MRC-GAN Deliverable 2.2: Outcomes of Virtual Trial Emulations

Abstract This report describes the final outcomes of the MRC-GAN research project, which is focused on the (**a**) evaluation of the synthetic data sampled from the trained virtual trial emulation model and (**b**) comparison of the outcomes from the virtual trial emulations against the LEAD-5 trials, and extended counterfactual emulations to prediction of 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 are focused on the replication of the existing LEAD-5 trials, and the second type of experiments attempt to emulate counterfactual scenarios where different drugs are applied to the same patients to support clinical decision making. When the patients meet the LEAD-5 patient baseline characteristics, the trial emulations produce the same ranking between the three drugs as LEAD-5 based on HbA_{1c} and BMI measurements – the effect sizes are estimated with both the average treatment effect (i.e. difference between pre- and post-treatment measurements) and the pairwise difference-in-differences between pairs of the drugs. Our experiments based on either independent sampling of virtual patients for the three treatment groups, counterfactual emulations on the same group of virtual patients, or counterfactual emulations on the real patients, 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. 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.

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

Project context 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.

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., $\boldsymbol{f}$) 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 $\boldsymbol{X}$, 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,

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

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

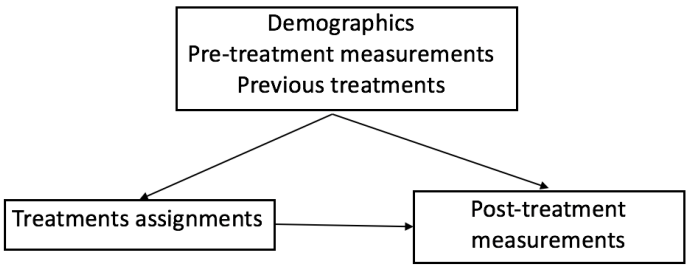

Figure 1 Overall causal structure between treatment, post measurements and confounders

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.

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 l^{th} 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:

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

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 j^{th} 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 $P\left(\widetilde{\boldsymbol{X}}\right)$. Specifically, the conditional generator at node i: $\tilde{X}_i = f_i\left(\boldsymbol{X}; \boldsymbol{Z}_i; W_i^1, W_i^2\right)$ takes samples from its conditional probability distribution $\boldsymbol{P}\left(X_i \middle| X_{\mathrm{pa}(i)}\right)$. Together, this allows sampling to generate samples $\widetilde{\boldsymbol{X}}$ from the underlying joint distribution $P(\boldsymbol{X})$, which is factorised with the local conditional probability distributions.

3. Trial Emulation

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 emulate an established clinical trial, namely, the LEAD-5 (Liraglutide Effect and Action in Diabetes), and compare 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* (i.e. age) 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]; *HBA1c* [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[ ].

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. the 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 1-1 below.

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

	HbA_{1c}		**Systolic Blood Pressure**		**BMI**	
	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 2 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 1-2 shows pairwise comparisons of the drugs on HbA1c, systolic blood pressure and BMI, which is calculated with difference-in-differences method[8].

All of Table 1-1, Table 1-2 and Figure 2 show that all 3 drugs have similar pre-treatment measurement and *GLP-1* produces the largest reduction of these 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 ranking in terms of the HbA_{1c} and BMI measurement. 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 1-2 LEAD-5 emulation results:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA_{1c}			**Systolic Blood Pressure**			**BMI**		
	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

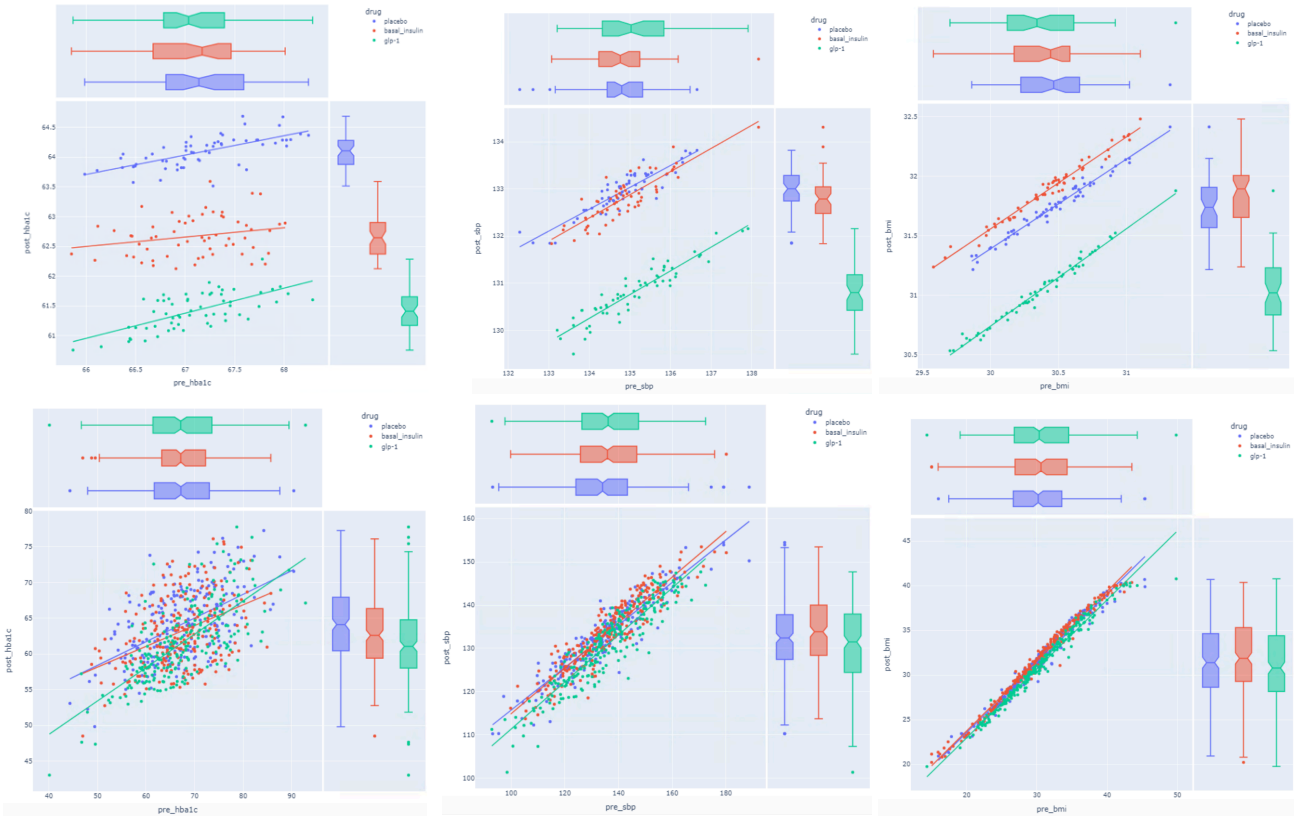

Figure 2 Outcomes of the virtual trial emulations: The performance difference between *GLP-1*, *basal insulin* and *placebo* in HbA_{1c}, systolic blood pressure and BMI measurements. Each drug is visualized by pre-treatment HbA_{1c} (x-axis) vs post-treatment HbA_{1c} (y-axis) and is colour coded. We compare the performance of these three drug 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: Comparison 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 this 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, which 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 the LEAD-5 outcomes. The discrepancy lies within the systolic blood pressure measurement. Our simulation 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 significance between *GLP-1*vs *basal insulin* reduction.

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 the patients meeting the inclusion criteria are recruited and randomly assigned to a drug-group (either *GLP-1*, *basal insulin* or *placebo*). So 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 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 the difference of the effects of the same patient with different drugs.

Table 2-1 shows the average treatment effect (from the *M=60* trials) of each of the three drugs estimated with the counterfactual emulations.

Table 2-1 Counterfactual emulation results:

Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA1c		**Systolic Blood Pressure**		**BMI**	
	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 2-2 shows pairwise difference-in-differences of the drug effects from the counterfactual emulation model.

Table 2-2 Counterfactual emulation results:

Pairwise comparisons of treatment effect with difference-in-differences

	HbA1c			**Systolic Blood Pressure**			**BMI**		
	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 previous section. As reported in Table 2-1 and Table 2-2, *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 is again has reproduced the ranking of the three drugs from 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 the same patient under multiple treatment scenarios to support clinical decision making. We do not need to match individual patients within different groups as conventionally required in the observational studies.

The next section further extends the counterfactual emulation to the data from real patients.

,

3.2 **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 administer 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*, *GLP-1* vs *basal insulin* and *basal insulin* vs *placebo*. The results of the counterfactual simulations are presented in Table 3-1(average treatment effect) and Table 3-2 (pairwise difference-in-differences) below.

Table 3-1a Counterfactual emulation on *GLP-1* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA1c		**Systolic Blood Pressure**		**BMI**	
	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 3-2a Counterfactual emulation on *GLP-1* drug group:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA1c			**Systolic Blood Pressure**			**BMI**		
	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 3-1a and Table 3-2a are the results for the *GLP-1* group. This is the group where patients took *GLP-1* in their treatments. 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} 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 3-1b Counterfactual emulation on *basial insulin* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA1c		**Systolic Blood Pressure**		**BMI**	
	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 3-2b Counterfactual emulation on *basal insulin* drug group:
Pairwise comparisons of treatment effect with difference-in-differences

	HbA1c			**Systolic Blood Pressure**			**BMI**		
	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 3-1b and Table 3-2b are the results for the *basal insulin* group. This is the group where patients took *basal insulin* in their treatments. 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} 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 3-1c Counterfactual emulation on *placebo* drug group:
Average treatment effect measured by differences between pre- and post-treatment measurements.

	HbA1c		**Systolic Blood Pressure**		**BMI**	
	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 3-2 a-c present pairwise comparisons of the treatment effect with difference-in-differences methods for each of the three drug groups.

Table 3-2c Counterfactual emulation on *placebo* drug group::
Pairwise comparisons of treatment effect with difference-in-differences

	HbA1c			**Systolic Blood Pressure**			**BMI**		
	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 3-1c and Table 3-2c are the results for the *placebo* group. This is the group where patients took *placebo* in their treatments. 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} 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 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.

Figure 3 visualise the results of the counterfactual emulation on the three groups, respectively.

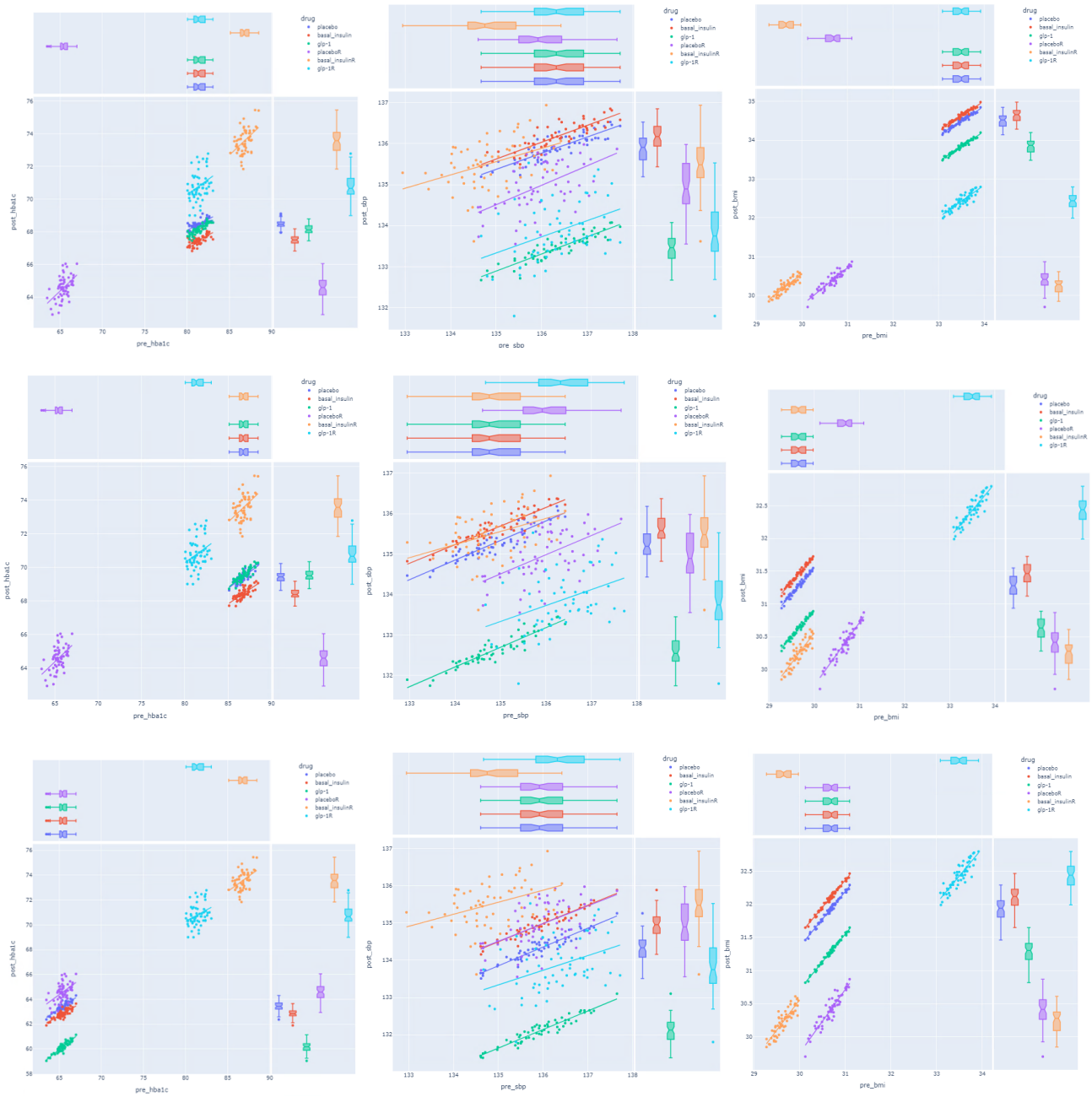

Figure 3 Visualisation of the counterfactual emulation results of the *GLP-1*, *basal insulin* and *placebo* groups. (a) Top row: *GLP-1* group results measured by (from left to right) HbA_{1c}, systolic blood pressure and BMI; (b) Middle row: *basal insulin* group results measured by (from left to right) 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 emulation on real patients in Table 3-1, Table 3-2 show that *GLP-1* remains as the drug with best performance over the *placebo* group. In fact, the emulation results on the *placebo* group agree well with LEAD-5 on the HbA1c measurement in terms of the ranking of the three drugs. 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} measurement. However, the results of the counterfactual emulation 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 normal distribution in LEAD-5. Interestingly, according to the emulation, *GLP-1* is not the best drug in the *GLP-1* group, however, *basal insulin* appeared to be the right medicine for the patients in the *basal insulin* group, Another observation towards the emulation vs the real data shows that the emulations have slightly over-predicted the average treatment outcomes on all the three drugs.

Further experiments on trial emulations with the real patient data in SCI diabetes data suggest that LEAD-5 results cannot be simply extrapolated to cover patients outside its baseline characteristics. When we directly applied the drugs to the patients that are randomly sampled from the datasets, the treatment outcomes of *GLP-1* are not aligned with LEAD-5. This suggests that *GLP-* is probably not the best drug to all patients. However, if we sample from the mean of the patient population with a much narrowed (0.5x) standard deviation, the *GLP-1* performance become much improved – see Figure 4. Our analysis shows that the narrowed data distribution is much closer to the LEAD-5 baseline characteristics.

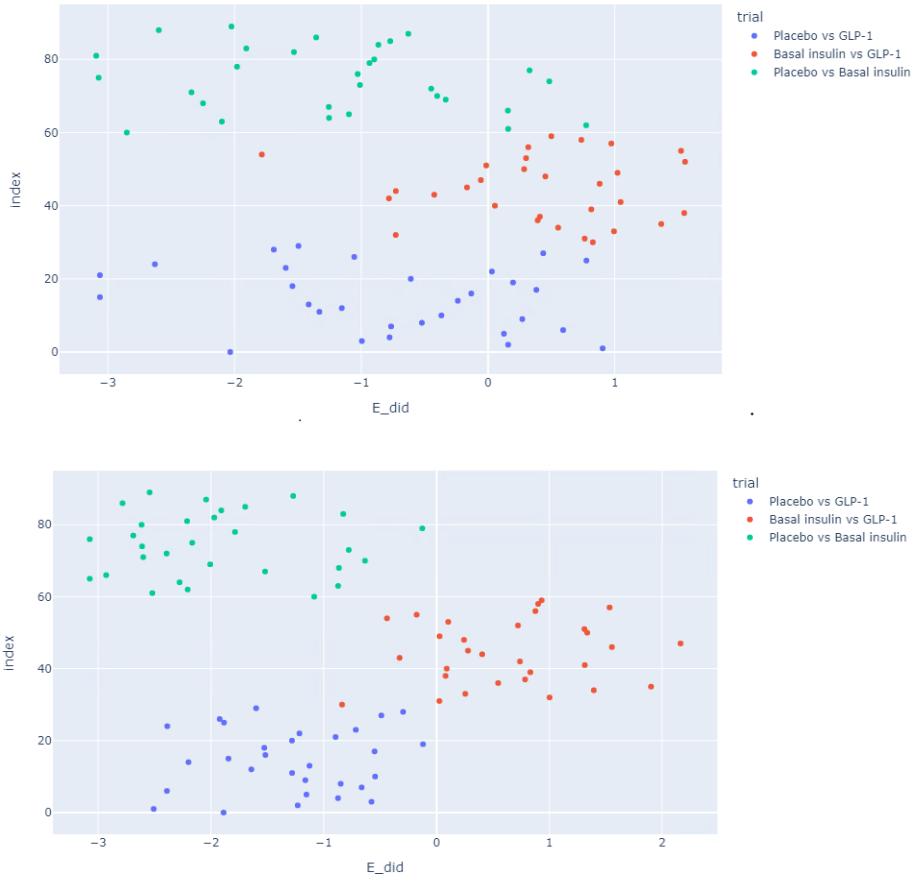

Figure 4 Comparisons of difference-in-differences results from three pairs of drugs in two trial emulations on real patient data. (**a**) Top row: patient data are sampled according to the mean and standard deviation of the real dataset; (**b**) Bottom row: patient data sampled from narrowed (0.5x) standard deviation.

4. Analysis and conclusion

We have mainly conducted two types of experiments on the virtual trial emulation in the context of T2DM treatment with three different drugs. The first type of experiments are focused on the replication of the existing LEAD-5 trials, and the second type of experiments attempt to emulate counterfactual scenarios where different drugs are applied to the same patients to support clinical decision making. When the patients meet the LEAD-5 patient baseline characteristics, the trial emulations produce the same ranking between the three drugs as LEAD-5 based on HbA_{1c} and BMI measurements – the effect sizes are estimated with both the average treatment effect (i.e. difference between pre-and post-treatment measurements) and the pairwise difference-in-differences between pairs of the drugs. Our experiments based on either independent sampling of virtual patients for the three treatment groups, counterfactual emulation on the same group of virtual patients, or counterfactual emulation on the real patients, 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. 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 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 agrees with LEAD-5 very well. 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 predicted significant reduction. In addition, our experiments of the trial emulation on real patients have compared the results with the real data (where the results are comparable). Most of the emulation have produced the right effect in positive and negative signs. However, we have also noticed over estimation in some of the cases.

We recon the difference between the trial and the emulations may come 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 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 follow the trial populations of LEAD-5 by sampling from the normal distributions of the 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, the variables involved in the virtual trial emulation are not able to fully cover several variables that are specified in the patient baseline characteristics of LEAD-5; we have only used *age* to describe the patient demographics; several clinical measurements such as blood pressures are based on yearly means, and so on and so forth. Another limitation within the real-world dataset is the pre-treatment history of the patients might also 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 emulation and trial results. Hidden confounders can be responsible for generating bias in the emulation. The current model didn't catch these factors.
- Practical errors in the technologies themselves (e.g. training and computational errors) may also contribute to the difference at certain degrees.

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