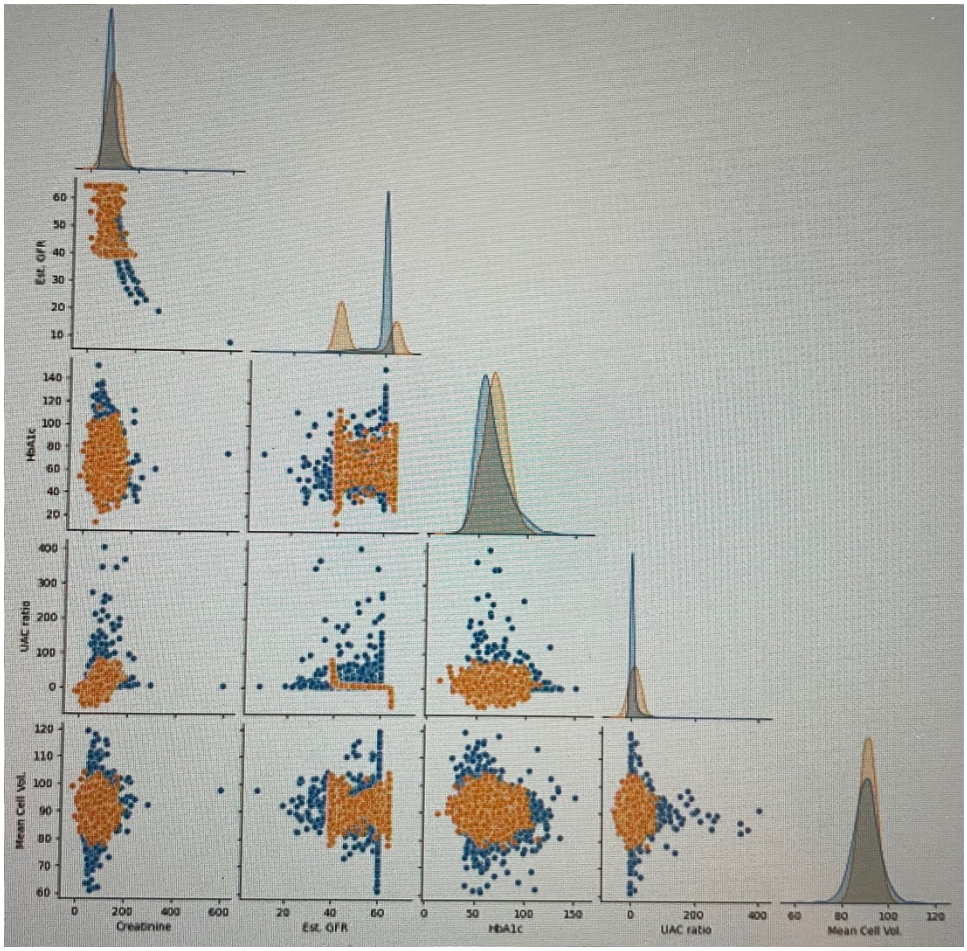


Figure B.1: *Boxplots for the synthetic data generated by the final generator weights.* In comparison to the results for the 12th iteration weights (Figure 4.2), we observe that mean cell volume, diastolic blood pressure, and systolic blood pressure are modelled significantly better with more training. On the other hand, we observe that BMI has collapsed into the median value of the real data.


patient
real
fake
Diastolic BP
BMI
Age
Systolic BP
Cholesterol

(a)



(b)

Figure B.2: *Marginal distributions for the generated data produced by the final weights, with (a) illustrating the first five features, and (b) the last five features.* Overall, we can see that the statistical properties have been improved over the marginals produced in the 12th iteration weights (Figure 4.3). However, as we will see, the model has lost relevant correlation structure between the features, in addition to all causal relationships with the drug classes.

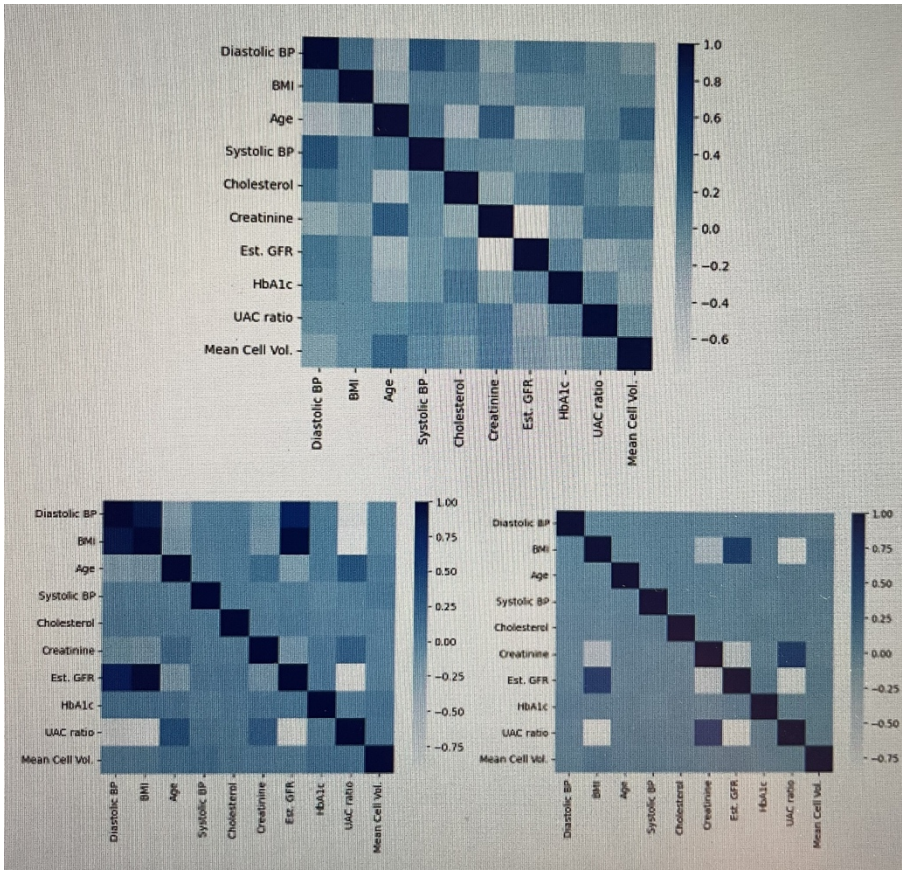


Figure B.3: *Correlation matrices for the real data (top), and synthetic data produced by the 12th iteration weights (bottom, left) and final iteration weights (bottom, right)*. Although we observed improved independent statistical structure, the additional training times has unfortunately removed much of the relevant correlation structure observed in the 12th iteration weights.

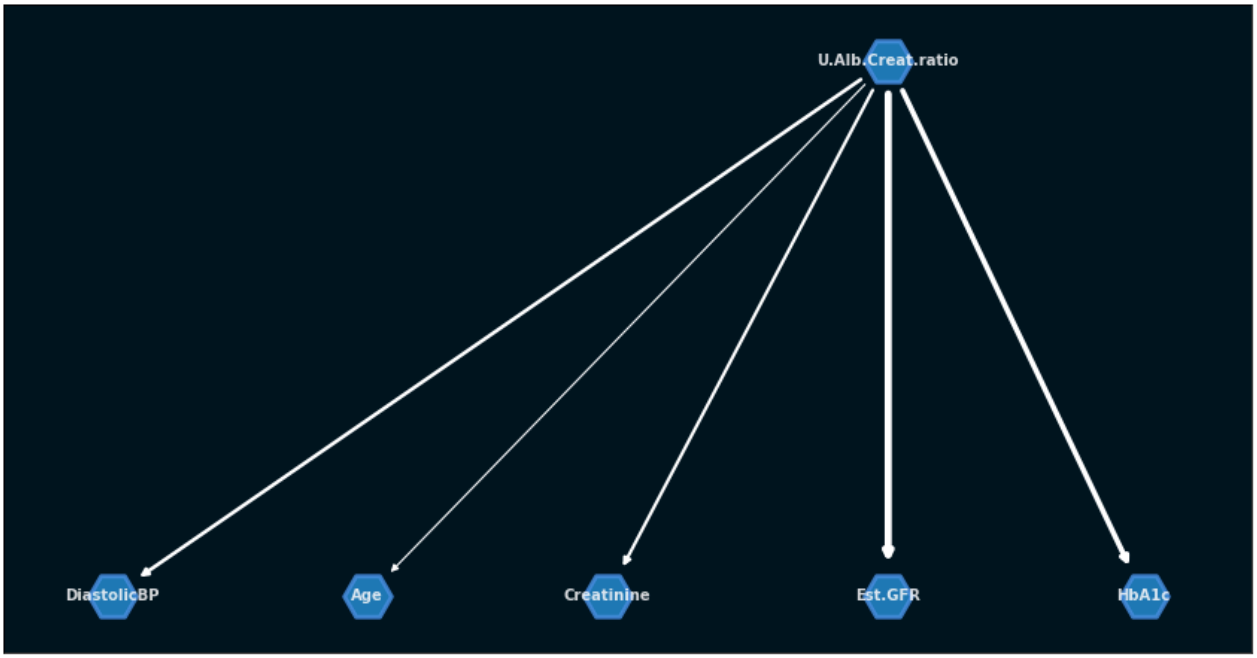


Figure B.4: *Causal graph obtained from the final generator weights.* In contrast to the graph produced from the weights at iteration 12, we no longer see connections between any of the drug classes and features.

C. Ablation experiments

The results presented in the main text were the product of a series of ablation experiments to rectify sources of mode collapse. Here, we provide the results to these experiments. Based on our findings, we elected to rely solely on the Wassterstein critic loss (I.e., removed MMD loss), and to use a dropout rate of 0.2 into all layers of the discriminator.

Note: *The choice for using dropout in the discriminator was to minimise the risk of overfitting and mode collapse, which we observed with a dataset of 5,000 patients and screening only metformin patients. For a larger dataset (e.g., 10,000 patients), and with changes to the dataset (e.g., time-series), we may discover that dropout is not needed. Furthermore, the following results were obtained prior to separating the drugs into their respective classes, and instead produced using our initial approach: encoding the drug as a single discrete variable. In this case, the drug could take on the value 0, representing all patients on metformin, and 1 for patients on both metformin and gliclazide. No consideration was given to any of the other drugs patients may have received.*

In in our initial experiments, we made use of the maximum mean discrepancy (MMD) loss that was promoted in a recent GAN-based causality learning work [DAG-GAN paper]. The intuition for using the MMD term in the loss function was to encourage the real and synthetic data distributions to lie close to one another for improving synthetic data quality. Since high data quality is an obviously desired property of our own model, we decided to introduce the MMD loss into our training algorithm. However, with closer inspection, we found a fundamental issue with the MMD term: in the context of GAN training, divergence minimisation between the two distributions occurs within data space, which is typically high-dimensional. This contrasts with variational autoencoders (VAEs), which perform minimisation within the lower-dimensional latent space. As a result of this high dimensionality, it is unlikely that the real and synthetic data distributions have overlapping support, leading to the aggravation of known GAN training pathologies (e.g., unstable training) and causing the discriminator to overfit because the fakes were too easy to detect.

To rectify this issue, we proposed to introduce dropout into each layer of the discriminator. The key idea was to prevent the discriminator from relying too heavily on features that would enable easy fake detections, forcing the model to learn new information about what discriminates real from fake. We therefore trained separate models

and monitored the synthetic features as training progressed. The results to our analysis are illustrated in Figures C.1 and C.2.

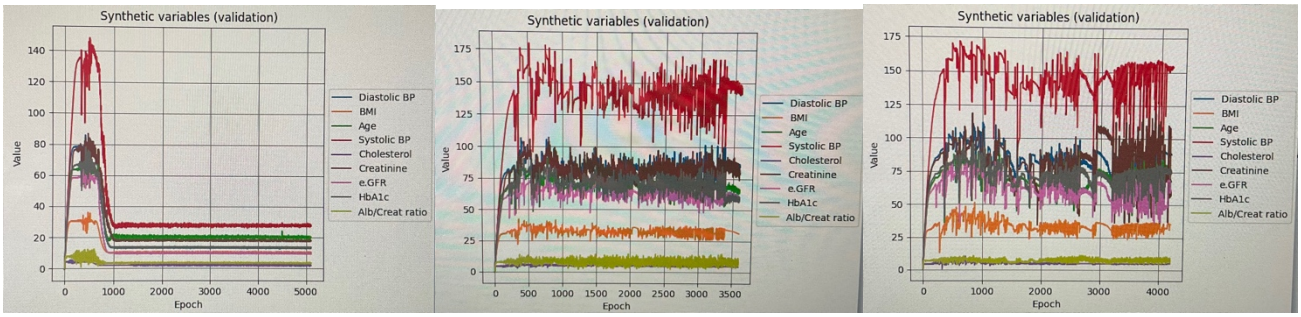


Figure C.1: *Evolution of the synthetic features during training when optimising the MMD loss.* Results using no dropout (left), a dropout of 0.2 (middle), and a dropout of 0.5 (right) in the discriminator. When training with the MMD loss, if we do not use any dropout in the discriminator (left), the model collapses to incorrect feature values after a few Lagrangian iterations. By introducing some dropout, we observe improved results: the degeneracy is removed together with greater sample diversity in the features, particularly as dropout is increased. However, we chose a dropout of 0.2 instead of 0.5 since we found that it struck a suitable balance between improving sample diversity and data quality without introducing too much noise into training.

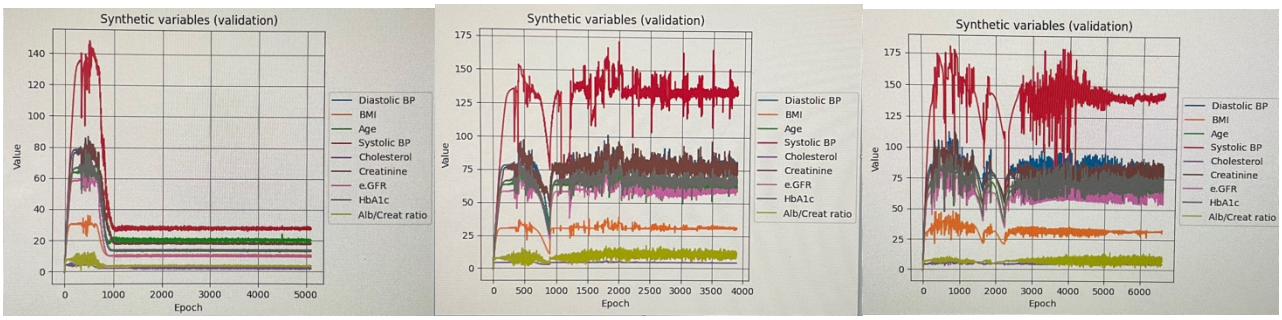


Figure C.2: *Evolution of the synthetic features during training when optimising the MMD loss.* Training with MMD loss and no dropout (left), using **no** MMD loss and no dropout (middle), and using **no** MMD loss but with a dropout of 0.5 in the discriminator (right). The degenerate behavior introduced by the MMD (left) is resolved by removing this loss term (middle). These benefits are then coupled with greater sample diversity when also using dropout (right), but potentially introducing other instabilities.