MRC-GAN: Virtual Trial Emulations and Outcomes

Abstract This report describes the virtual trial emulation outcomes of the MRC-GAN research project, including (**a**) evaluation of the synthetic data sampled from the virtual trial emulation models and (**b**) comparison of the outcomes from the virtual trial emulations against the LEAD-5 trial and (**c**) extended counterfactual emulations to predict the effect sizes on real patient data. We have mainly conducted two types of experiments on the virtual trial emulations in the context of T2DM treatment with three different drugs, namely *GLP-1, basal insulin and placebo*. The first type of experiments is focused on the replication of the existing LEAD-5 trial, and the second type of experiments attempts to emulate counterfactual scenarios where different drugs are applied to the same patients to supply evidence for clinical decision making. The effect sizes are estimated with both average treatment effect and difference-in-differences between pairwise drugs. Our trial emulations show that when the patients meet the LEAD-5 patient baseline characteristics, the trial emulations produce the same ranking between the three drugs as what LEAD-5 have concluded. We have experimented with independent sampling of virtual patients for the three treatment (drug) groups, counterfactual emulations on the same group of virtual patients, and counterfactual emulations on real patients. All results have all suggested that *GLP-1* has the best performance in terms of HbA$_{1c}$, systolic blood pressure and BMI reduction if the patients meet the inclusion criteria of LEAD-5. However, the experiments with real patients who do not fall into the baseline characteristics of LEAD-5 have presented different performance rankings between the drugs. These results suggest that the LEAD-5 trial outcomes cannot be simply extrapolated to cover other patient populations. To this end, the virtual trial emulation models and tools are potentially very useful in terms of providing evidence to support the extrapolation of clinical trials for real-world clinical practice.

The report is organised as follows: Section 1 gives the general background of the project. Section 2 presents the results from the assessment of the synthetic data quality ; and Section 3 shows trial emulation process and their results. Section 4 draw the conclusions.

1. Project Background

Health data contain important knowledge that enables clinical research to assess treatment effect in real-world settings. However, there are significant limitations in real-world health data: 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. More importantly, observational studies with real-world health data do not involve hypothetical interventions, and researchers cannot test their hypotheses on treatment effect from different drugs and treatments with the data that are collected retrospectively.

The MRC-GAN project is designed to investigate an alternative approach to support clinical research through the use of synthetic data. We study the feasibility of running virtual clinical trial emulations to extrapolate randomised clinical trials to cover real-world populations, which supports experiments with hypothetical virtual interventions to answer a range of clinical questions with respect to treatment effects. The emulations generate synthetic populations that preserve the same value for research as real patient data under the support of the latest generative AI and causality learning technology. We have studied the feasibility of this trial emulation approach through a specific use case in the context of Type 2 diabetes mellites (T2DM). The AI model has been trained with the SCI Diabetes data on the Safe Haven platform [1]. 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.

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 these research questions, we have carried out experiments to assess the synthetic data quality and compare the trial emulation outcomes with the LEAD-5 trial [2], which is an existing trial that we have tried to emulate under a confirmative study framework to test the trial emulation models.

2. The Emulation Model

2.1 Learning core causality model

To run trial emulations with either observational or synthetic data, we need to build a simulation model that captures causal relations between multiple variables through causality learning. In this project, this is achieved by learning how to generate data in a generative process.

Notation: We use a random vector $\boldsymbol{X} \in \mathbb{R}^d$ to denote an observation with *d* variables, X_i, *i* =1,…,*d*, and $\widetilde{\boldsymbol{X}} \in \mathbb{R}^d$ to denote synthetic data. $P(\boldsymbol{X})$ and $P(\widetilde{\boldsymbol{X}})$ are their distributions.

Further, let $G = (V, E)$ denote a directed acyclic graph (DAG) with *d* nodes in space $\mathbb{D}$. $A \in \mathbb{R}^{d \times d}$ is the adjacent matrix to represent *G*, where $[A]_{ij} \neq 0$ indicate the existence of a weighted directed edge between vertex i and j. $\boldsymbol{f}(\boldsymbol{X}; \boldsymbol{Z}; W^1, \dots, W^L)$ denotes a vector function with input $\boldsymbol{X}$ and $\boldsymbol{Z}$, and $W^1, \dots, W^L$ are its parameters (i.e. weights in a L-layer neural network). Note, bold text (e.g., ***f***) stands for a vector with its scalar components f_i. Namely $\boldsymbol{f} = (f_1, \dots, f_d)$. *m* denotes the dimension of noise vector $\boldsymbol{Z}_i$.

Formally, we describe the causality learning problem to build the model, together with the assumptions involved and the learning process as follows:

Problem statement: Given *n* independent and identically distributed observations ***X***, we learn a DAG $G \in \mathbb{D}$ to match the underlying joint distribution $P(\boldsymbol{X})$ of the observations. *G* entails a structural equation model (SEM) that describes the data generative process f_i: $\mathbb{R}^d \rightarrow \mathbb{R}$ at each node X_i,

$$\mathbb{E}[X_i \mid X_{\text{pa}(i)}] = g_i(f_i(\boldsymbol{X})) \tag{1}$$

where pa(*i*) denotes the parents of node X_i in *G*. g_i: $\mathbb{R} \rightarrow \mathbb{R}$ is the so-called *link function.*

Assumptions: We make several basic assumptions for the causality learning (**a**) *Faithfulness*: The variables in the dataset are probabilistically dependent if they are causally connected in the underlying causal graph. This assumption allows us to learn causal graph from the data distributions; (**b**) *Causal sufficiency*: There is no unobserved confounder that produces bias in the estimated causal effect. This assumption allows us to infer causal graph with observed data distributions only; (**c**) *Model identifiability*: the link function within the causal model is an additive noise model (ANM). Namely, we select link function g_i in Eqn. (1) to be an additive noise model to realise noise $\boldsymbol{Z} \in \mathbb{R}^{d \times m}$ $\boldsymbol{Z}_i \in \mathbb{R}^m$ sampling in the generative process:

$$\widetilde{\boldsymbol{X}}(\tilde{X}_1, \dots, \tilde{X}_d) = \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{Z}; W^1, \dots, W^L) = \boldsymbol{f}(\boldsymbol{X}; W^1, \dots, W^L) + \boldsymbol{Z} \tag{2}$$

$$\tilde{X}_i = f_i(\boldsymbol{X}; \boldsymbol{Z}_i; W_i^1, \dots, W_i^L), i = 1, \dots d$$

where $\widetilde{\boldsymbol{X}}$ are synthetic data samples. According to [Hoyer et al 2009], this additive noise nonlinear model is identifiable if f_i is three times differentiable and none-liner. With this identifiability assumption, we can uniquely identify the underlying DAG from the data distribution.

Temporal constraints: Temporal information provides a natural causal order (i.e., the proceeding variable X_j of X_i cannot be its cause). This imposes the so-called temporal causal constraint.

By applying the temporal causal constraints, we divide the variables into 3 main categories, including treatments, post-treatment measurements and the confounders that involve patient demographics, pre-treatment measurements and pre(vious)-treatments. The overall causal structure is shown *in Figure 1*. In addition to the direct causal link between treatment and post-treatment measurements, we account for confounding effects from demographics, pre-treatment measurements and previous treatment to the treatment assignments and post-treatment measurements. Through learning from the data distribution $\boldsymbol{X}$, we infer the exact causal graph structure between the variables, together with the structural equations $\boldsymbol{f}$ that are associated to the graph to enable synthetic data generation for trial emulation.

Most of the confounders and post-treatment measurements are continuous variables. We consider each drug as a separate discrete variable. To reduce the number of possible drug combinations, and to account for the confounding effects imposed by other prescribed drugs, we categorise all drugs into their corresponding classes. This allows 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. Thus, the treatment (drug) nodes are represented by binary variables. Each drug is associated with one designated binary node. The node is set to 1 if the associated drug is applied to patients in the treatment, and 0 otherwise.

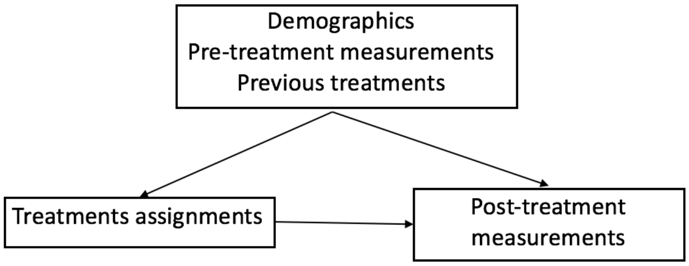


Figure 1 Overall causal structure between treatment, post measurements and confounders

Causality learning with adversarial loss To learn the model from $\boldsymbol{X}$, we combine causality learning and generative adversarial learning to simultaneously learn causal structure and functions for synthetic data generation, We use the Wasserstein GAN with a gradient penalty loss[3]. 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. The causality learning follows the general framework that was recently proposed by [4] by applying acyclic constraints to the DAGs learning. Under the faithfulness and model identifiability assumptions, the generative process can only produce same data distribution $\widetilde{\boldsymbol{X}}$ equivalent to $\boldsymbol{X}$ if the model entails the correct causal structure.

Conditional Generator: Without loss of generality, we use neural networks to approximate and learn $\boldsymbol{f}$ in the generative model Eqn.(2). Specifically, each variable X_i is modelled with a fully connected neural network of L hidden layers $f_i(\boldsymbol{X}; W_i^1, \dots, W_i^L)$, where W_i^l is the weights (&bias) of the lth layer. Given observations $\boldsymbol{X}$, $\boldsymbol{f} = (f_1, \dots, f_d)$ is learned through optimisation of the function parameters $\left(W_i^1, \dots, W_i^L\right)$, $\forall i = 1, \dots d$. $f_i\left(\boldsymbol{X}; Z_i; W_i^1, \dots, W_i^L\right)$ is a conditional generator**.**

Discriminator: The discriminator D_θ (θ denotes its parameters) takes either $\boldsymbol{X}$ or $\widetilde{\boldsymbol{X}}$ to measure the distance between the distributions $P(\boldsymbol{X})$ and $P\left(\widetilde{\boldsymbol{X}}\right)$.

Loss function: The loss function L involves the generative adversarial loss term. It also involves a gradient penalty term (WGAN-GP) [Gulrajani2017] as follows:

$$L = \mathbb{E}_{\boldsymbol{X} \sim P(\boldsymbol{X})}[D_\theta(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{Z} \sim P(\boldsymbol{Z})}\left[D_\theta\left(\boldsymbol{f}(\boldsymbol{Z})\right)\right] + \lambda \mathbb{E}_{\widetilde{\boldsymbol{X}} \sim P(\widetilde{\boldsymbol{X}})}\left[(\| \nabla_{\tilde{x}} D_\theta\left(\widetilde{\boldsymbol{X}}\right) \|_2 - 1)^2\right] \tag{3}$$

where $\boldsymbol{Z} \sim P(\boldsymbol{Z})$ is a process to sample Z_i at each generator. As explained in Eqn.(3), the noise sampling is implemented with ANM in this work.

The adversarial loss training minimises the difference between the true data distribution $P(\boldsymbol{X})$ and synthetic data distribution $P(\widetilde{\boldsymbol{X}})$ by discovering the right causal structure (DAG) in the generative process (Eqn.(1) and (2)). Under the ANM assumption (which is an identifiable model), we can only achieve global minimum (i.e. $P(\boldsymbol{X}) = P(\widetilde{\boldsymbol{X}})$if a true causal structure is discovered.

Acyclic constraints: The optimisation is subject to an acyclic constraint $h(A) = tr(\exp(A \circ A)) - d$ [4] or $h(A) = tr\left(\left(\mathrm{I} + A \circ \frac{A}{d}\right)^d\right) - d$ [5,6], where A is the adjacent matrix to represent G. Similar to [6, 7]. The adjacency matrix A that represents G is defined implicitly through the weights of these neural networks – more specifically, we follow the method in [6] by defining $[A]_{ij}$ as the l^2norm of the jth column in W_i^1, which determines whether X_j is a cause of X_i.

Synthetic data generation through sampling After training, the generative causal model (Eqn.(1) and (2)) allows synthetic data generation via data sampling from the learned distribution $P(\tilde{\boldsymbol{X}})$. Specifically, the conditional generator at node i: $\tilde{X}_i = f_i(\boldsymbol{X}; \boldsymbol{Z}_i; W_i^1, W_i^2)$ takes samples from its conditional probability distribution $\boldsymbol{P}(X_i | X_{\mathrm{pa}(i)})$. Together, this allows sampling to generate samples $\tilde{\boldsymbol{X}}$ from the underlying joint distribution $P(\boldsymbol{X})$, which is factorised with the local conditional probability distributions.

2.2 Training & Technical validation

2.2.1 Datasets

The dataset consists of an array of clinical variables, drug prescriptions, and demographics variables for 56,476 unique patients with T2DM. Table 1 provides an overview of the datasets involved in the study. 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.

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.

Pre-processing. For our analysis, we prepared the SCI-Diabetes dataset to represent each patient's record as a series of pre- and post-treatment measurements collected over time in response to different treatments, where each new treatment marked a distinct point in time. Note some patients received multiple drugs at one time. Pre-measurements were collected from 9 months prior to treatment, and post-measurements from 12 months following treatment. If multiple measurements were taken during these periods, we use the median values. On this basis, we made the simplifying assumption that each set of pre- and post-features were independent of those at previous time-steps, but whose pre- to post- change is confounded by all treatments given up to that time. This allowed us to accumulate the treatments over time to form a new 'pre-treatment' feature to encode the passing of time since the beginning of treatment.

Missing data. After pre-processing, we observed a significant proportion of missing data. Performing a hard removal of these elements, where patients with either pre- or post-features missing from any category were removed from the dataset, reduced the number of available datapoints from 72,958 to 6,674. From this set we randomly chose 6,500 samples for model training, upon which we performed a 90/10 split to partition the set into training and validation subsets for optimisation and evaluation, respectively.

2.2.2 Training

Model and training parameters. To help mitigate the risk of mode collapse, we implemented a PAC-based discriminator [9] with a PAC value of 10. Both discriminator and generator networks were optimised with Adam using initial learning rates of $3\text{x}10^{-4}$, that were decayed to $7\text{x}10^{-6}$ using a cosine annealing schedule. The nature of the Lagrangian optimisation algorithm imposes a cyclical process for updating the Lagrangian multiplier that repeats every *N* epochs. We chose to set *N* to 300 and found that performing a warm restart of the learning rates at the beginning of each new cycle yielded more favourable learning dynamics, and a model with better causal structure, compared to training under a simple step decay. We refer the reader to Appendix A for a comprehensive list of our hyper-parameter settings.

We monitored the training behaviour of the networks using the gradient penalty (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. Figure 2 shows the training and validation loss curves for MRC-GAN.

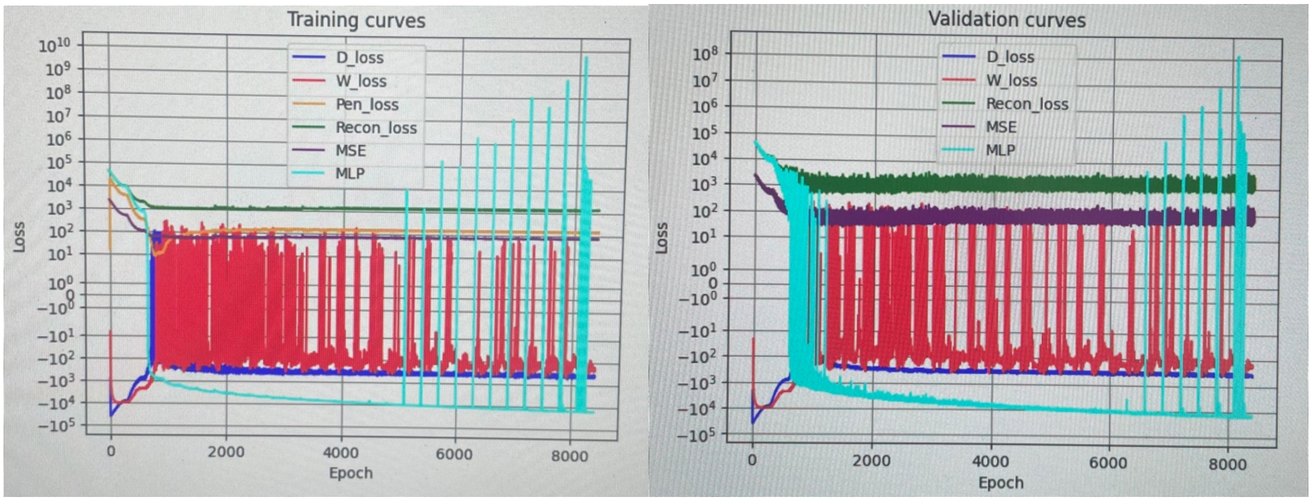


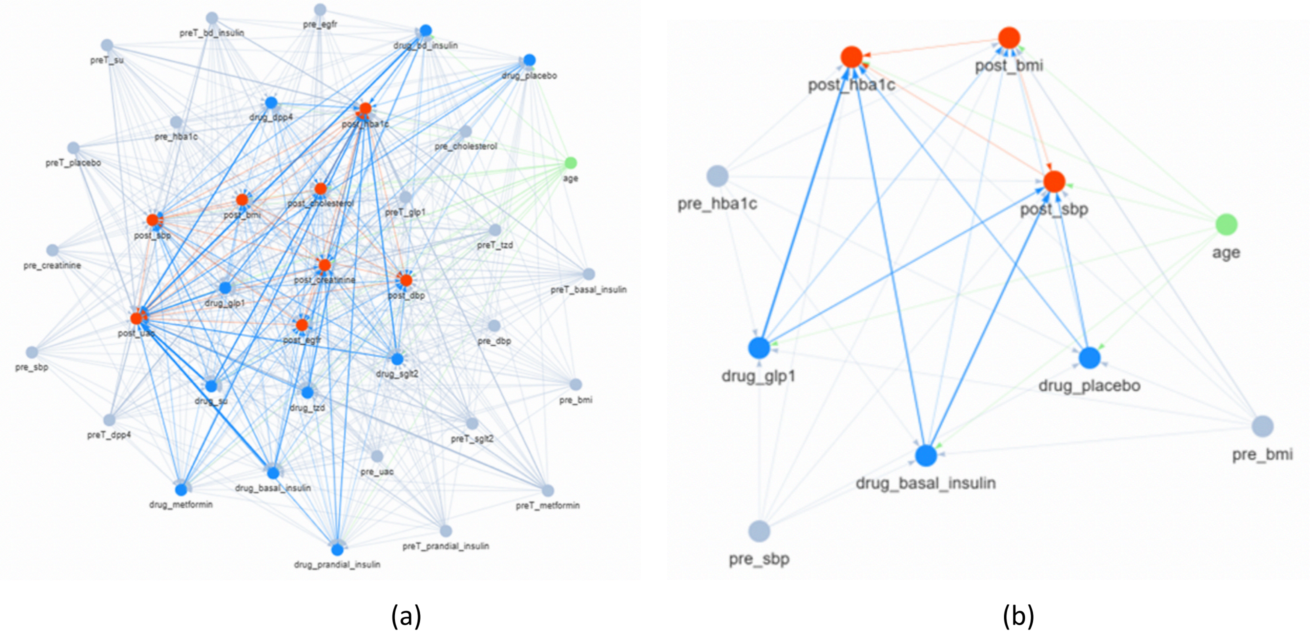
Figure 2. Learning curves for MRC-GAN.

2.2.3 Validation

This is to answer the first question on synthetic data and generative AI models:

- *Can we use generative AI models to generate synthetic data that preserve the same value for research as real-world health data?*

Causal graph. The graph illustrated in Figure 3 depicts the learned connections in the adjacency matrix inferred from the generator weights.



(a) (b)

Figure 3: Causal graph learned by MRC-GAN, where confounders are shown in grey and green, together with drug (blue) and post-measurements (red). (a) All nodes and edges. (b) Subset of edges relevant to LEAD-5. Line thickness conveys the magnitude of the connection strength between nodes.

The primary goal for MRC-GAN is to learn functions that are able to predict the pre- to post- effect size of each drug, while adjusting for all known confounders in the data. This requires that the model learns connections between all drugs and post-features, while at the same time, maintaining edges from the age, pre-features, and pre-treatment into both drug and post-feature nodes. We evaluate the model outputs in both aspects. Namely, we: (**a**) visually examine the learned causal structure graph that underlines the data distribution; and (**b**) evaluate the quality of the data generated from this graph structure. The generator weights selected for study were obtained at the end of the 23rd (of 28) Lagrangian iteration (~6000 epochs), based on the best validation loss terms and prior to the observed spikes in the learning curves (Fig. 2), in addition to preserving connections between the post-features. However, we observed no significant difference in the subsequent trial emulation results when using weights beyond this point in training.

From the global view of the entire learned graph in Figure 3a, we can see that this structure has been successfully captured by MRC-GAN. Specifically, we observe that all relevant drug to post-feature edges have been preserved (blue), with edges from all confounders (grey and green) retained, in addition to connections within the post-features (red), to adjust for confounding. In Figure 3b, we can draw some initial insights about the effects of each drug that are of relevance to the LEAD-5 trial. Most importantly, we observe that stronger edges appear to exist between *GLP-1* and post-HbA_{1c} than between either basal insulin or placebo and HbA_{1c}. Further, *GLP-1* also appears to have a stronger connection to post-BMI than both glargine and placebo, but shares a similar connection strength to systolic blood pressure. Although the weights of the network depicted in Figure 3 are not commensurate with effect size, this nevertheless gives initial evidence that the model has correctly learned to associate the *GLP-1* class of drugs (which include liraglutide) with a similar set of trends as the LEAD-5 clinical trial. We examine this hypothesis in more detail in Section 3 using virtual trial emulations.

Synthetic data generation. To investigate the quality of the data generated by the graph in Figure 3, we sampled a cohort of synthetic patients from the model to compare with patients from the real dataset. To generate the synthetic patients, we used a random sampling-based approach that assumes the confounders are each independently normally distributed according to their respective criteria from the LEAD-5 trial (see Section 3: inclusion criteria). This ensures that all patients begin the trial with similar features, and to adjust for any confounding effect between the confounders themselves.. Given the randomly sampled confounders, we manually set the pre-treatments and drug nodes and generate the post-measurements for our analysis.

We performed two analyses on the post-measurements generated from MRC-GAN: (**a**) comparing the statistical structure between the real and synthetic features; (**b**) comparing the regression performance of random forest modelling when trained using the real and synthetic features.

Statistical structure

We selected real patients from each treatment arm and generated an equivalently sized cohort of synthetic patients to compare with. Based on the available training data, this resulted in the generation of 391 patients for the placebo, 382 for basal insulin, and 482 patients for the *GLP-1* treatment arms. We then computed the joint distributions between the pre- and post-measurements in both the real and synthetic patients, together with their respective marginal distributions using kernel density estimation (KDE). The results are illustrated in Figure 4.

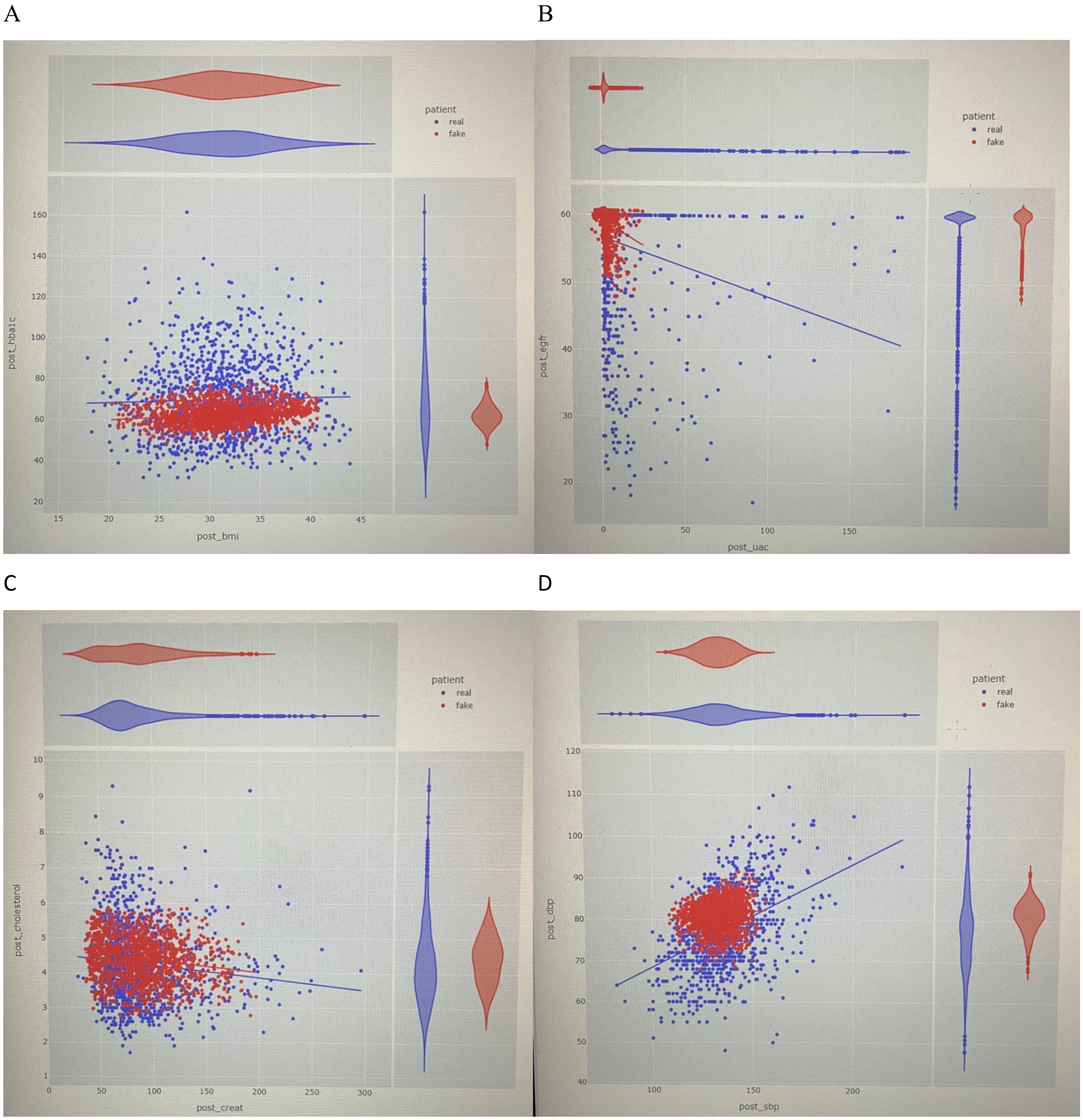


Figure 4. Statistical comparisons between post-measurements from the real and synthetic data. **A**. BMI (p=0.07868) and Hba1c (p<0.0001). **B**. UAC (p<0.0001) and EGFR (p=0.07857). **C.** Creatinine (p<0.0001) and Cholesterol (p<0.0001). **D**. Systolic (p=0.0029) and diastolic (p<0.0001) blood pressure. Significant differences are interpreted as p<0.05, which were computed using the Mann-Whitney non-parametric t-test. Trendlines were modelled with ordinary least squares (OLS) and are included to study correlations between the post-features.

Inspecting the marginal distributions above reveals that, overall, the generator learns a PDF for each post-feature that has overlapping support with the corresponding true PDF. In particular no significant differences were found between the real and synthetic BMI (p=0.07868) and EGFR (p=0.07857) features, although it is clear visually that the distributions for systolic blood pressure (Fig. 4D) and cholesterol (Fig. 4C) closely resemble the true distributions. However, despite converging around their true median values, we observe poorer results for important features such as HbA1c (Fig. 4A) and diastolic blood pressure (Fig. 4D) owing to their smaller sample diversity.

The datapoints in Figure 4 also expose insights into the correlation structure between the post-features, which is useful for understanding whether the model preserves the expected feature dependencies at the output space. From the OLS trendlines provided (formulated as $y = mx + c$), we observe that the relationships between the post-features have been well-preserved overall, especially for HbA1c and BMI (real: m=0.127, c=65,7, R^2=0.001, fake: m=0.247, c=54.9, R^2=0.04), and cholesterol and creatine (real: m=-0.003, c=4.57, R^2=0.01; fake: m=-0.003, c=4.71, R^2=0.03).

Machine learning regression analysis

Lastly, we perform regression analysis on both the real and synthetic data to evaluate the efficacy of the generated post-features for predicting clinical outcomes. We implemented random forest (RF) regression to predict clinical outcomes related to the LEAD-5 trial, given the remaining post-measurements. We trained separate RF models on the real and synthetic datasets, and then assigned a separate hold-out set of real data for evaluating the performance of both trained RF models. This enabled us to compare the ability of both RF models to make predictions given the same test data. For each RF, we used 1000 decision trees with a max depth (number of splits) of 5 per decision tree.

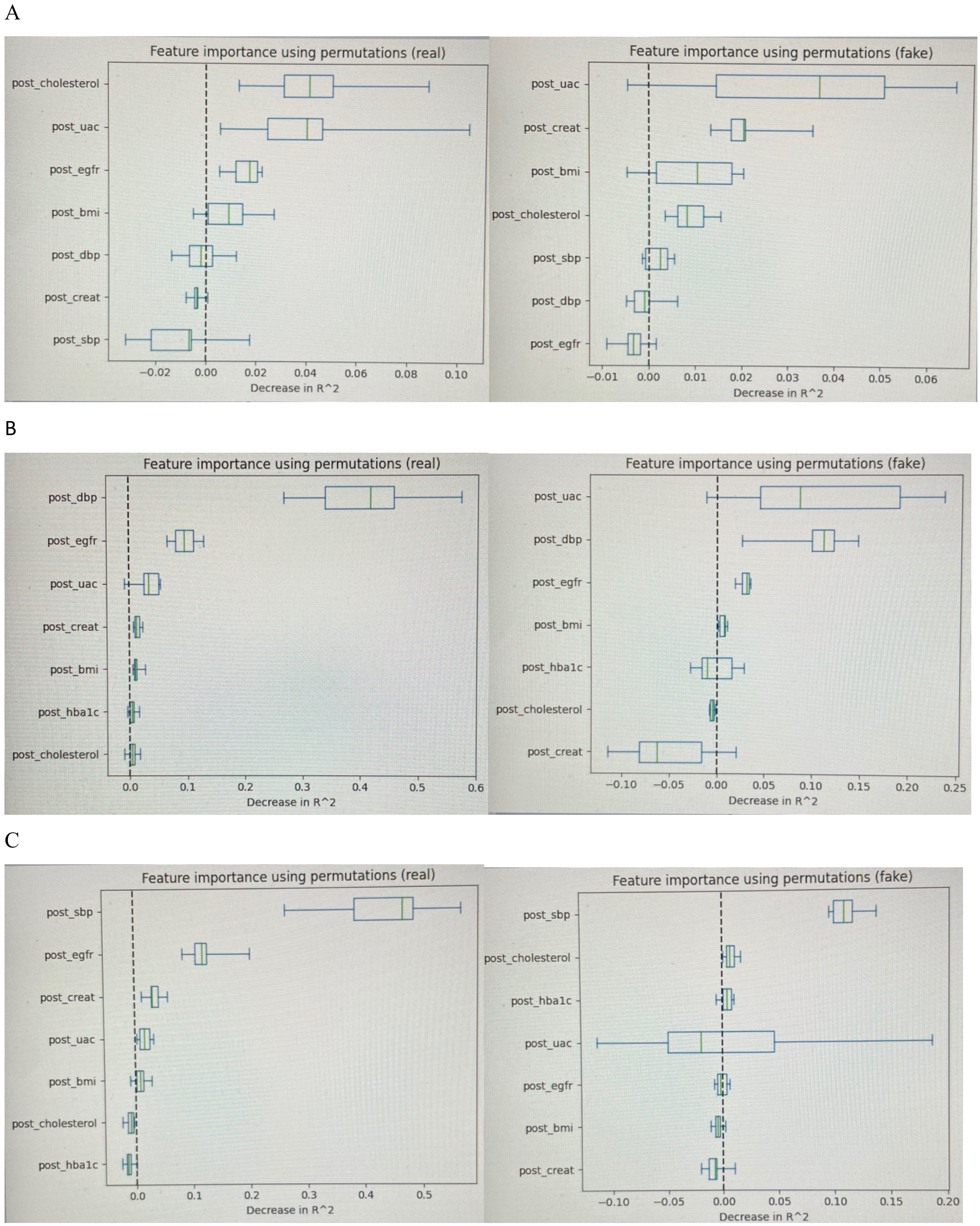
Regression performance. We quantify the performance of both real and synthetic RF models using the mean squared error (MSE) and R^2 metrics, which were computed between the ground truth (y_r) and predicted targets from the real ($\widehat{y_r}$) and synthetic ($\widehat{y_f}$) RF model. The results are provided in Table 2.

Overall, we observe that the synthetic RF model yields similar predictive accuracy to the real RF model when asked to predict HbA1c and BMI, based on their corresponding MSE and R^2values. However, in the case of both systolic and diastolic blood pressure, we observe that the targets show little to no correlation with the fake predictions, indicating that the fake RF struggles to model the dependencies to the blood pressure. Since we expect the confounding variables (i.e., age and pre-features) to contain additional predictive information, we provide results in Appendix B to show the effects of including these features on the real and fake RF performance, and whether the fake RF model captures the relationships between pre- and post-features.

Table 2: Regression performance of the real and synthetic RF models on a test set of real *post-measurement* data. Outcomes align with those studied in the LEAD-5 trial. HbA1c: blood glucose [mmol/mol]. SBP: systolic blood pressure [mmHg]. DBP: diastolic blood pressure [mmHg]. BMI: body mass index [kg/m2].

	HbA1c		**SBP**		**DBP**		**BMI**	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Real	0.057	411.27	0.273	169.41	0.298	56.51	0.036	22.02
Synthetic	-0.052	459.26	-0.039	242.37	-0.248	100.58	-0.078	24.65

Feature importance. In this section, we examine the importance of the test set features used by both RF models to discern whether the fake RF model makes use of clinically sensible features in its predictions. Assuming the features used by the real RF model are the ground truth, we interrogate the fake RF model by permuting each test feature independently, re-computing the outcome, and observing changes in the R^2 value. The output from this process has an intuitive interpretation: features with more importance cause the predictions to be more correlated with the targets, meaning that a decrease to R^2 corresponds to greater feature importance and vice versa. From the results in Figure 5, we can observe that, although the features are not precisely matched, the synthetic RF model generally uses similar clinical features to the real model across all outcome predictions.


A
Feature importance using permutations (real)
post_cholesterol
post_uac
post_egfr
post_bmi
post_dbp
post_creat
post_sbp
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
Decrease in R^2
Feature importance using permutations (fake)
post_uac
post_creat
post_bmi
post_cholesterol
post_sbp
post_dbp
post_egfr
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Decrease in R^2
B
Feature importance using permutations (real)
post_dbp
post_egfr
post_uac
post_creat
post_bmi
post_hba1c
post_cholesterol
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Decrease in R^2
Feature importance using permutations (fake)
post_uac
post_dbp
post_egfr
post_bmi
post_hba1c
post_cholesterol
post_creat
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
Decrease in R^2
C
Feature importance using permutations (real)
post_sbp
post_egfr
post_creat
post_uac
post_bmi
post_cholesterol
post_hba1c
0.0
0.1
0.2
0.3
0.4
0.5
Decrease in R^2
Feature importance using permutations (fake)
post_sbp
post_cholesterol
post_hba1c
post_uac
post_egfr
post_bmi
post_creat
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
Decrease in R^2

D

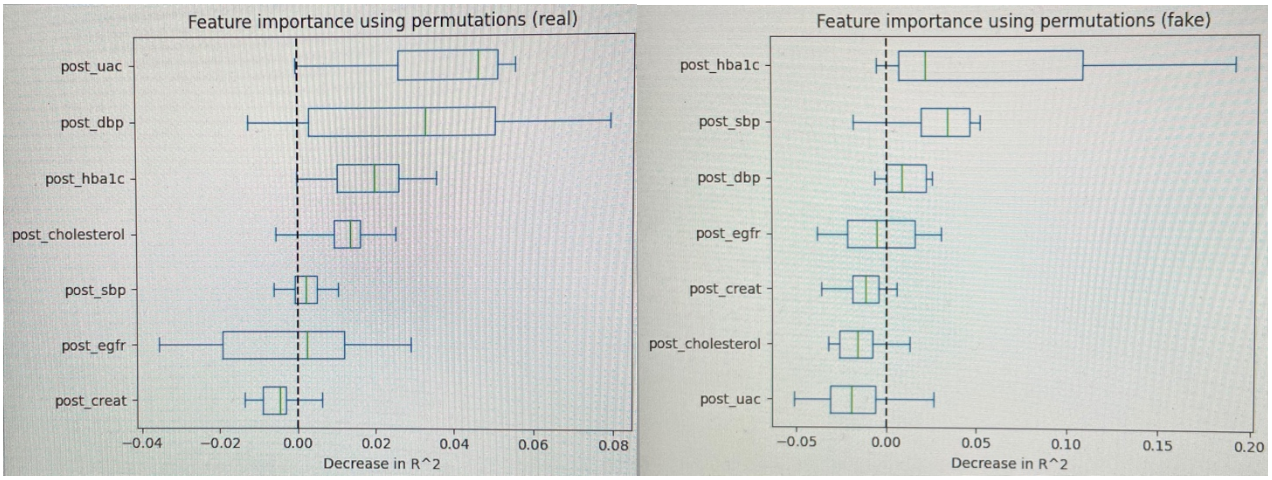


Figure 5. Ranking the most important features (descending order) used by the random forest models, trained on real (left) and synthetic (right) post-measurement data, for predicting the LEAD-5 outcomes: **A.** HbA1c. **B.** Systolic blood pressure**. C.** Diastolic blood pressure. **D**. BMI**.** Boxplots summarise the change in model performance over 10 random and independent permutations to each feature. The dashed line illustrates whether removing the feature is likely to worsen performance (right-hand side) or improve performance (left-hand side), indicating that the feature is more or less important for predictions, respectively. Results when confounding features are included are available in Appendix B.

2.3. Summary and limitations

In our implementation, we constrained the model to only generate post-measurements (e.g., post-hba1c), and found that, overall, the learned causal graph (Fig. 3) produces synthetic data that contains similar statistical structure (Fig. 4) and predictive information (Tab. 2 and Fig. 5) to that of the real data. However, we identified two key limitations to our analysis. First, we sampled synthetic patients from the model using a simplistic random-based approach that assumes no confounding between the pre-measurements. However, this will be an important factor for synthetic data generation since we expect there are such confounding effects in the real patients. A more complete and principled generation approach would be to instead learn the causal structure amongst the pre-features first, and to then generate them on this basis before producing the post-measurements. Second, the training dataset was limited to a sample size of around 7000 due to a large proportion of missing data. This led to the drug classes of interest (i.e., placebo, basal insulin, and *GLP-1*) being under-represented, with only 382, 391, and 482 samples in each, respectively, compared to others (e.g., metformin with c. 1500 samples). Contemporary deep learning typically depends upon datasets of an order magnitude greater than this, together with balanced classes, meaning that with our current experimental setup we limits the amount of causal structure for the model to learn from. To add to this issue, many of the remaining patients in our primary drug classes received additional drugs within a short period of time (~few weeks from beginning of treatment), meaning the true drug effects in such patients could be confounded.

3. Trial Emulation

To evaluate the outcomes of the trial emulations, we have made comparison of the outcomes from the virtual trial emulations against an existing trial (Section 3.1 and Section 3.2), and also conducted extended counterfactual emulations to predict the effect sizes on real patient data (Section 3.3)

3.1 LEAD-5 Emulation

The experiment with trial emulations is to answer the second research question:

- "*Can we perform virtual clinical trial emulations by discovering correct causal relations from the synthetic data?*"

This experiment is performed as confirmatory study in a specific Type 2 diabetes mellites (T2DM) use case. We have emulated an established clinical trial, namely, LEAD-5 (Liraglutide Effect and Action in Diabetes), and compared the emulation outcomes against the published LEAD-5 outcomes to check whether the "virtual" outcomes are similar to the "true" outcomes. The LEAD-5 trial measures the effect of Liraglutide, a *GLP-1* receptor agonist. More details about the LEAD-5 trial can be found in the following Abstract extracted from [2] :

> Liraglutide vs insulin glargine and placebo in combination with metformin and sulfonylurea therapy in type 2 diabetes mellitus (LEAD-5 met+SU): a randomised controlled trial
>
> **Aims/hypothesis** The aim of the study was to compare the efficacy and safety of liraglutide in type 2 diabetes mellitus vs placebo and insulin glargine (A21Gly,B31Arg,B32Arg human insulin), all in combination with metformin and glimepiride.
>
> **Methods** This randomised (using a telephone or web-based randomisation system), parallel-group, controlled 26 week trial of 581 patients with type 2 diabetes mellitus on prior monotherapy (HbA$_{1c}$ 7.5–10%) and combination therapy (7.0–10%) was conducted in 107 centres in 17 countries. The primary endpoint was HbA$_{1c}$. Patients were randomised (2:1:2) to liraglutide 1.8 mg once daily (n = 232), liraglutide placebo (n = 115) and open-label insulin glargine (n = 234), all in combination with metformin (1 g twice daily) and glimepiride (4 mg once daily). Investigators, participants and study monitors were blinded to the treatment status of the liraglutide and placebo groups at all times.
>
> **Results** The number of patients analysed as intention to treat were: liraglutide n=230, placebo n=114, insulin glargine n= 232. Liraglutide reduced HbA$_{1c}$ significantly vs glargine (1.33% vs 1.09%; −0.24% difference, 95% CI 0.08, 0.39; p= 0.0015) and placebo (−1.09% difference, 95% CI 0.90, 1.28; p < 0.0001). There was greater weight loss with liraglutide vs placebo (treatment difference –1.39 kg, 95% CI 2.10, 0.69; p = 0.0001), and vs glargine (treatment difference −3.43 kg, 95% CI 4.00, 2.86; p < 0.0001). Liraglutide reduced systolic BP (−4.0 mmHg) vs glargine (+0.5 mmHg; −4.5 mmHg difference, 95% CI 6.8, −2.2; p = 0.0001) but not vs placebo (p = 0.0791). Rates of hypoglycaemic episodes (major, minor and symptoms only, respectively) were 0.06, 1.2 and 1.0 events/patient/year, respectively, in the liraglutide group (vs 0, 1.3, 1.8 and 0, 1.0, 0.5 with glargine and placebo, respectively). A slightly higher number of adverse events (including nausea at 14%) were reported with liraglutide, but only 9.8% of participants in the group receiving liraglutide developed anti-liraglutide antibodies. Conclusions/interpretation Liraglutide added to metformin and sulfonylurea produced significant improvement in glycaemic control and bodyweight compared with placebo and insulin glargine. The difference vs insulin glargine in HbA$_{1c}$ was within the predefined non-inferiority margin.

Inclusion criteria According to the patient baseline characteristics in LEAD-5[2], the inclusion criteria of the virtual trial emulations are set based on the normal distribution of the *demographic variable* and *pre-treatment clinical measurements* as follows:

Age [57.6, 9.5]; *BMI* [30.4, 5.3]; *Cholesterol* [4.47, 1.17]; *Creatine* [84.02, 31.45]; *Diastolic blood pressure* [80.8, 9.1]; *Systolic blood pressure* [135, 15]; *HbA$_{1c}$* [67.2, 7.5]; *UAC* [2.2, 1.1, lower=1.1, upper = 5.7]; *EGFR* [59.5, 1.0, lower=59.5, upper=60].

Two numbers in the brackets stand for the mean and standard deviation of the normal distributions of the variables. Note *UAC* and *EGFR* receive truncated normal distributions with lower and upper bounds specified above.

Effect size and drug comparisons The effect of the drugs (*GLP-1, basal insulin and placebo*) are measured specifically with clinical measurements including HbA_{1c}, systolic blood pressure and body mass index (BMI). For each drug, we compare the difference between the pre-treatment and post-treatment measurements. The trials involve three virtual patient groups, including the *GLP-1* group, *insulin* group and *placebo* group. The effect size between different drugs are compared with the difference-in-differences method [8].

Emulation Each trial group contains N = 232 patients, who are sampled from the normal distribution of the confounding variables according to the inclusion criteria and they are randomly assigned to one of the three groups. Thus, each virtual individual patient is given either *GLP-1*, *basal insulin* or *placebo* according to its group. In addition, all the patients have a history of using *metformin* and *sulfonylurea* in previous treatments. As the demographics and pre-treatment measurements are randomly sampled according to the inclusion criteria and the treatments are randomly assigned, all the virtual patients have very similar conditions before the treatments and their treatment decisions (i.e., use of the drugs) are not confounded. This prevents bias in the estimation of the drug effect and enables meaningful comparisons between the drugs.

Once the confounding and drug variables are assigned, the model calculates the post-treatment measurements of the virtual patients, yielding full records of the virtual patients in each group. This computation is stochastic since our model is generative involving the sample of noise variable $\boldsymbol{Z}$**.** This allows us to make comparisons of the treatment effects of *GLP-1*, *basal insulin* and *placebo*. To guarantee repeatability, we conduct the above trials multiple M = 60 times, and the trial results are computed as the mean and standard deviations of the multiple trials – see Table 3-1 below.

Trial outcomes Table 3-1 shows the average treatment effect (from *M=60* trials) of each of the three drugs estimated based on the difference between pre- and post-treatment measurements.

Table 3-1 LEAD-5 emulation results:
Average treatment effect measured by differences between pre- and post-treatment measurements

	HbA_{1c} [mmol/mol]		**Systolic Blood Pressure**[mmHg]		**BMI** [kg/m2]	
	Mean	Std	Mean	Std	Mean	Std
GLP-1	-5.68	0.36	-4.31	0.57	0.68	0.07
Basal insulin	-4.43	0.53	-2.04	0.51	1.46	0.09
Placebo	-3.09	0.41	-1.87	0.50	1.28	0.09

Figure 6 present visualization of the treatment effects of the three drugs (colour coded). Each visualization shows the pre-treatment measurements in *x-axis* versus post-treatment measurements in *y-axis*, together with regression lines to indicate the overall trends in the relationships. We have indicated the range of the measurements with box plot.

Table 3-2 shows pairwise comparisons of the drugs on HbA_{1c}, systolic blood pressure and BMI, which is calculated with the difference-in-differences method [8].

The figures in Table 3-1, Table 3-2 and Figure 6 show that all of the three drug groups have similar pre-treatment measurement and *GLP-1* produces the largest reduction of the clinical measurements after treatments among the three drugs; *Basal insulin* performs the second best in HbA_{1c} and systolic blood pressure measurements; *Placebo* is the second best according to the BMI measurements. This is in a good agreement with the LEAD-5 outcomes, which suggested the same drug ranking in terms of the HbA_{1c} and BMI measurements. In the systolic blood pressure measurements, LEAD-5 also suggested more reduction from *GLP-1* than *basal insulin*. However, there should be no significant difference between *GLP-1* and *placebo*, which contradicted to our experiments in this aspect.

Table 3-2 LEAD-5 emulation results:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA$_{1c}$ [mmol/mol]			**Systolic Blood Pressure** [mmHg]			**BMI** [kg/m2]		
	Expected	95% Conf Int	P-value	Expected	95% Conf	P-value	Expected	95% Conf	P-value
GLP-1 vs *placebo*	-2.58	-2.78, -2.37	<0.001	-2.38	-2.73, -2.03	<0.001	-0.61	-0.77, -0.45	0.001
GLP-1 vs *Insulin*	-1.21	-1.42, -1.0	<0.001	-2.99	-2.62, -1.89	<0.001	-0.79	-9.96, -0.63	0.001

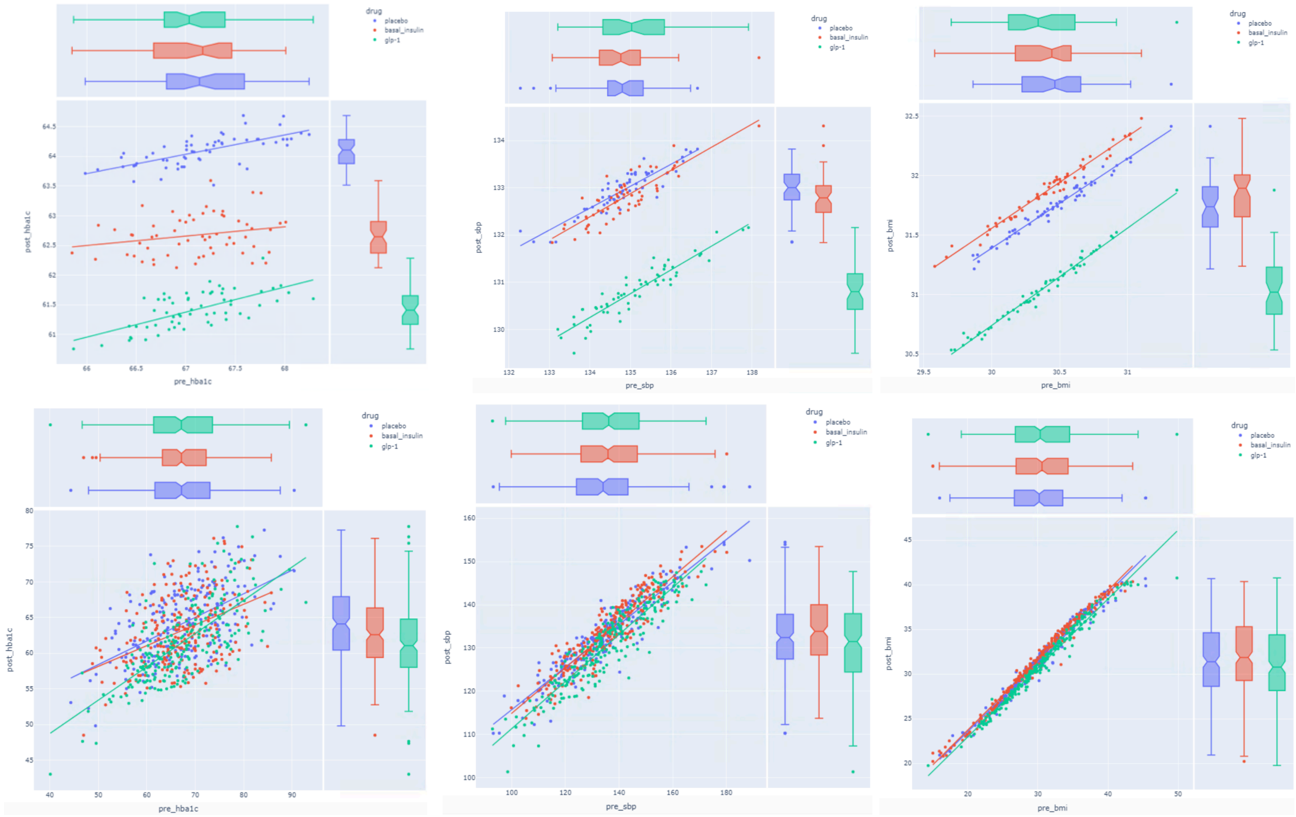


Figure 6 Outcomes of the virtual trial emulations: The performance difference between *GLP-1*, *basal insulin* and *placebo* in HbA$_{1c}$, systolic blood pressure and BMI measurements. Each drug is visualized by pre-treatment HbA$_{1c}$ (*x-axis*) vs post-treatment HbA$_{1c}$ (*y-axis*) and is colour coded. We compare the performance of these three drugs in each diagram. (**a**) Top row: overall performance comparison of *M=60* trials between the three drugs on, from left to right, HbA$_{1c}$ systolic blood pressure and BMI measurements; (**b**) Bottom row: Comparisons between the three drugs in three randomly selected trials from *M=60* trials on, from left to right, HbA$_{1c}$ systolic blood pressure and BMI measurements.

These results of the emulation show that:

- *GLP-1* is the best according to the measurements in HbA$_{1c}$, systolic blood pressure and BMI reduction in comparison with *basal insulin* and *placebo*. Through pairwise comparisons with difference-in-differences, the HbA$_{1c}$ reduction by *GLP-1* vs *placebo* is -2.58, 95% CI[-2.78, -2.37] p< 0.001 ; by *GLP-1*vs *basal insulin* is -1.21, 95% CI[-1.42, -1.0] p< 0.001. This is generally in a good agreement with the LEAD-5 outcomes, i.e., "*Liraglutide reduced HbA$_{1C}$ significantly vs glargine (1.33% vs 1.09%; −0.24% difference, 95% CI [0.08, 0.39]; p= 0.0015); Liraglutide reduced HbA$_{1C}$ significantly vs placebo (−1.09% difference, 95% CI [0.90, 1.28]; p < 0.0001)*".

- *GLP-1* also shows similar significance in the reduction of systolic pressure and BMI in the virtual trial emulations. The agreement with the outcomes from LEAD-5 is mixed in this aspect, as LEAD-5 reported " *weight loss with liraglutide vs glargine (treatment difference −3.43 kg, 95% CI [4.00, 2.86]; p < 0.0001; weight loss with liraglutide vs placebo (treatment difference –1.39 kg, 95% CI [2.10, 0.69]; p = 0.0001; Liraglutide reduced systolic BP (−4.0 mmHg) vs glargine (+0.5 mmHg); −4.5 mmHg difference, 95% CI [6.8, −2.2]; p = 0.0001); but not Liraglutide vs placebo (p = 0.0791)*". The ranking of the three drugs on the BMI measurement agrees with LEAD-5. The discrepancy lies within the systolic blood pressure measurement. Our simulations suggest *GLP-1* vs *placebo* reduction -2.38, 95% CI[-2.72,-2.03], p<0.001 and *GLP-1*vs *basal insulin* reduction -2.99, 95% CI[-2.62, -1.89], p<0.001. However, LEAD-5 only reported reduction significance between *GLP-1*vs *basal insulin*.

3.2 **LEAD-5 counterfactual emulations**

This experiment is conducted as a complement to the trial emulation described in Section 3.1. The trial emulation in Section 3.1 has emulated a real trial scenario in which patients who meet the inclusion criteria are recruited and randomly assigned to a drug-group (either *GLP-1*, *basal insulin* or *placebo*). Thus in that case we have three groups with different patient cohorts (but with similar conditions in demographics, pre-treatment measures and pre-treatment medical history to allow meaningful comparisons between them for the effect of the drugs). In contrast, the counterfactual trial emulations allow us to examine the clinical questions about the drug effects in a counterfactual scenario by computing the post-treatment measures as the effects of a set of hypothetical treatment with different drugs. This allows us to answer clinical questions such as "*What would be the clinical outcomes if the patient had been given a different treatment?*". The experiment still follows the same inclusion criteria as LEAD-5 (with the same normal distribution). All the other settings, including the group size (*N=232*) also remain the same. The only difference here is that instead of creating three groups of virtual patients with very similar conditions, we only use one patient group but giving them three different drugs (*GLP-1*, *basal insulin* or *placebo)* in each trial, hence measure difference drug effects of the same patients.

Table 4-1 shows the average treatment effect (from the *M=60* trials) of each of the three drugs estimated with the counterfactual emulations.

Table 4-1 Counterfactual emulation results:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA$_{1c}$ [mmol/mol]		**Systolic Blood Pressure**[mmHg]		**BMI** [kg/m2]	
	Mean	Std	Mean	Std	Mean	Std
GLP1	-5.72	0.40	-4.21	0.49	0.65	0.09
Basal insulin	-4.48	0.48	-2.14	0.50	1.45	0.09
Placebo	-3.06	-0.44	-1.96	0.49	1.27	0.09

Table 4-2 shows pairwise difference-in-differences of the drug effects from the counterfactual emulations.

Table 4-2 Counterfactual emulation results:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA$_{1c}$ [mmol/mol]			**Systolic Blood Pressure**[mmHg]			**BMI** [kg/m2]		
	Expected	Conf Int	P-value	Expected	Conf	P-value	Expected	Conf	P-value
GLP1 vs placebo	-2.65	-2.85, -2.45	<0.001	-2.25	-2.65, -1.85	<0.001	-0.61	-0.77, -0.46	<0.001

GLP1 vs insulin	-1.24	-1.44, -1.03	<0.001	-2.07	-2.47, -1.66	<0.001	-0.80	-0.96, -0.64	<0.001

In fact, the counterfactual emulations of LEAD-5 provide very similar results to the emulations presented in the Section 3.1. As reported in Table 4-1 and Table 4-2, *GLP-1* is the best performed drug as it shows significant reduction in HbA$_{1c}$, systolic blood pressure and BMI in comparison with both *placebo* and *basal insulin*. This again has reproduced the ranking of the three drugs in LEAD-5, except the discrepancy on systolic blood pressure.

Counterfactual emulation is one of the main advantages that we have in our emulation approach, which is based on the learning of the structural equations in the causal model. With the counterfactual ability, we can view consequences of treating the same patient under multiple clinical scenarios to support clinical decision making. We do not need to match individual patients between different groups as conventionally required in the observational studies. The next section further extends the counterfactual emulation to the data from real patients.

3.3 **Extended counterfactual emulation on real patients**

In this experiment, we demonstrate the use of trial emulation to seek answers to "counterfactual" clinical questions in clinical practice. Specifically, we emulate hypothetical and counterfactual treatments where different drugs are applied to the same real patients. These virtual trial emulations are designed to find out how different the clinical outcomes would be if the patients had taken different treatment pathways. In our experiments we have identified patients in the SCI diabetes dataset according to the drugs they took in their treatments. The patients have been placed in three groups according to the drugs they have had, namely *GLP-1*, *basal insulin* and *placebo*. Then for each patient, the virtual trial emulation administers three drugs, one is the real drug that the patient took in reality, and the other two drugs are counterfactual. The emulations then calculate the average treatment effect in each drug group, and we have also estimated pairwise difference-in-differences in each drug group, namely *GLP-1* vs *placebo* and *GLP-1* vs *basal insulin*. The results of the counterfactual simulations are presented in Table 5-1(average treatment effect) and Table 5-2 (pairwise difference-in-differences) below.

Table 5-1a Counterfactual emulation on *GLP-1* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA$_{1c}$ [mmol/mol]		**Systolic Blood Pressure**[mmHg]		**BMI** [kg/m2]	
	Mean	Std	Mean	Std	Mean	Std
GLP-1	**-13.26**	0.53	**-2.88**	0.45	**0.34**	0.06
Basal insulin	-13.90	0.61	-0.16	0.46	1.14	0.06
Placebo	-12.93	-0.59	-0.43	0.47	1.00	0.06
Real GLP-1	**-10.61**	0.95	**-2.47**	0.79	**-1.07**	0.10

Table 5-2a Counterfactual emulation on *GLP-1* drug group:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA$_{1c}$ [mmol/mol]			**Systolic Blood Pressure**[mmHg]			**BMI** [kg/m2]		
	Expected	Conf Int	P-value	Expected	Conf	P-value	Expected	Conf	P-value
GLP-1 vs *placebo*	-0.32	-0.62, -0.02	0.03	-2.44	-2.72, -2.17	<0.001	-0.66	-0.76, -0.55	<0.001
GLP-1 vs *insulin*	0.64	0.34, 0.94	<0.001	-2.72	-2.99, -2.44	<0.001	-0.80	-0.90, -0.70	<0.001

Table 5-1a and Table 5-2a are the results for the *GLP-1* group. This is the group in which the patients took *GLP-1* in reality. We add another row in Table 3-1a to show the average treatment effects of *GLP-1* calculated using the real patient data, and we highlight the counterfactual and real results that are comparable. We can see that the emulations have produced the right direction (i.e., positive vs negative) for the treatment effect except on the BMI results. On HbA$_{1c}$ performance, the ranking is *basal insulin>GLP-1>placebo*, which does not agree with LEAD-5, however, the pre-treatment HbA$_{1c}$ in this group are outside the range of LEAD-5 inclusion criteria. On systolic blood pressure, *GLP-1>placebo>basal insulin*, which partially agrees with LEAD-5 (i.e. *GLP-1=placebo>basal insulin*) except on *placebo*, note the pre-treatment systolic blood pressure within this group is within the LEAD-5 inclusion range. On BMI, *GLP-1>placebo>basal insulin*, which agrees well with LEAD-5, and the pre-treatment measurement on BMI are also within the LEAD-5 inclusion range.

Table 5-1b Counterfactual emulation on *basial insulin* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA$_{1c}$ [mmol/mol]		**Systolic Blood Pressure**[mmHg]		**BMI** [kg/m2]	
	Mean	Std	Mean	Std	Mean	Std
GLP-1	-17.23	0.47	-2.22	0.41	0.97	0.04
Basal insulin	**-18.33**	0.51	**0.78**	0.43	**1.81**	0.05
Placebo	-17.33	0.45	0.42	0.41	1.62	0.05
Real insulin	**-13.21**	0.75	**0.66**	0.74	**0.59**	0.09

Table 5-2b Counterfactual emulation on *basal insulin* drug group:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA$_{1c}$ [mmol/mol]			**Systolic Blood Pressure**[mmHg]			**BMI** [kg/m2]		
	Expected	Conf Int	P-value	Expected	Conf	P-value	Expected	Conf	P-value
GLP-1 vs placebo	0.1	-0.17, 0.37	0.472	-2.64	-2.94, -2.34	<0.001	-0.65	-0.74, -0.56	<0.001
GLP-1 vs insulin	1.1	0.83, 1.36	<0.001	-3.00	-3.3, -2.7	<0.001	-0.833	-0.93, -0.74	<0.001

Table 5-1b and Table 5-2b are the results for the *basal insulin* group. This is the group in which the patients took *basal insulin* in reality. We add another row in Table 3-1b to show the average treatment effects of *basal insulin* calculated using the real patient data, and we highlight the counterfactual and real results that are comparable. We can see that the emulations have produced the right direction (i.e., positive vs negative) for the treatment effect. On HbA$_{1c}$ performance, the ranking is *basal insulin>GLP-1>placebo*, which does not agree with LEAD-5, however, the pre-treatment HbA$_{1c}$ in this group are outside the range of LEAD-5 inclusion criteria. On systolic blood pressure, *GLP-1>placebo>basal insulin*, which partially agrees with LEAD-5 (i.e. *GLP-1=placebo>basal insulin*) except on *placebo*, note the pre-treatment systolic blood pressure within this group is within the LEAD-5 inclusion range. On BMI, *GLP-1>placebo>basal insulin*, which agrees well with LEAD-5, and the pre-treatment measurement on BMI are also within the LEAD-5 inclusion range.

Table 5-1c Counterfactual emulation on *placebo* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA$_{1c}$ [mmol/mol]		**Systolic Blood Pressure**[mmHg]		**BMI** [kg/m2]	
	Mean	Std	Mean	Std	Mean	Std
GLP-1	-5.24	0.37	-3.84	0.36	0.61	0.04
Basal insulin	-2.57	0.47	-0.98	0.37	1.42	0.05

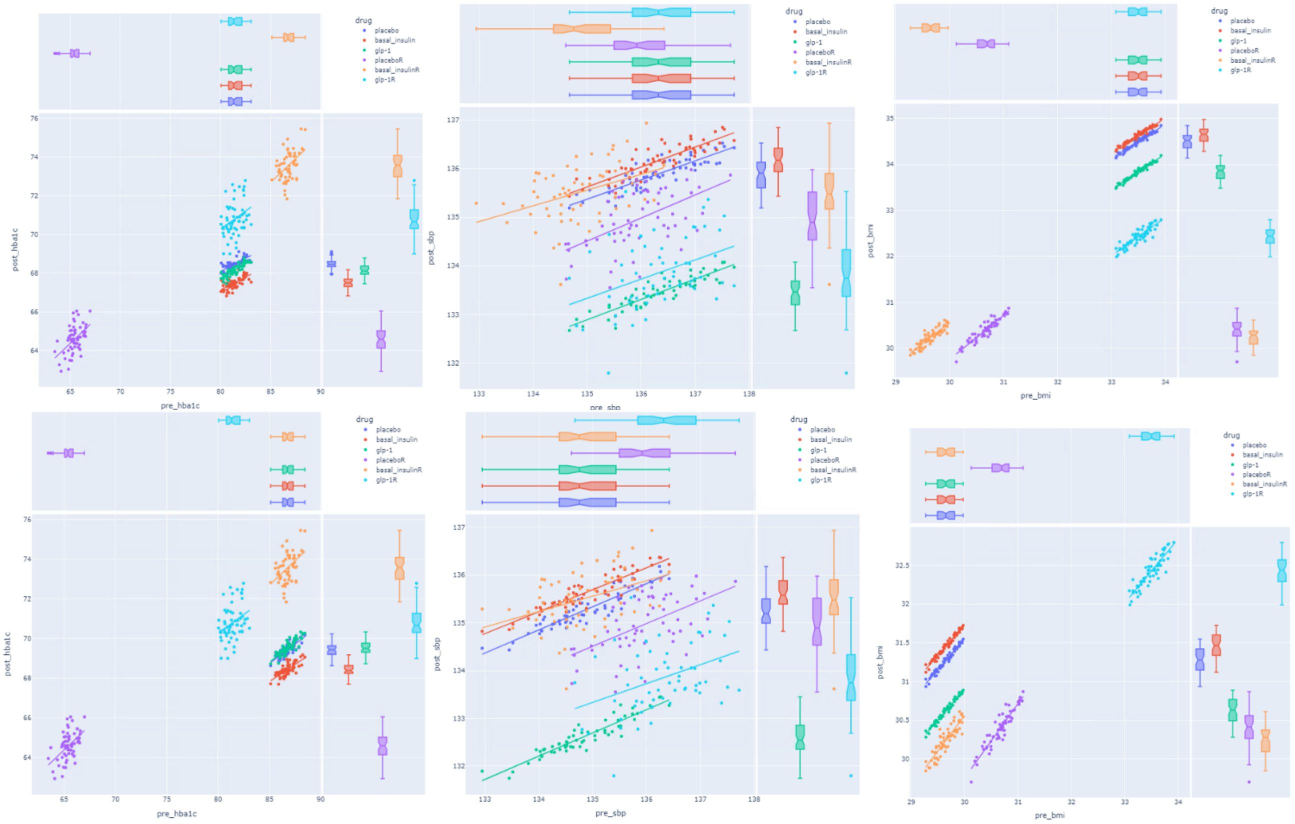
Placebo	**-1.99**	0.42	**-1.62**	0.36	**1.24**	0.04
Real placebo	**-0.83**	0.71	**-0.99**	0.62	**-0.26**	0.08

Table 5-2c Counterfactual emulation on *placebo* drug group::
Pairwise comparisons of treatment effect with difference-in-differences

	HbA$_{1c}$ [mmol/mol]			**Systolic Blood Pressure**[mmHg]			**BMI** [kg/m2]		
	Expected	Conf Int	P-value	Expected	Conf	P-value	Expected	Conf	P-value
GLP-1 vs *placebo*	-3.25	-3.5, -3	<0.001	-2.22	-2.48, -1.96	<0.001	-0.64	-0.76, -0.53	<0.001
GLP-1 vs *insulin*	-2.67	-2.92, -2.42	<0.001	-2.86	-3.12, -2.59	<0.001	-0.82	-0.93, -0.70	<0.001

Table 5-1c and Table 5-2c are the results for the *placebo* group. This is the group in which the patients took *placebo* in reality. We add another row in Table 3-1c to show the average treatment effects of *placebo* calculated using the real patient data, and we highlight the counterfactual and real results that are comparable. We can see that the emulations have produced the right direction (i.e. positive vs negative) for the treatment effect except on BMI. On HbA$_{1c}$ performance, the ranking is *GLP-1>basal insulin >placebo*, which agrees with LEAD-5, and the pre-treatment HbA$_{1c}$ in this group are within the range of LEAD-5 inclusion criteria. On systolic blood pressure, *GLP-1>placebo>basal insulin*, which partially agrees with LEAD-5 (i.e. *GLP-1=placebo>basal insulin*) except on *placebo*, note the pre-treatment systolic blood pressure in this group is within the LEAD-5 inclusion range. On BMI, *GLP-1>placebo>basal insulin*, which agrees well with LEAD-5, and the pre-treatment measurement on BMI are also within the LEAD-5 inclusion range.

Figure 7 visualise the results of the counterfactual emulations on the three groups.



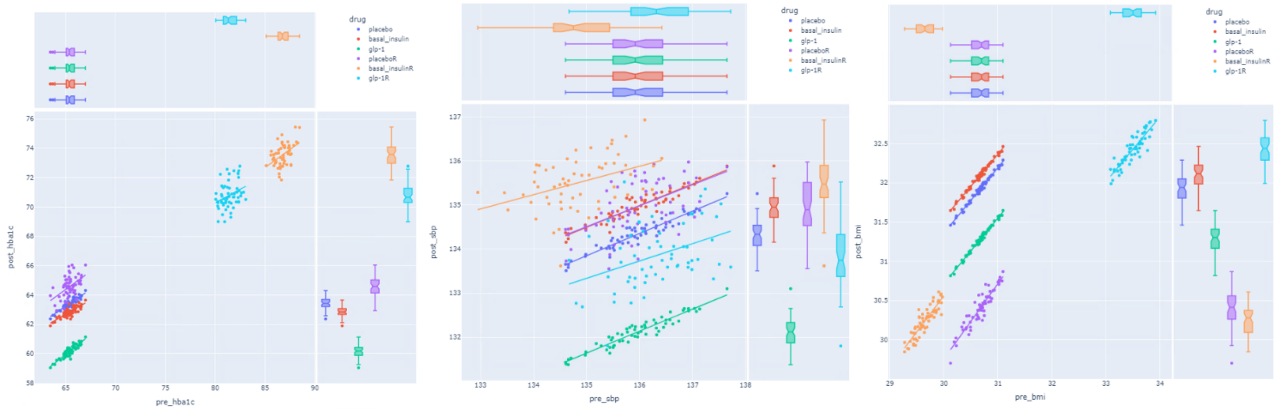


Figure 7 Visualisation of the counterfactual emulation results of the *GLP-1*, *basal insulin* and *placebo* groups. (a) Top row: *GLP-1* group results measured by (from left to right) HbA_{1c}, systolic blood pressure and BMI; (b) Middle row: *basal insulin* group results measured by (from left to right) HbA_{1c}, systolic blood pressure and BMI; (c)Bottom row: *placebo* group results measured by (from left to right) HbA_{1c}, systolic blood pressure and BMI;

The results of the counterfactual emulations on real patients in Table 5-1, Table 5-2 show that *GLP-1* remains as the drug with best performance over the *placebo* group of real patients. In fact, this result agrees well with LEAD-5 in terms of the ranking of the three drugs on HbA_{1c} and BMI measurements. Our analysis shows that patients in the *placebo* group are very close to the LEAD-5 patient baseline characteristics in terms of the HbA_{1c} measurements. However, the results of the counterfactual emulations on the other two groups are very different. The ranking of the three drugs are different from the LEAD-5 results. Our analysis shows that patients in these two groups are very different from the LEAD-5 patient baseline characteristics – for example, their pre-treatment measurements on HbA_{1c} are not within the range of the LEAD-5 criteria. Remarkably, according to the emulations, *GLP-1* is not the best drug for the *GLP-1* group patients, instead, *basal insulin* appear to be the right medicine for the patients in

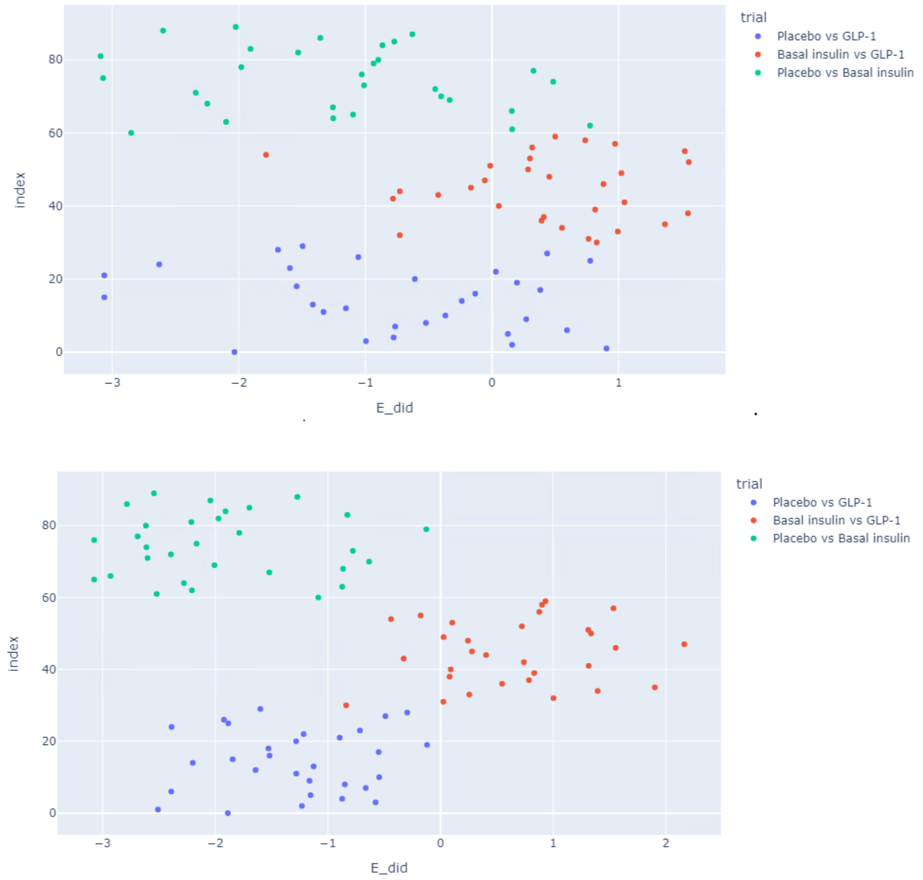


Figure 8 Comparisons of difference-in-differences results from three pairs of drugs in two trial emulations on real patient data. (**a**) Top row: patient data are sampled according to the mean and standard deviation of the real dataset; (**b**) Bottom row: patient data sampled from the real dataset with narrowed (0.5x) standard deviation.

the *basal insulin* group. Another observation towards the emulations vs the real data is that the emulations often slightly over-predict the average treatment outcomes on all of the three drugs.

Further experiments on trial emulations with the real patient data (SCI diabetes) suggest that the LEAD-5 results cannot be simply extrapolated to cover patients outside its baseline characteristics. When we directly apply the drugs to the patients that are randomly sampled from the datasets, the treatment outcomes of *GLP-1* are not aligned with the results in LEAD-5, which suggests that *GLP-1* is probably not the most suitable drug to all patients. However, if we sample from the dataset with a much narrowed (0.5x) standard deviation, the *GLP-1* performance become much more aligned with LEAD-5 – see Figure 8. Our analysis shows that the narrowed data distribution is much closer to the LEAD-5 patient baseline characteristics.

4. Analysis and conclusion

We have mainly conducted two types of experiments on the virtual trial emulations in the context of T2DM treatment with three different drugs. The first type of experiments is focused on the replication of the existing LEAD-5 trial, and the second type of experiments attempts to emulate counterfactual scenarios where different drugs are applied to the same patients to support clinical decision making. The effect sizes are estimated with both average treatment effects (i.e., difference between pre-and post-treatment measurements) and difference-in-differences between pairwise drugs. When the patients meet the LEAD-5 patient baseline characteristics, the trial emulations produce the same ranking between the three drugs as LEAD-5. Our experiments are conducted based on independent sampling of virtual patients for the three treatment groups, counterfactual emulations on the same group of virtual patients, and counterfactual emulations on the real patients. All the results have suggested that *GLP-1* has the best performance in terms of HbA_{1c}, systolic blood pressure and BMI reduction if the patients meet the inclusion criteria of LEAD-5. However, the experiments with real patients who do not fall into the baseline characteristics of LEAD-5 have presented different performance rankings between the drugs. This suggests that LEAD-5 trial outcomes cannot be simply extrapolated to cover other patient populations. To this end, the virtual trial emulation models and tools are potentially very useful in terms of providing evidence to support the extrapolation of clinical trials for real-world clinical practice.

Overall, the trial emulations have replicated LEAD-5 very well on the HbA_{1c} and BMI measurements, which are the most important clinical measures in T2DM. *GLP-1* vs *basal insulin* on systolic blood pressure also well agrees with LEAD-5. There is a discrepancy on the effect of *GLP-1* vs *placebo* on systolic blood pressure, where LEAD-5 shows no significance while our emulation still predicts significant reduction. In addition, our experiments with the trial emulations on real patients have compared the results with the real data (where the results are comparable). Most of the emulation have produced the right direction for the effect. However, we have also noticed over estimation in some of the cases.

We recon the difference between the real trial and the emulations potentially comes from the following sources:

- The data presents the drug as *GLP-1* , which is more general than the specific drug (i.e. Liraglutide) tested in LEAD-5. Hence the training of the emulation model and the testing with the trial emulations have addressed aggregated/mean effect of several *GLP-1* drugs as a drug category. Also, both *basal insulin* and *placebo* are involved as drug categories in the emulations, which are not identical to the corresponding drugs involved in the original trials in LEAD-5.
- Although the trial emulations have tried to create the LEAD-5 trial populations by sampling from the normal distributions of the LEAD-5 patient baseline characteristics, it is not possible to replicate a completely identical cohort due to different settings between the trial and the real-world clinical practice. For example, several variables involved in the virtual trial emulation are not identical to the variables that are used in the LEAD-5 patient baseline characteristics, for example, we have only used *age* to describe the patient demographics; several clinical measurements such as blood pressures are based on yearly mean measurements, and so on and so forth. Another limitation within the real-world dataset is the pre-treatment history of the patients might not be complete.
- Other difference in the settings between LEAD-5 and the real-world clinical practice may have also contributed to the discrepancy between the emulations and real trial results. Hidden confounders can be responsible for generating bias in the emulations. The current model does not take these into considerations.

- Practical errors in the technologies (e.g., training and computational errors) may also contribute to the difference at certain degrees.

Limitations: Due to time constraint, this study has its limitations in a number of aspects in both the technology and clinical dimensions. These include:

- As a proof-of-concept study, it only has replicated a single clinical trial in T2DM (i.e.LEAD-5). To further validate the concept, more evidences are needed from further studies on a wider range of clinical trials.
- The study has evaluated the emulation with a real trial that we already know the outcomes. For the future use of the trial emulations to address real-world clinical questions where the ground truth is unknown, we need standardised quality metrics to calibrate the emulation outcomes.
- The current model has made several assumptions about the causal structures with in the data in order to recover them. These include (**a**) the model identifiably assumption, where we assume the data was generated from a model in additive noise form. This is quite a big assumption as the model underlying the real world data may be different. The true model might not be in an identifiable form, in which case we cannot completely identify the causal structure from the data. Our current work has not taken into account the uncertainty from unidentifiable models; (**b**) the causal sufficiency assumption, where we assume there is no unobserved confounders. This is not realistic as real-world measurement is always limited and there is a possibility of influence from unobserved variables.
- We have made several simplifications in data pre-processing. For example, only *age* variable is used to represent the patient demographics; the pre-treatment and post-treatment measurements are taken as yearly average/median values – all these might have contributed to errors in the study. Also, we have simply dropped all the missing data in the training.
- The differential privacy framework has not worked compatibly with the causality learning framework in the current model. Although we have used similarity metrics to show that each synthetic patient is different from the real patients in the training data, further investigation is worthwhile to explore the differential privacy framework further. One potential solution is to involve a recent model in differentiable DAG sampling[10], where we can offer more protection to the training data from the data generative process.

References

1. NHS Greater Glasgow and Clyde & NHS Tayside Health Board area data https://www.nhsggc.scot/staff-recruitment/staff-resources/research-and-innovation/nhsggc-safe-haven/
2. Russell-Jones, D., Vaag, A., Schmitz, O., Sethi, B K, Lalic, N, Antic, S., Zdravkovic, M., Ravn, G M, Simó, R; Rigato,M.,Fadini,G.P.(2014), Liraglutide vs insulin glargine and placebo in combination with metformin and sulfonylurea therapy in type 2 diabetes mellitus (LEAD-5 met+SU): a randomised controlled trial, *Randomized Controlled Trial, Diabetologia* 2009 Oct;52(10):2046-55.
3. Petkov, H. Hanley, C., Dong, F., Causality Learning with Wasserstein Generative Adversarial Networks, International *Journal of Artificial Intelligence and Applications (IJAIA*),13(3), 2022
4. 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* .
5. Yu, Y., Chen, J., Gao, T., and Yu, M., DAG-GNN: Dag structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098,* 2019.
6. Zheng, X., Dan, C., , Aragam, B., Ravikumar, P., and Xing, E.P., Learning Sparse Nonparametric DAGs, *AISTATS* 2020
7. Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S.,. Gradient-Based Neural DAG Learning. *arXiv preprint arXiv:1906.02226*, 2019.
8. Neal, B. (2020), Introduction to Causal Inference from a Machine Learning Perspective, https://www.bradyneal.com/causal-inference-course
9. Lin, Z., Khetan, A., Fanti, G., and Oh, S.: PacGAN: the power of two samples in generative adversarial networks. *NIPS* 2018.
10. Bertrand Charpentier, Simon Kibler, Stephan Günnemann, Differentiable DAG sampling, *International Conference on Learning Representations* (ICLR), 2022.

Appendix

We provide supplementary material to support the discussions in the main text: Section A includes the list of hyper-parameters; Section B includes additional results.

A. Hyper-parameters

Table A: List of model and optimisation hyper-parameters. **Note**: the (+1) in the generator network indicates the concatenation of the 371dim 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 37 synthetic variables (hence x37).

Hyper-parameter	**Architecture**	
	Generator	**Discriminator**
Learning rate	3e-4	3e-4
Batch size	200	200
First moment ($\boldsymbol{\beta_1}$)	0.9	0.5
Second moment ($\boldsymbol{\beta_2}$)	0.999	0.9
Dropout (all layers)	0.0	0.5
Weight decay	1e-6	1e-6
PAC	n/a	10
$\boldsymbol{c_A}$: Coefficient for absolute value of h(A)	1.0	n/a
$\boldsymbol{\lambda_A}$: DAG constraint for h(A)	0.0	n/a
Latent dim (z)	1	n/a
Network structure (MLP)	FC1: 37-1369 FC2: 37(+1) -1 (x37)	370-256-256-10
Updates per mini batch	1	1
Epochs	300	300

Learning rate scheduler. We also scheduled learning rate decay using cosine annealing with warm restarts, which decayed from LR=3e-4 to 7e-6 over 300 epochs and performed restarts at the end of this period.

B. Additional results

This section is structured as follows: section B.1 performs a sensitivity analysis on the pre-measurements, where we fix all features and perturb each one in turn to observe the change in the post-measurements; section B.2 contains the results to the model when we turn the edges into pre-features back on; section B.3 includes the confounding features in the regression task, to supplement the random forest results in the main text; section B.4 contains the results when we use differential privacy with the No-Tears model.

B.1. Sensitivity analysis of the pre-measurements

The ability of MRC-GAN to perform simulations on the learned causal structure allows us to ask counterfactual questions, such as: *'What would a real patient's outcome most likely be if they were given a different drug?',* which we explored in the main text (Sec. 3.2).

In this section, we pursue a different but related question: *'Given a patient on a particular drug, how do their post-features change when we permute their pre-measurements?'*. This is to study the extent to which the model accounts for changes in the confounders under the same trial conditions. For example: *'If the patient with a particular set of pre-features were 20 years older, how effective would each of the drugs be?'*, or: '*If the patient's blood pressure was initially much higher, how effective would the drug be?'*. Since we do not have the ground truth information for such cases, we rely upon the opinions of our clinical experts.

Experimental setup. We selected two real patients from the training dataset, and for each patient, we fixed all input pre-measurements and predicted their post-measurements with MRC-GAN in response to variations in their age. We then repeated this process for each pre-measurement, holding the remaining inputs fixed at their baseline values (Table B.1) and for each of the LEAD-5 drug classes: placebo, basal insulin. GLP-1. This enabled us to study how the effectiveness of each drug changes as we vary our patient's clinical data, and by extension, which drug is preferable under different conditions.

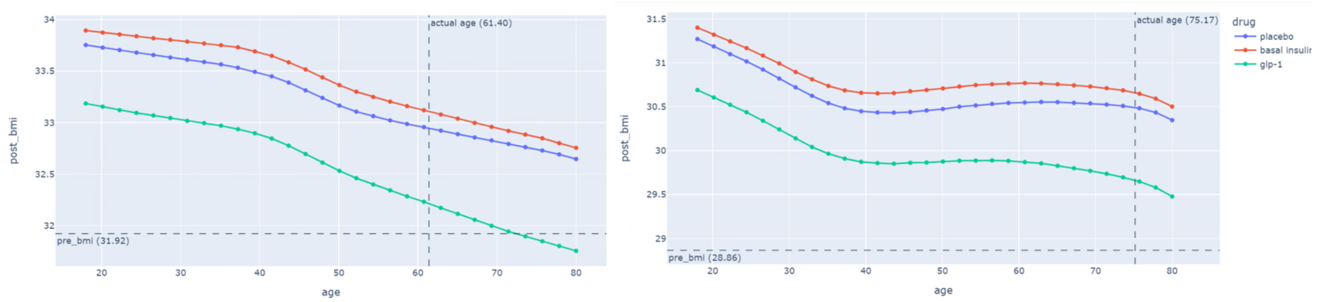
Table B.1: Baseline characteristics of two patients randomly selected from the SCI dataset for the sensitivity analysis. Clinical features are the pre-measurements (i.e., collected before treatment).

Patient		Age	BMI	Cholesterol	Creatinine	DBP	SBP	UAC	HbA1c	EGFR
1		61.40	31.92	4.6	58.0	77.0	135.0	13.2	63.0	60.0
2	75.16	28.86	3.8	81.0	56.0	106.0	5.8	103.0	60.0	

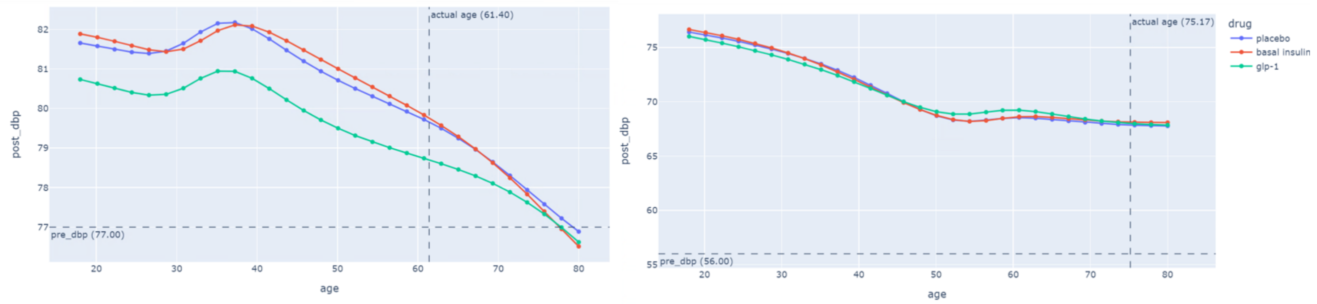
Results. The results for variations to the input age, body mass index (BMI), systolic blood pressure (SBP), and HbA1c in patients 1 and 2 are illustrated in Figures B.1 to B.4, respectively.

<u>Age</u>

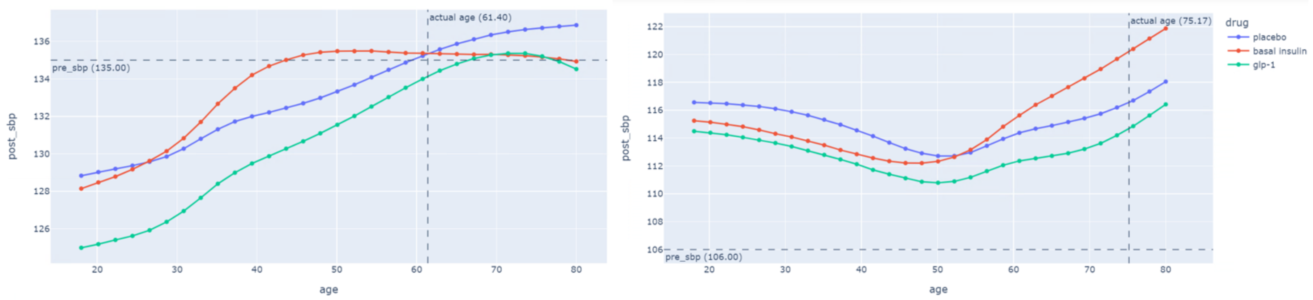
A.



B.



C.



D.

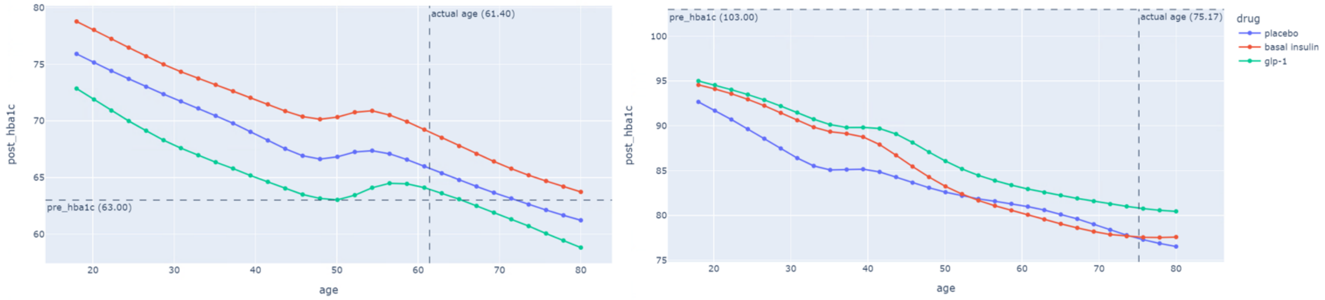
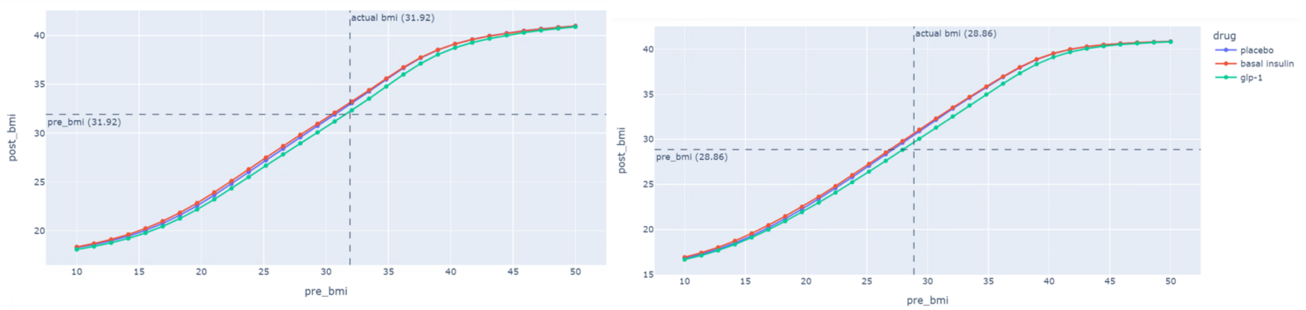


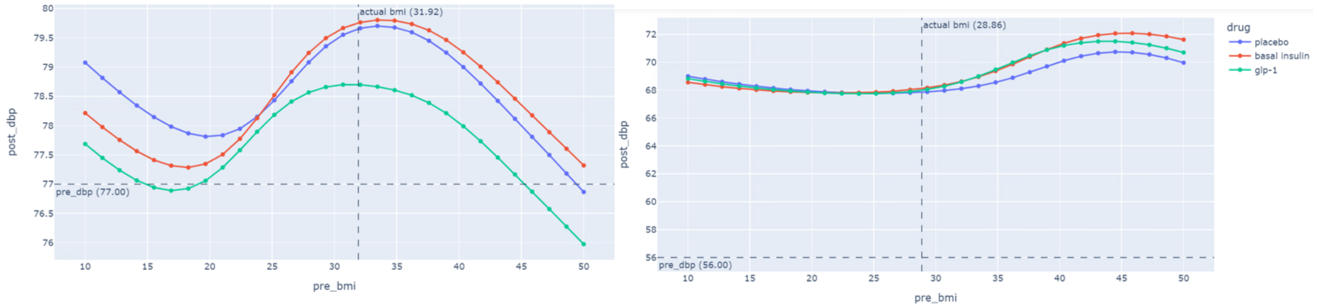
Figure B.1. Sensitivity analysis on the initial *age* for patient 1 (left) and patient 2 (right). Post-treatment outcomes predicted by MRC-GAN: **A.** BMI [kg/m^2]. **B.** DBP [mmHg]. **C.** SBP [mmHg]. **D.** HbA1c [mmol/mol]. Baseline age for each patient is shown by the dashed vertical line, and initial (pre-measurement) value for each feature shown by the dashed horizontal line.

<u>BMI</u>

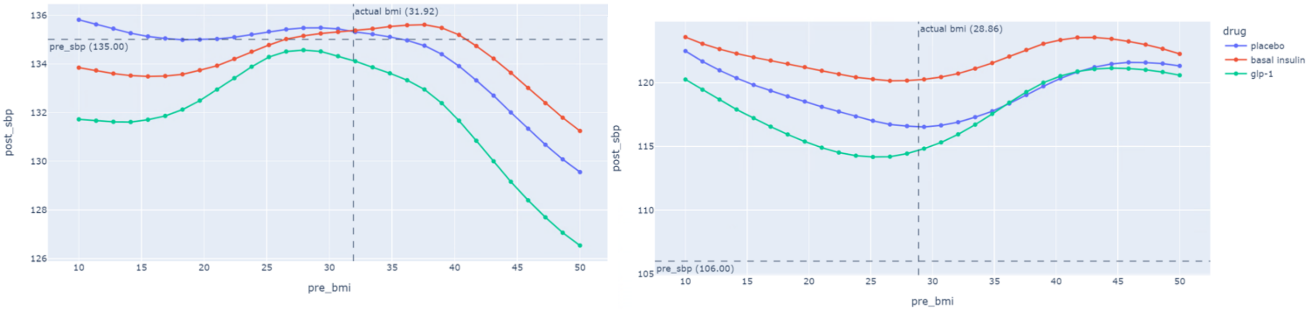
A.



B.



C.



D.

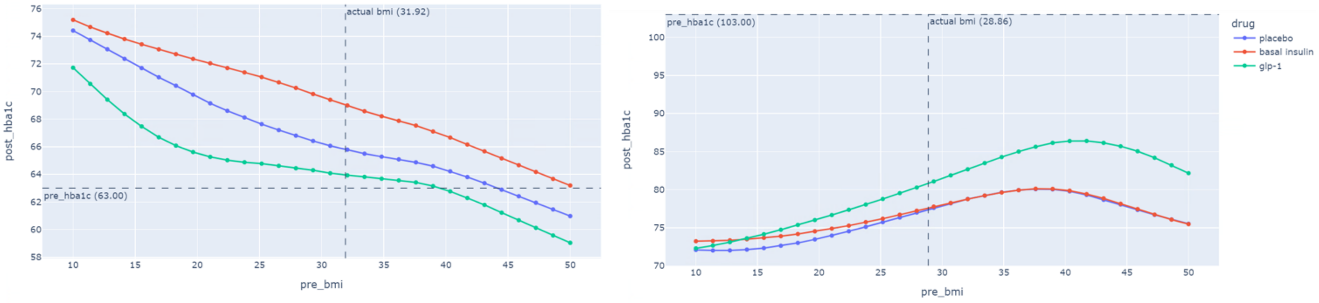
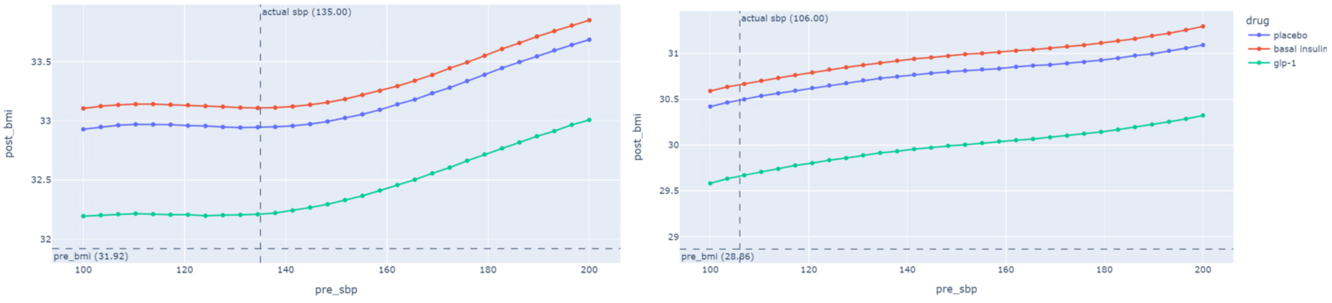


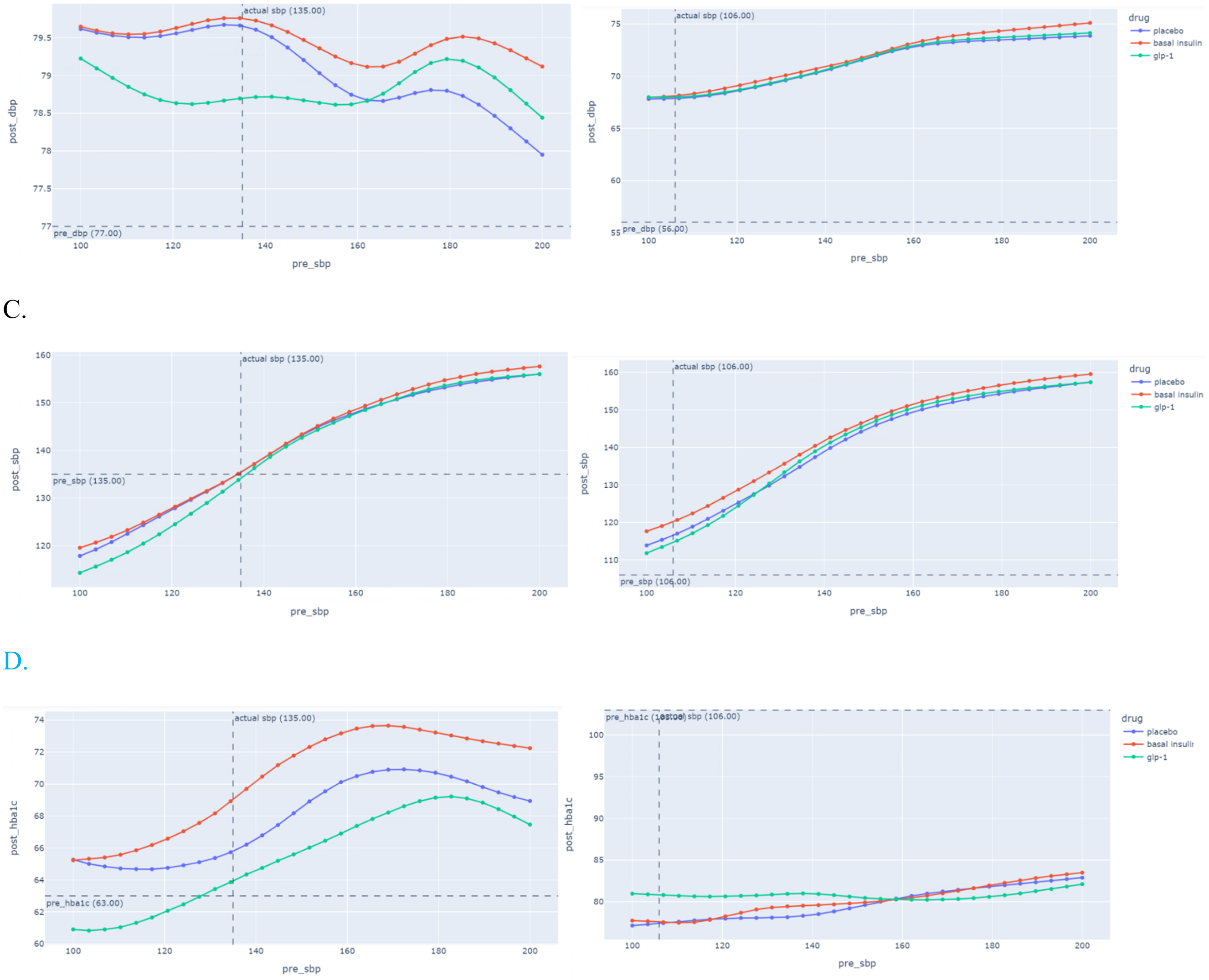
Figure B.2. Sensitivity analysis on the initial *body mass index* (*BMI)*. Post-treatment outcomes predicted by MRC-GAN: **A.** BMI [kg/m^2]. **B.** DBP [mmHg]. **C.** SBP [mmHg]. **D.** HbA1c [mmol/mol].

Systolic blood pressure

A.



B.



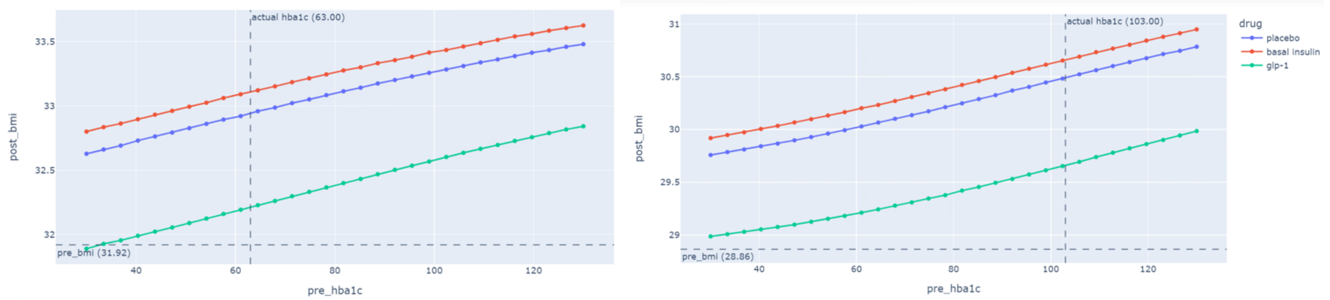
C.

D.

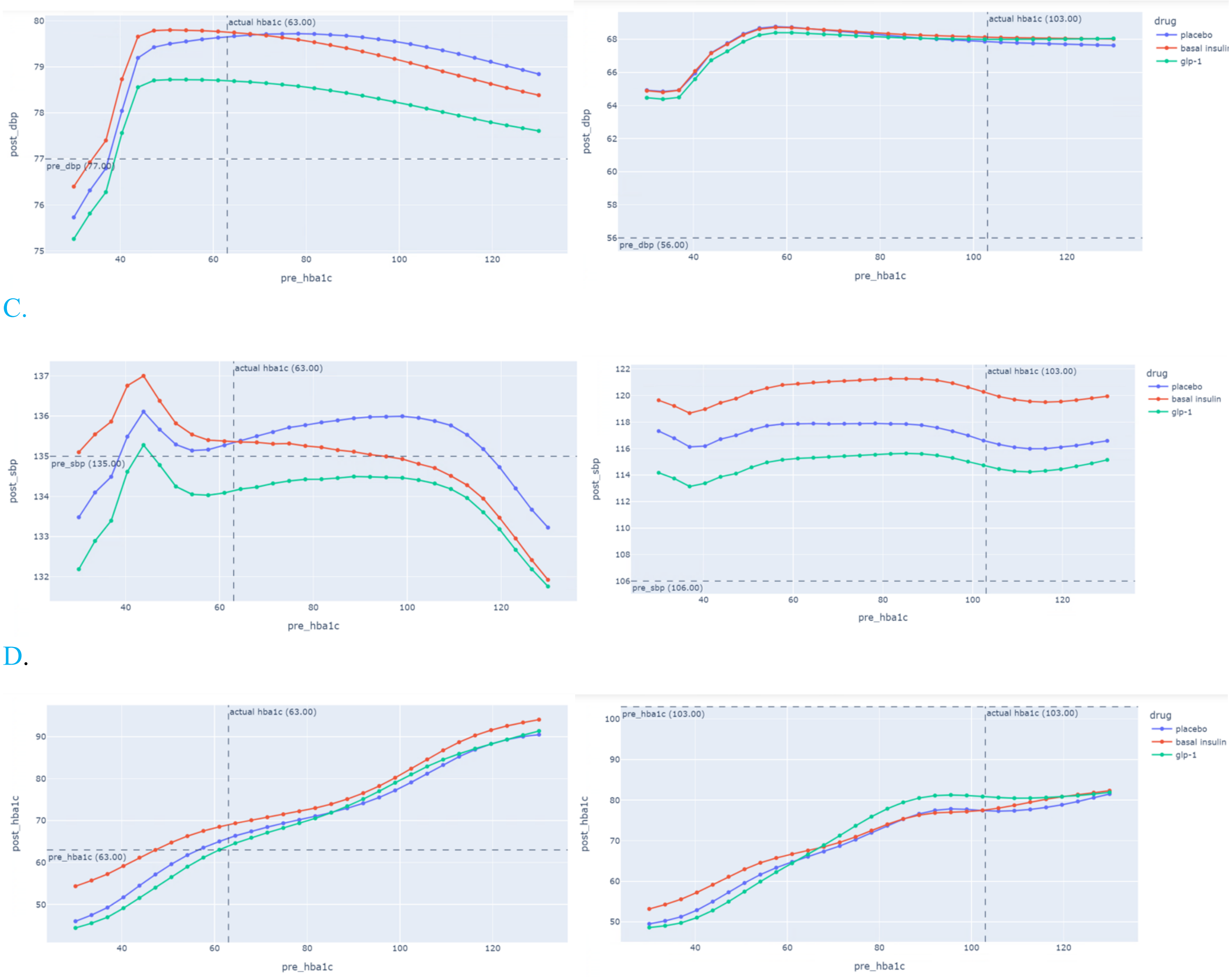
Figure B.3. Sensitivity analysis on the initial *systolic blood pressure (SBP)*. Post-treatment outcomes predicted by MRC-GAN: **A.** BMI [kg/m^2]. **B.** DBP [mmHg]. **C.** SBP [mmHg]. **D.** HbA1c [mmol/mol].

<u>HbA1c</u>

A.



B.



C.

D.

Figure B.4. Sensitivity analysis on the initial *HbA1c*. Post-treatment outcomes predicted by MRC-GAN: **A.** BMI [kg/m^2]. **B.** DBP [mmHg]. **C.** SBP [mmHg]. **D.** HbA1c [mmol/mol].

B.2. Memorization-informed distribution distance

We have used a modification from the MiFID[1] distance to measure the similarity between synthetic and real patient populations used for training. The new definition is called *MiMMD* (memorisation-informed MMD), which is based on the use of Maximum Mean Discrepancy (MMD)[2] to measure the distribution distance between the real and synthetic data distributions. In addition to MMD, we take the sample memorization into account. The memorisation distance is defined as the minimum distance of a synthetic patient with the most similar individual in the real patient population. We have used cosine similarity to identify the most similar real patient and Euclidean distance to compute their difference. MiMMD is then calculated by dividing MMD by the memorisation distance

[1] 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,

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

(to penalise too similar individual samples). The smaller MiMMD is, the better the synthetic populations are. Table B.2 shows the similarity of 6 synthetic sample cohorts towards the real patient population.

Table B.2 MiMMD similarities between a real patient cohort and 6 synthetic sample cohorts

	Memorisation Distance	**MMD**	**MiMMD**
Sample set 1	2.145	0.0694	0.0324
Sample set 2	2.093	0.0726	0.0347
Sample set 3	2.165	0.0669	0.0309
Sample set 4	2.181	0.0702	0.0322
Sample set 5	2.186	0.0677	0.0310
Sample set 6	2.191	0.0656	0.0300

B.3. Machine learning regression analysis with the confounding features

Using the same experimental setup as in Section 2.2.3, we include the confounding features in the prediction of the post-features.

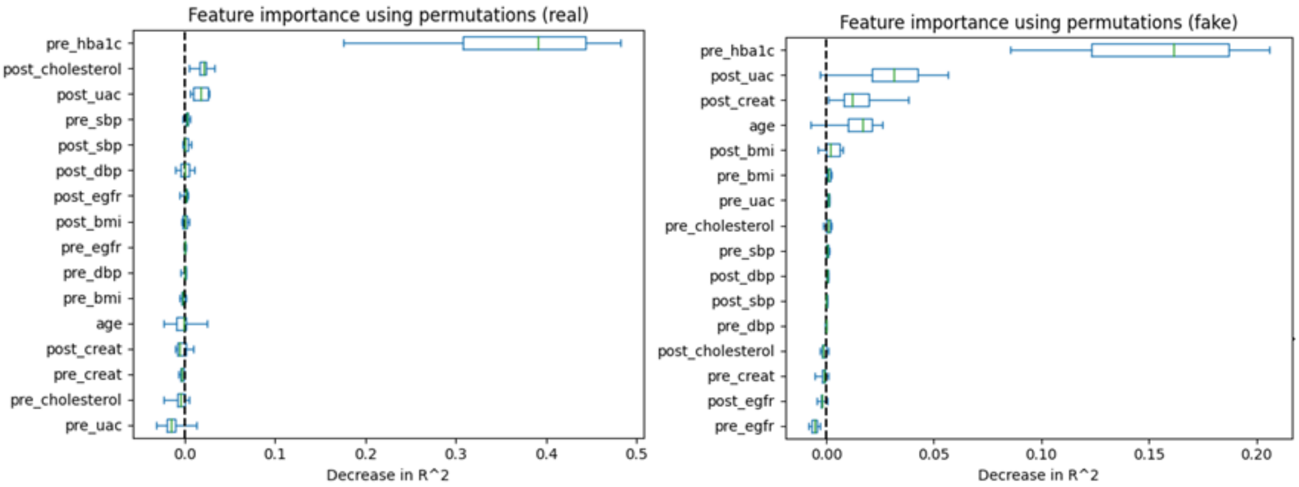
Regression performance. We observe that, unsurprisingly, the performance of both real and synthetic RF models improves with the inclusion of the confounding features (i.e., age and pre-measurements). In particular, the synthetic RF model closely matches the real RF model in the prediction of BMI, where the fake predictions correlate strongly with the correct targets ($R^2 = 0.74$, Table B.2).

Table B.3: Regression performance of the real and synthetic RF models on a test set of real data. Outcomes align with those studied in the LEAD-5 trial. HbA1c: blood glucose [mmol/mol]. SBP: systolic blood pressure [mmHg]. DBP: diastolic blood pressure [mmHg]. BMI: body mass index [kg/m2].

	HbA1c		**SBP**		**DBP**		**BMI**	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Real	0.209	344.89	0.421	134.92	0.415	47.13	0.814	4.23
Synthetic	0.037	420.11	0.174	192.48	0.096	72.86	0.743	5.86

Feature importance. When we examine the importance of the features used to make the above predictions, we observe that in all post-features the fake RF model correctly makes use of the corresponding pre-feature (e.g., pre-systolic BP is highly predictive of post-systolic BP).

A



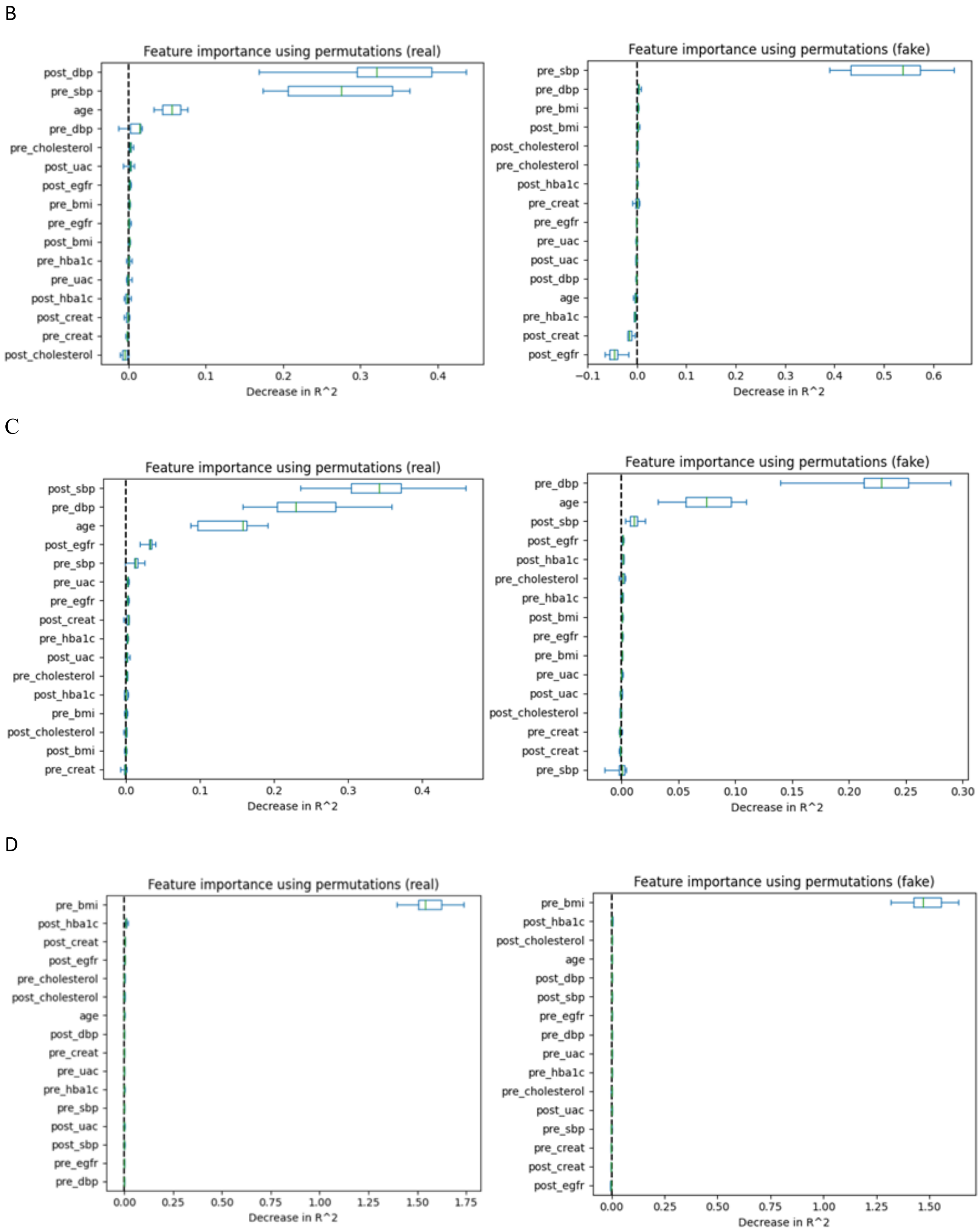


Figure B.2. Ranking the most important features (descending order) used by the random forest models, trained on real (left) and synthetic (right) *post-measurement and confounding* data, for predicting the LEAD-5 outcomes: **A.** HbA1c. **B.** Systolic blood pressure**. C.** Diastolic blood pressure. **D**. BMI**.** Boxplots summarise the change in model performance over 10 random and independent permutations to each feature. The dashed line illustrates whether

removing the feature is likely to worsen performance (right-hand side) or improve performance (left-hand side), indicating that the feature is more or less important for predictions, respectively.