A framework for detecting causal effects of risk factors at an individual level based on principles of Mendelian randomization: Applications to cardiovascular medicine

Yujia SHI[1], Yong XIANG[1], Yuxin YE[1], Tingwei HE[1], Hon-Cheong SO[1-7]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong

[2]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China

[3] Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong

[4] CUHK Shenzhen Research Institute, Shenzhen, China

[5]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong

[6]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China

[7]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China

This Version: 8 Aug 2022

Abstract

Mendelian Randomization (MR), which uses genetic variants as instruments for causal inference, has become a popular approach to evaluate effects of risk factors. However, most MR studies only focus on the *average* causal effects in the population. With the advent of precision medicine, the *individualized* treatment effect (ITE) is often of greater interest. For example, a risk factor may confer greater risk to some individuals but not others, and the corresponding benefit of treating the risk factor may differ across different individuals.

We propose a new framework extending the idea of MR to study *individualized* causal effects. We presented several ITE estimation approaches under an MR framework, mainly based on principles of the R-learner. We also proposed how one may combine ITE predictions from different learners. Permutation testing methods were proposed to assess the presence of heterogeneity of causal effects. We employed polygenic risk scores (PRS) as the instrument, and showed that removing potential pleiotropic SNPs may improve accuracy of ITE estimates. Our approach was verified by extensive simulations.

We further applied our framework to study the individualized causal effect of several lipid-related traits, including LDL, HDL, triglycerides and total cholesterol, on the risk of coronary heart disease (CHD) and stroke in the UK Biobank. We showed that an increased level of LDL is causally associated with elevated risks of CHD and stroke, and the effect is significantly heterogeneous for CHD. Similar results are also found for triglycerides. We also revealed clinical factors contributing to the heterogeneity of ITE based on Shapley value analysis.

1 Introduction

The rising incidence and mortality of chronic diseases are placing a heavy burden on many countries in the past decades[1]. Therefore, identifying potential causal risk factors and designing appropriate interventions becomes a high priority. In the past days, epidemiologists were only interested in studying the average causal effect of candidate intervention, which neglects the importance of population heterogeneity. The existence of heterogeneity implies that people may benefit differently from some interventions. For instance, metformin treatment may have a heterogeneous effect on diabetes patients[2]. For example, patients with higher diabetes risk exhibit a larger absolute risk reduction than low diabetes risk patients. For lower-risk patients, metformin treatment may even cause harmful effect[3]. These studies emphasize the importance of estimating the individualized treatment effects (ITE). Otherwise, applying drug to patients with poor response will introduce an extra financial burden and may even cause unforeseeable health problems. To maximize the efficiency of intervention in the population and to minimize the cost of controlling the disease, estimating how much a given patient may benefit from an intervention (or prevention of a risk factor) is of utmost importance. In this study, we aimed to estimate the individualized causal treatment effect (ITE) of a given intervention to individual patients, leveraging principles of Mendelian randomization (MR).

It is widely accepted that the most accurate approach to estimate the causal effect is to set up a randomized controlled trial (RCT), in which the confounding factors are controlled, sufficient participants are recruited, and the randomness of the treatment assignment is ensured[4]. However, it is challenging to meet RCT's conditions because of the limitations like ethical concerns or logistic constraints[5,6]. For instance, if people want to study whether drinking alcohol has a causal effect on the occurrence of stroke, they need to allocate the subjects into treatment and control groups randomly, which is impractical and ethically questionable.

In practice, conducting RCTs is usually not feasible due to the limitations discussed above. Therefore, people mostly use observational data to infer the causal effect, especially in epidemiology studies. Unlike RCT, a major concern of observational studies is that unmeasured confounding may affect causal inference. Mendelian Randomization (MR) is a useful approach to reduce the risk of unmeasured confounding and it is also largely immune to reverse causality. In MR, genetic variants can be used as instruments to represent the exposure[7].

After years of development and innovation, various statistical methods have been developed invented for instrumental variable estimation, including the Wald ratio method, two-stage least squares, and semi-parametric methods[8]. The ratio method, also called the Wald method, is the most straightforward approach to estimating the causal effect, in which the ratio estimate of the causal effect is defined as the ratio between the coefficient of regressing the exposure on the IV $\hat{\beta}_{X|Z}$ and the coefficient of regressing the outcome on the IV $\hat{\beta}_{Y|Z}$ [9]. Another widely-used IV approach in MR is the two-stage method, which consists of two regression stages, including the regression of the exposure on the IVs and the regression of the outcome on the fitted value estimated from the first stage regression model[10]. Although

these methods are robust and flexible, they still have limitations. An important one is that they can only estimate an average causal effect without considering the heterogeneity of the population. In addition, the usage of the weak instrument may introduce bias in the estimator, especially for methods like the two-stage least square (2SLS) approach.

With the weak instrument limitation, a possible solution is to use the allele scores (also called polygenic risk score, PRS), a score that summarizes multiple genetic variants associated with the exposure[11]. Theoretically, using a polygenic risk score (PRS) as the instrument may be affected by pleiotropy. In other words, if we include the pleiotropic genetic variants in the PRS calculation, the PRS itself may violate the valid instrument assumption discussed above.

In order to achieve the goal of estimating the individualized causal effect under the MR scheme, we introduce a novel framework capable of inferring the *individualized* causal effects in observational studies taking the polygenic risk score as an instrument. We presented several ITE estimation methodologies under the MR framework, based on principles of "R-learners"[12]. These methods are also flexible as they leverage machine learning approaches for modelling. In addition, we proposed how one may combine ITE predictions from different learners.

To ensure the valid instrument assumption for MR, we applied the contamination mixture approach to remove the pleiotropic SNPs before calculating the PRS[13]. We also presented permutation-based approaches to assess the presence of heterogeneity. Our framework has several advantages. First, we can obtain the *individualized* causal treatment effect for every patient, which may facilitate precision medicine as intervention on the risk factors may be tailored for each person based on predicted ITE. Our proposed permutation-based test provides flexible and robust testing of the presence of heterogeneity. To assess the feasibility and validity of our framework, we set up a simulation study and further applied our framework to study the causal effect of several lipid traits on the risk of heart disease and stroke.

2 Methods

2.1 Set-up and notation

2.1.1 Rubin's causal model

It is widely accepted that a causal model needs to be formalized first to identify the potential causal effect. A well-established and popular choice is the Neyman-Rubin causal model, also called the potential outcome framework[14,15]. We consider a dataset with N units, indexed by $i = 1, \dots, N$. Following the potential outcome framework, we define the potential outcome for unit i in treatment and control status as Y_{i1} and Y_{i0}, respectively. For each unit, we let X_i be a vector consisting of M covariates and Z_i be a continuous instrument variable. We further define $W_i \in \{0,1\}$ as a binary indicator for the treatment, where $W_i = 0$ means that the unit i does not receive any treatment and $W_i = 1$ means that the unit i is receiving the treatment. Given the formalization above, our data can be regarded as a set of quadruple data point $\left(Y_i^{Status}, W_i, X_i, Z_i\right)$ units, indexed from 1 to N. In this case, we further define the unit-level causal effect as the difference between two potential outcomes Y_{i1} and Y_{i0}, $\tau_i = Y_{i1} - Y_{i0}$.

The framework discussed above is under a binary treatment setting. However, in many epidemiology studies, the risk factors are continuous variables, and it is difficult to define an arbitrary cut-off to partition the population into treatment and control groups. The main difference between binary and continuous settings is that we no longer define a binary indicator for the treatment[16]. In this case, the unit-level causal effect is defined as the effect of unit increment of treatment on the outcome, $\tau_i = Y_i^{W+\Delta W} - Y_i^W$.

2.1.2 Latent conditional average treatment effects

With the formalized causal model, we can specify two main goals in the causal inference with the instrument: estimating the local average treatment effect (LATE) and the conditional local average treatment effect (CLATE), where LATE is defined as, $LATE = E[Y_{i1} - Y_{i0}|Z_i]$ and CLATE is defined as, $CLATE = E[Y_{i1} - Y_{i0}|X_i, Z_i]$. The main focus of our study is on estimating the conditional local average treatment effect $\tau(x)$ of interested risk factors.

2.2 Generalized Random Forest (GRF)

The generalized random forest (GRF) is a forest-based estimator introduced by Athey et al. It preserves the main components of classic random forest, including recursive partitioning and subsampling[17]. For details of the methodology and notations please refer to[17]. An important improvement of GRF is that the final estimation is not obtained via averaging the predictions from each ensemble. Instead, the GRF developed an adaptive nearest neighbor estimator allowing a more problem-specific estimation of weights $\alpha_i(x)$ that can be used in local estimation equations $\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \underset{\theta,\nu}{argmin} \left|\left|\sum_i^n \alpha_i(x)\psi_{\theta,\nu}(O_i)\right|\right|$ [18]. We can think of the weight $\alpha_i(x)$ as a measurement of the relevance between ith training sample and given sample x in fitting $\theta(\cdot)$, and GRF use the frequency of ith training sample and given sample x are located in the same leaf to measure such relevance, defined as, $\alpha_{bi}(x) = \frac{1(X_i \in L_b(x))}{|L_b(x)|}$, $\alpha_i(x) = \frac{1}{B}\sum_{b=1}^B \alpha_{bi}(x)$.

The instrumental causal forest is an important application of GRF, in which the gradient-based labelling ρ_i is modified to fit the instrumental case. In the regular GRF causal forest, the gradient-based labeling is defined as, $\rho_i = Y_i - \bar{Y}_p - \left(W_i - \bar{W}_p\right)\hat{\tau}_p$, while in the instrumental causal forest, the gradient-based labeling is defined as, $\rho_i = \left(Z_i - \bar{Z}_p\right)\left(\left(Y_i - \bar{Y}_p\right) - \left(W_i - \bar{W}_p\right)\hat{\tau}_p\right)$. The algorithm will further run a standard CART regression split by splitting the parent node into two child nodes C_1 and C_2 with maximizing the criterion:

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{\left|i: X_i \in C_j\right|}\left(\sum_{i:X_i \in C_j} \rho_i\right)^2$$

In addition to the difference in making splits, the instrumental GRF modifies the prediction step by introducing the forest-weighted two-stage least squares approach, allowing a more accurate causal effect estimation in the instrumental case.

2.4 Double robustness instrumental variable estimator (DRIV)

Unlike instrumental causal forest, DRIV mainly focuses on designing a novel loss function used in machine learning[19], which starts by minimizing an empirical analog of the square loss,

$$min_{\theta\in\Theta} L^1(\theta; E[Y|X], E[T|Z,X], E[T|X]) \coloneqq E\left[\left(Y - E[Y|X] - \theta(X)(E[T|Z,X] - E[T|X])\right)^2\right]$$

where Z is the instrumental variable and θ represents estimate of the individualized treatment effect. The square loss can be derived from the classic moment function proposed by Chernozhukov et al. aims to estimate the ATE with instrument[20],

$$E\left[\left(Y - E[Y|X] - \theta(T - E[T|X])\right)(Z - E[Z|X])\right] = 0$$

The nuisance terms $E[Y|X], E[T|Z,X]$ and $E[T|X]$ should be estimated from another training set separately. However, the $\hat{\theta}$ estimated from the first step is not a consistent estimate if effect heterogeneity with respect to X or heterogeneous compliance exists. A double robustness approach can be applied to resolve the issue,

$$min_{\theta\in\Theta} L^2\left(\theta; \theta_{pre}, \beta, E[Y|X], E[Z|X], E[T|X]\right) \coloneqq E\left[\left(\theta_{pre}(X) + \frac{\{\tilde{Y} - \theta_p(X)\tilde{T}\}\tilde{Z}}{\beta(X)} - \theta(X)\right)^2\right]$$

where $\beta = E[(T - E[T|X])(Z - E[Z|X])|X]$. Note that we can further apply SHAP analysis directly to assess the effect modifier importance since the DRIV allows a flexible specification of machine learning models in the training step.

2.5 Polygenic risk score

A single trait-associated genetic variant can only explain a small proportion of variation on phenotype, and it is insufficient to measure the potential risk of an individual getting disease or exhibiting an abnormal phenotype with only a single genetic variant. Enormous evidence suggests that many common disease/phenotype is mediated by multiple genetic variants simultaneously, where each individual genetic variant only contributes a small effect[21,22]. Polygenic risk score can be used to summarize the estimated effect of multiple trait-associated genetic variants on an individual's phenotype, which is usually defined as a weighted sum of trait-associated risk alleles across multiple genetic loci[23].

In our study, we apply PRSice-2 to calculate the individualized PRS as an instrument for further analysis. PRSice-2 is an efficient PRS calculating software that can automate PRS analyses, which can identify the precise P-value threshold by calculating the PRS at multiple threshold[24].

2.6 Contamination mixture model

A major concern of applying an allelic score as an instrument in MR is that it may violate the assumption of an instrumental variable. Including pleiotropic SNPs in calculating the allelic score may significantly affect the overall performance of the results[11]. Therefore, it is necessary to remove the pleiotropic SNPs to ensure the calculated PRS is a valid instrument. To achieve the goal, we apply the contamination mixture (ConMix) approach to identify

invalid SNPs in our framework[13]. The contamination mixture approach assumes that the valid genetic variants follow a normal distribution centered at the true causal parameter θ with standard deviation equal to the ratio estimate's standard error $se(\hat{\theta}_j)$, while the invalid genetic variant is normally distributed around 0 with variance comprising of the uncertainty in the ratio estimates (which is $se(\hat{\theta}_j)$) and proposed variability in the invalid estimand (which denoted as ψ^2). The ConMix approach first specify a likelihood function integrating information from all genetic variants, defined as,

$$
\begin{aligned}
L(\theta, \xi) &= \prod_j \xi_j L_{V,j} + (1 - \xi_j) L_{F,j} \\
&= \prod_j \xi_j \times \frac{1}{\sqrt{2\pi se(\hat{\theta}_j)^2}} exp\left(- \frac{(\theta - \hat{\theta}_j)^2}{2se(\hat{\theta}_j)^2} \right) \\
&\quad + (1 - \xi_j) \times \frac{1}{\sqrt{2\pi \left(\psi^2 + se(\hat{\theta}_j)^2 \right)}} exp\left(- \frac{-\hat{\theta}_j{}^2}{2\left(\psi^2 + se(\hat{\theta}_j)^2 \right)} \right)
\end{aligned}
$$

where ξ is a vector denoting the configuration of the valid and invalid instruments. In the likelihood, $\xi_j = 1$ if genetic variant j is a valid instrument, while $\xi_j = 0$ if genetic variant j is invalid. Instead of making inferences with this likelihood, the ConMix method applies a profile likelihood approach, in which the causal estimate θ is assumed to be fixed. When θ is fixed, the optimal θ that maximizes the aforementioned likelihood function can be easily inferred, that is, ξ_j equals to 1 when $L_{V,j}$ greater than $L_{F,j}$ for genetic variant j. The ConMix approach will iterate a range of values of θ and picks the one maximizing the profile likelihood. In summary, the ConMix approach tries to identify the invalid SNPs based on the distance between variant-specific estimate $\hat{\theta}_j$ and the proposed causal parameter θ, in other words, if the $\hat{\theta}_j$ is closer to the θ, the SNP j is more likely to be a valid SNP. Here we removed the SNPs that are likely pleiotropic based on the ConMix approach before construction of the PRS as instrument.

2.7 Aggregating multi-model prediction with the R-loss

Currently dozens of methods are available for inferring individual treatment effect, and some of them may give a better performance in some specific cases. Therefore, applying an approach to aggregate the results from different methods can help get improved prediction accuracy[25]. In our framework, we apply the R-stack approach to averaging the model predictions[12]. R-stack is mainly developed based on the R-loss, defined as,

$$\hat{\tau}(x) = \sum_{k=1}^{K} \alpha_k \hat{\tau}_k(x)$$

$$(\hat{b}, \hat{\alpha}) = \underset{b,\alpha}{\operatorname{argmin}}\{\sum_{i=1}^{n}\Bigg[[\Big(Y_i - \hat{m}^{(-i)}(X_i)\Big) - b - \Bigg(\sum_{k=1}^{K}\alpha_k\hat{\tau}_k(x)\Bigg)\Big(W_i - \hat{e}^{(-i)}(X_i)\Big]\Big(Z_i - \hat{Z}^{(-i)}(X_i)\Big)\Bigg]^2 : \alpha_k \geq 0\}$$

where b is an intercept to absorb potential bias of $\hat{m}$, and α_k is the coefficient for method k.

2.8 Assessing the presence of treatment effect heterogeneity

In addition to estimating the individual treatment effect, we also presented two permutation-based methods to assess the presence of heterogeneity among the estimated treatment effect. Typically, the heterogeneity of treatment effect can be defined as a non-random, explainable variability in the direction and magnitude of individualized treatment effect obtained from a population[26]. Another direction to think of heterogeneity of treatment effects is that whether the predicted treatment effects are different from the average effect by chance[27]. These definitions offer valuable insight in developing methods for us to estimate ITE and assess heterogeneity. For example, in each split, the causal tree will try to maximize the variance of the estimated treatment effect across leaves as well as penalize the uncertainty of estimated treatment effects[28]. In this case, if covariates cannot contribute to the heterogeneity, the variance of predicted individual treatment effect across leaves should be smaller than splitting on covariates that contribute to the heterogeneity, inspiring us to develop the heterogeneity testing methods based on covariate permutation.

Here we present two permutation-based methods, including the permutation-variance and permutation-τ-risk test, to evaluate whether there is significant heterogeneity. While the same principles can also be applied in ordinary HTE models (see ref[29], Chapter 4), here the tests are specifically catered for our framework with instrumental variables, especially for the test based on the (modified) τ-risk.

The two tests are briefly described below. For the permutation-variance test, we aimed to compare the variance of predicted individualized treatment effect estimated from original set of covariates and permuted covariates. The rationale behind this approach the if we permute the covariates, they should no longer to modify the treatment effects and contribute to the τ heterogeneity. In this case, if there is heterogeneity that the covariates can explain, we will have $\widehat{Var}_{observed} \geq \widehat{Var}_{perm}$, where Var denotes variance of τ. We can further set up a statistical test by repeating the permutation procedure *N* times, and the p-value can be obtained by $pvalue = \frac{\#(\widehat{Var}_{observed} \leq \widehat{Var}_{perm})}{N}$ .Note that we did not re-calculate the residualized Y and W; we only permute the covariates when the covariates are fit into the causal forest (or other ITE-finding procedures). The reason is that we still wish to correct Y and W for possible confounders; the permutation is employed to mimic the case in which no covariates serve as effect modifiers.

Another method we proposed is the permutation-τ-risk test, which aims to test whether introducing heterogeneity in treatment effects can lead to a better goodness-of-fit than

assuming homogenous treatment effects. To assess the goodness-of-fit of a causal model, Nie et al. propose a loss function based on Robinson transformation[12,30], defined as,

$$\tau^*(\cdot) = arg\ \min_{\tau}\{E([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\tau(X_i)]^2)\}$$

It is also referred to as the R-loss function. In the instrumental case, we propose a modification of the R-loss to

$$\hat{L}\{\tau(\cdot)\} = \sum_{i}\left([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\tau(X_i)](Z_i - E[Z_i|X_i])\right)^2$$

We also termed this loss as τ risk.

The rationale behind the permutation-τ-risk test is that we expect the total R-loss of a heterogeneous model to be smaller than homogeneous model. In this case, we define the improvement of introducing heterogeneity as,

$$\hat{L}_{improve} = \sum_{i}\left([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\bar{\tau}(X_i)](Z_i - E[Z_i|X_i])\right)^2 - \sum_{i}\left([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\hat{\tau}(X_i)](Z_i - E[Z_i|X_i])\right)^2$$

Where $\bar{\tau}(X_i)$ represents the average treatment effect (i.e. assume no heterogeneity in treatment effects). We then permute the covariates model the null distribution of the $\hat{L}_{improve}$, and the permutation p-value can be computed as

$\mathrm{P}value = \frac{\#(\hat{L}_{imporve-observed} \le \hat{L}_{improve-per})}{N}$. Similar to the above, the residualized Y, W and Z would not be re-calculated; the covariates were permuted when they were fit into the forest (or other ITE-finding procedure).

2.9 Measuring variable importance

In addition to identifying the individualized treatment effect, we aimed to identify which covariate may contribute more to the heterogeneity. We mainly used two approaches to achieve the goal. The first approach is mainly designed for the generalized random forest algorithm since it calculates the variable importance based on the split frequencies. The idea is firstly proposed by Eoghan et al.[31], in which they define the variable importance as the measurement of how often the variable is selected by the causal forest algorithm to perform the split. Theoretically, the feature that can maximize the variance of treatment effect across leaves is tended to be selected in each split; thus, the selection frequency can represent the importance of features. The grf package also adopts the split-count idea and implements a simple but efficient algorithm to calculate the feature's importance[17], where the importance of variable x_i of interest up to a depth of n is defined as,

$$\mathrm{imp}(x_j) = \frac{\sum_{k=1}^{n}\left[\frac{\sum_{\text{all trees}} \text{number depth-}k \text{ splits on } x_j}{\sum_{\text{all trees}} \text{total number of depth-}k \text{ splits}}\right]k^{-2}}{\sum_{k=1}^{n} k^{-2}}$$

Although the split-frequency-based approach is a flexible method to identify variable importance, it has a few drawbacks. First, it can only be applied to machine learning models with node splitting procedures, such as decision trees and random forests. Another concern is that the split-frequency approach cannot provide a reliable measurement of global feature importance owing to the inconsistency[32]. In other words, the feature's importance calculated via ways like gain[33] and split frequency[34] may sometimes not be able to demonstrate the feature's true impact on the model's output.

Besides, considering the coming age of precision medicine, we are also interested in how much a covariate would contribute to the individual prediction, which is impossible to interpret via a split-frequency-based approach. To avoid the drawbacks of the split-frequency-based approach, we apply SHAP (SHapley Additive exPlanations) values proposed by Lundberg et al. in our framework, which can provide a robust and consistent measurement of feature importance[35]. Lundberg et al. further developed TreeExplainer to efficiently compute SHAP value for tree-based models like random forest and XGBoost[36], which will be used by our framework to interpret the causal forest model. Briefly, the Shapley value captures the contribution of each feature after consideration of the rest of the features, and Shapley value can be computed for each individual as well as for the whole sample.

2.10 Individualized treatment effect estimation framework

To summarize, this project aims to present a novel causal analytic framework to study the causal effect of risk factors and potential drugs for chronic disease. We integrate the idea of Mendelian Randomization (MR) in identifying individual treatment effects, which provides a relatively robust individualized causal effect estimation in observational studies.

Our framework starts by identifying valid SNPs, in which we minimize the risk of pleiotropic effect. Next, we will use the valid SNPs to estimate the polygenic risk score (PRS), which will be further set as the instrument. With a proper estimation of PRS, we apply approaches, including GRF and DRIV, to estimate the individualized treatment effect. Two permutation-based testing methods will then be used to test the presence of heterogeneity.

2.11 Simulation study

To illustrate the validity of the proposed framework, we conducted two simulations to assess the performance of our proposed framework in estimating ITE and the power of our proposed heterogeneity testing methods.

We first set up a simulation study with different pleiotropic scenarios to compare our proposed framework's performance with regular causal forest. Three different pleiotropic scenarios were included in our simulation[13]:

1. Balanced pleiotropy: some genetic variants directly affect the outcome, and the effects are all positive or negative.
2. Directional pleiotropy: some genetic variants directly affect the outcome, and the effects are all positive.
3. Pleiotropy via a confounder: some genetic variants affect the outcome via a confounder, and these effects relate to the instrument strength.

The simulation is set up following a similar idea from Burgess et al., and the data is generated as follows:

$$U_i = \sum_{j=1}^{J} \xi_j G_{ij} + \epsilon_{Ui}$$
$$T_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_{Xi}$$
$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \tau(X_i) T_i + U_i + \epsilon_{Yi}$$

where,

$$G_{ij} \sim_{i.i.d} Binomial(2,0.3)$$
$$\epsilon_{Ui}, \epsilon_{Xi}, \epsilon_{Yi} \sim_{i.i.d} \mathrm{N}(0,1)$$
$$\gamma_j \sim_{i.i.d} Uniform(0.03,0.1).$$

We included eight different treatment effect functions $\tau(X_i)$ s; details can be found in Appendix A[37]. We simulate 100 genetic variants in each scenario and consider three cases with 20, 40, and 60 invalid instruments, respectively. We set the α_j and ξ_j to 0 for valid instruments. For invalid instruments, α_j and ξ_j will be set differently for different scenarios. In the balanced pleiotropy scenario (scenario 1), the α_j is simulated from a uniform distribution on the interval from -0.1 and 0.1, and the ξ_j will be set to 0. For the directional pleiotropy scenario (scenario 2), we simulate the α_j from a uniform distribution on the interval from 0 to 0.1, and we set ξ_j to 0. In scenario 3 (Pleiotropy via a confounder), the α_j will be set to 0 and ξ_j will be draw from a uniform distribution on the interval from -0.1 to 0.1. As we aimed to use the polygenic risk score as the instrument to infer the causal effect, we further fit a simple linear regression model to model the relationship between exposure and instruments. We will use the prediction from the linear regression model as the instrument in the simulation. The simulation will be repeated 50 times in each scenario. To avoid overfitting issues, we simulate four independent datasets in each case and are used for different steps of our framework.

The performance of our proposed HTE-testing methods is evaluated through simulation as well. We set up the simulation data following similar approaches discussed above, but only 40 invalid SNPs were included. Similarly, we repeat the simulation 50 times for each scenario.

2.12 Applications in real data: Lipid-related risk factors for coronary artery disease and stroke

Using data from the UK-Biobank study, we applied our framework to study the heterogeneous treatment effect for several lipid-related risk factors on coronary heart disease (CHD) and stroke. UK-Biobank is a large-scale biomedical database consisting of genetic and clinical data from 502,682 participants. The included clinical variables used for different traits are summarized in Table 1. We also convert the discrete variables to dummy variables.

Considering that we are using polygenic risk score as instrument, an extra data-cleaning on genetic data is necessary. We follow the recommended quality control pipeline of PRSice-2 to ensure the target data meets standards implemented in GWAS studies, in which we remove SNPs with low genotyping rate, low minor allele frequency, and individuals with low genotyping rate[38,39] following the default. Only variants significantly associated with the exposure will be included in the further analysis (P-value < 1e-8). Furthermore, considering that the UK Biobank dataset is large (*N* ~500,000) but number of stroke/CHD cases are small, we adopted a propensity-score matching approach to drawing healthy samples such that a case-control ratio of 1:1 is maintained in the study population[40]. This approach also greatly accelerates computational speed. Overall, there are 10938 patients included in the stroke study and 22158 patients included in the CHD study. In our real-data application, we also compared IV-GRF and DRIV approaches with a more standard (non-instrumental) approach in which the risk factor was directly modelled without using any genetic instruments. The causal forest (CF) approach (implemented in the R package GRF) was used to model ITE of the risk factors.

3 Results

3.1 Simulation Study

Figure 1 presents the simulation results for different treatment effect scenarios under various invalid SNPs counts settings. Under the balanced pleiotropy scenario, IV-GRF significantly outperforms the regular causal forest, which achieves a much better mean squared error (MSE) (Fig.1A). However, whether removing invalid pleiotropic SNPs seems not to affect the performance as we do not observe a significant difference in all treatment effect cases in the balanced pleiotropy scenario. We hypothesize this may be related to a mixture of positive and negative pleiotropic effects from invalid SNPs, and pleiotropic effects on the outcome can be cancelled out by each other. In this case, whether removing pleiotropic SNPs may not have a significant impact on the results.

Similarly, we observe a significant improvement of the MSE performance using IV-GRF in the other two pleiotropy scenarios, direction pleiotropy and pleiotropy via a confounder. As expected, IV-GRF outperforms regular causal forest in all treatment effect scenarios while the improvement gradually decreases with the increment of invalid SNPs counts. Furthermore, unlike the balanced pleiotropy scenarios, removing invalid pleiotropic SNPs can significantly improve the performance of the IV-GRF (Fig.1B, Fig.1C). The simulation results indicate that it is important to include a proper invalid pleiotropic SNPs removal step in the framework. We further assess the performance of the ConMix approach in removing invalid SNPs, and our results show that the ConMix approach can achieve an accuracy of around 80% in detecting invalid SNPs.

In addition to benchmarking the performance of applying PRS as an instrument in inferring heterogeneous treatment effects, we also perform a simulation to assess the validity of our proposed heterogeneity detecting methods. The results are summarized in Table 1. We notice that both methods achieve a good type 1 error control in a scenario with no heterogeneity (scenario 1). Also, both methods hold good power in several simple scenarios (scenario 2-5, 8), in which the heterogeneous treatment effect functions ($\tau(\cdot)$) are simple linear combinations of the same types of covariates or weak non-linear effects without interaction between different types of covariates. However, the results show that the permutation τ-risk test outperforms the permutation-variance test when an interaction between different types of covariates exists (Scenario 6). Surprisingly, the permutation-variance test has a poor power in scenario 7, where a strong non-linear effect exists in the $\tau(\cdot)$, while the permutation τ-risk test still holds relatively good power in this scenario.

3.2 Treatment effect of Lipid-related traits on coronary heart disease and stroke

3.2.1 Baseline Characteristics of included participants

The baseline characteristics of the study's continuous and dummy covariates are summarized in Tables 3 and 4, respectively. We also perform a simple partial F-test to measure whether the polygenic risk score is a weak instrument[41], and our results show that PRS can be regarded as a strong instrument as F-statistics in all cases are far greater than 10 (Table 5). The

covariates included for different lipid-related traits are summarized in Table 2. Most covariates are shared among different traits.

3.2.2 LDL

3.2.2.1 LDL-C imposes heterogeneous effects on CHD

We first applied our framework to study the causal relationship between LDL and two diseases, CHD and stroke, under a continuous exposure scenario. Our results confirmed that the increment of LDL is causally associated with increased risk of these two diseases (Fig. 2).

For CHD, we observe that IV-GRF, DRIV, and causal forest (CF) predict positive treatment effects on all included participants, while IV-GRF and DRIV give higher predictions than CF. Note that we modelled the original risk factor without using instruments under the CF approach, as described earlier. Analysis using IV-GRF and DRIV and LDL-C as a continuous treatment show that per unit increment of LDL-C will increase the individual CHD risk by around 8%, while for CF the risk increase was less than 5% (Fig. 2A).

Under the setting of a binary 'treatment' (hyperlipidaemia vs normal lipid levels), IV-GRF also gives higher treatment effect predictions. The result shows that lowering LDL-C level to a normal range will lead to around 30% relative risk reduction. Interestingly, we observe that CF only detects 10% CHD relative risk reduction (Supp Fig. 1A). It is unclear which method's prediction under the binary scenario is closer to the truth. We therefore search for other research works on the potential of applying statin therapy for cardiovascular events prevention. A meta-analysis shows that statin can efficiently decrease the occurrence of coronary heart events (RR = 0.73) [42]. Toth et al. estimate the statin-treatment-related reduction in coronary heart disease risk in untreated high-risk individuals with dyslipidaemia and find that using moderate and high dose statin can decrease coronary heart disease by 27% and 47%, respectively[43]. These studies' relative risk reduction level is consistent with our IV-GRF findings in the binary setting.

Importantly, compared to ordinary MR methods, the proposed IV-GRF and DRIV approaches allows a causal effect estimate for *each individual*.

Similar to CHD, the treatment effects predicted from IV-GRF and DRIV are higher than CF (which was not based on instruments) in stroke. We notice that CF failed to detect a causal relationship between LDL and stroke, however studies have suggested LDL-C as a risk factor for stroke[44,45] (Fig.2B, Supp Fig.1B). When LDL-C was modelled as a continuous trait, we observed that the treatment effect is positive, but weaker in stroke than in CAD. When LDL-C was modelled as a binary exposure, our results are consistent with mainstream studies. For instance, a meta-analysis study of statin therapy on stroke prevention shows similar results as we observed here (Supp Fig.1), in which statin therapy results in a relative risk (RR) of 0.84[46]. Similar RR can also be found in several other meta-analysis studies[47,48].

We further assess whether the treatment effects are indeed *heterogeneous* using our proposed permutation-based test. We find that modifying LDL-C impose heterogeneous effects on CHD, while we did not find evidence for such heterogeneity in stroke (Table 6). Few RCT studies also support our findings. For example, pravastatin was reported to show

heterogenous efficacy in reducing coronary events[49]. Another interesting support is that Oscar et al. identified the heterogeneity in cost-effectiveness ratios after adjusting for absolute risk[50], emphasizing the importance of a tailored statin prescription policy for CAD prevention in the population. A lipid-lowering therapy (LLT) meta-analysis showed that the relative risk reduction of stroke by statins is significant but not heterogeneous[51], which is consistent with our findings that modifying LDL will lead to a homogeneous treatment effect on morbidity of stroke.

3.2.2.2 Features that contribute to the heterogeneity of LDL on CHD and stroke

In addition to identifying the heterogeneity in the effect of LDL on CHD and stroke, we are interested in which covariates contribute more to the heterogeneity. Fig.6A shows the top 10 important features identified via the DRIV model of LDL on CHD and stroke. Our results show that patients with higher testosterone levels may exhibit a higher treatment effect due to increase in LDL; such results may also indicate that males are more vulnerable to CHD caused by an increased LDL level. We found that the SHAP value of testosterone starts increasing at around 10 nmol/L, which is the lower bound of normal testosterone range in males (Fig.7A, 7B Testosterone). The result is also consistent with several findings showing the difference between males and females in response to the high concentration of LDL. For instance, Noda et al. found an association between higher concentrations of LDL-C and the increased risk of mortality from CHD in males with low cholesterol, while such association cannot be observed in females[52]. Another Mendelian Randomization study also pointed out that the causal effect of LDL-C on CVD risk was stronger in males[53].

Triglycerides were found to be the second important variable of the CHD DRIV model. Our results show that people with higher triglycerides and BMI may be less vulnerable to the increased risk of CHD caused by LDL increase. We observe a peak in triglycerides and BMI plots (Fig.7A, 7B Triglycerides, BMI). In other words, the causal effect of LDL on CHD will first gradually increase with rising triglycerides (TG) or BMI (when TG/BMI are on the low side), and after reaching the peak, the effect will gradually decrease with the increase of triglycerides or BMI. The findings suggest that LDL may exert a larger causal effect when the person does not have other metabolic risk factors (e.g. when TG/BMI are on the low side), but the effects is smaller in the presence of other metabolic risk factors.

Another interesting finding is that phosphate acts as the top clinical variable of the model. Phosphate is an electrically charged particle containing mineral phosphorus, and an abnormal phosphate level is usually a sign of kidney disease or several other diseases. Other reports show that serum phosphate is a risk factor for cardiovascular events[54]. Similar to BMI and triglycerides, our studies show that the harmful effect caused by LDL increase may be mitigated in patients with higher serum phosphate as they already have a high baseline CHD risk.

In our analysis on stroke (Fig. 6A Stroke), testosterone is also the top clinical variable, indicating its unique role in estimating the individual treatment effect of LDL on CHD and stroke. Like CHD, our results show that people with higher testosterone levels are associated with a higher LDL treatment effect. In addition to testosterone, we show that smoking is

another critical variable. Smoking is associated with a stronger treatment effect by LDL on stroke (Fig.7C, 7D Non-smoker). Hozawa et al. show that smoking is a stronger risk factor for CHD in people with high LDL-C[55], and considering the close association between CHD and stroke, this may further explain why we observe smoking as a top variable in the model. Furthermore, a higher LDL treatment effect is observed for people with higher systolic blood pressure (Fig 7C, 7D systolic blood pressure). The result is consistent with the finding that cardiovascular event risk was significantly higher among patients with higher baseline SBP and LDL-C[56]. Another study also shows that LDL-C reduction is associated with the reduction in stroke events for subjects with a lower 6-month systolic blood pressure level[57], indicating that the treatment effect of LDL on stroke risk is lower in patients with a lower level of SBP. This is in line with our current findings.

3.2.3 HDL

3.2.3.1 HDL shows heterogeneous protective effects on CHD and homogeneous effects on stroke

HDL is another important lipoprotein circulating in the blood, which can transport blood cholesterol back to the liver and thus help decrease the accumulation of atherosclerosis[58]. It is widely accepted that HDL cholesterol has an important role as a predictor of CHD and stroke risk[59,60], while whether the effect is causal or heterogeneous remains not well established. Therefore, we applied our framework in studying the causal effect of HDL on CHD and stroke.

We first assess the causal relationship between HDL and CHD, and our results show that most people have negative treatment effects (i.e. increase in HDL is a causal protective factor). Interestingly, we observe that a rather small group of participants have positive or no treatment effects (as detected by DRIV), suggesting that increasing HDL in these people may not reduce CHD risk (Fig.3A). However, the group of subjects is small in our analysis and further replication in larger samples is required.

The role of HDL on CHD is not entirely clear yet, and varying effects have been reported in different epidemiology studies. For example, a meta-analysis on fibrates did not find a significant effect of fibrates on cardiac mortality, despite the fact that fibrates can effectively increase the HDL-C levels[61]. However, another placebo-controlled study on gemfibrozil shows that raising HDL-C levels by 11% through gemfibrozil could decrease the incidence of CHD after 5 years[62]. In addition to fibrates, the randomized control trial of Torcetrapib, a CETP inhibitor helping raise HDL-C concentrations, was suspended due to the increased coronary heart disease events[63]. The discrepancy in results across studies is unclear, but one possible explanation is that HDL may have heterogeneous effects on different individuals.

Notably, the current ITE analysis indeed showed significant heterogeneity for the effects of HDL (Table 6, Perm-Risk test < 0.05).

One of the drawbacks of our approach is that we tried to linearize the response curve and estimate a constant derivative as the treatment effect. However, some studies identified a non-linear U-shape effect of HDL-C on CHD risk, meaning that people with extremely high HDL-C and low HDL-C may exhibit an increased CHD risk[64]. In this case, a linear effect may

mask the true causal relationship between HDL and CHD. We ran a non-linear MR applying a localized estimation approach to test possible non-linear effects. We assigned people with desirable HDL concentrations (1.6-2.4 mmol/L) as a reference. We ran our framework in a binary setting with people having high-level HDL (>2.4 mmol/L) and low-level HDL (<1.5 mmol/L) as the treatment group, respectively. We found that both increase of HDL to very high levels or decreasing HDL to low levels may causally increase the risk of CHD (Supp Fig.3). However, we note the limited size of subjects with very high HDL and the results

Unlike CHD, our studies show that increasing HDL can decrease the risk of stroke, though we do not observe significant heterogeneity in treatment effects (Perm-Risk P-value > 0.05). Our results are consistent with several prospective cohort studies focusing on the relationship between HDL and stroke incidence. For example, Yoshiyuki et al. found that a lower HDL-C level is significantly related to increased risk of stroke[65]. Another multi-ethnic study also confirmed the relationship between low HDL-C and increased incidence of stroke[66].

3.2.3.2 Features that contribute to the heterogeneity of HDL on CHD and stroke

Fig.6B shows the top 10 important covariates (covariates that contribute most to the effect heterogeneity) in the HDL on CHD model. LDL-C is one of the model's most important variables, indicating a strong interaction between HDL and LDL in CHD development. The scatter plot shows that high LDL-C patients have positive SHAP values, meaning that high LDL-C may reduce the protective effect of HDL on CHD (Fig.8A,8B LDL-C). The result implies that for controlling CHD risks, it may be preferable to lower LDL-C and raise HDL-C simultaneously. For example, Brown et al. show that the combination of simvastatin and niacin can significantly reduce LDL-C and elevate the HDL-C level, which introduces a plausible protective effect against CHD[67]. An interesting finding is that the Townsend deprivation index is one of the most important modifiers of HDL's effect on CHD, and our results show that a high deprivation level is associated with a lower protective HDL treatment effect on CHD. Furthermore, we also identified lipoprotein A as the second most important variable for the CHD model, and the pattern of lipoprotein A is consistent with the LDL-C (Fig.8A, 8B Lipoprotein A), which is reasonable as a close relationship exists between LDL and lipoprotein A. Another observation is that Vitamin D and sex hormone binding globulin (SHBG) are two of the top 5 important clinical variables. The results are consistent with the finding that SHBG is positively correlated with the HDL, indicating a similar effect of SHBG on CHD risk as HDL[68,69].

For the stroke model, besides lipoprotein A and Townsend index, we identified IGF-1 as another top feature, and high level of IGF-1 can strength the protective effect of HDL on stroke (Fig.8C, 8D IGF-1). It is known that the low-level IGF-1 is associated with atherosclerosis, one of the causes of stroke[70]. There is also report showing that IGF-1 acts as an independent modulator for HDL-C, which is consistent with our finding that IGF-1 is one of the top covariables in the model[71]. We further notice that aspartate aminotransferase (AST) also contributes to ITE. Our results shows that the high level of AST will impair the treatment effect of HDL (SHAP value > 0 in high AST patients), which is consistent with the finding that level of AST is associated with increased risk of stroke[72] (Fig.8C, 8D Aspartate aminotransferase). Furthermore, an elevated AST is often found in non-alcoholic fatty liver disease (NAFLD) patients[73], and the severity of NAFLD is usually considered as an independent risk factor of

stroke[74]. These results may explain why we observe the AST in the top clinical variables of the model.

3.2.4 Triglycerides

3.2.4.1 Increased triglycerides (TG) is causally associated with CHD risk and the effect is heterogeneous

Dozens of studies pointed out that the imbalance of triglycerides metabolism is associated with different diseases, including heart disease and diabetes[75-77]. Our study mainly focuses on the causal relationship between triglycerides and coronary heart disease and stroke. We show that the unit increase of triglycerides is associated with around a 3.5% absolute risk increase in CHD and 3% in stroke (Fig.5). Several studies support the results. For example, the REDUCE-IT study observed a 4.8% absolute between-group difference in cardiovascular events between patients treated with icosapent ethyl (a drug that lowers TG) and placebo[78]. The VA-HIT study shows that gemfibrozil, a triglyceride-modifying drug, is associated with a 4.4% absolute risk reduction in CHD death[79]. For both CHD and stroke, the CF approach that was not based on instruments failed to detect the treatment effect. Similar to HDL, we only observe significant treatment effect heterogeneity in CHD but not in stroke.

The effect of TG on CHD showed significant heterogeneity (perm_risk p-value<0.05; Table 6).

3.2.4.2 Features that contribute to the heterogeneity of triglycerides on CHD and stroke

Following the same procedure as LDL and HDL, we assess the variable importance of the model via the SHAP model. Our results for the CHD causal model show that testosterone is the model's most important effect modifier. We observe that patients with high testosterone have a negative SHAP value (Fig.9A, 9B), meaning that the risk conferred by high TG is lower in those with high testosterone. We further identified BMI and waist-hip-ratio (WHR) as the second and third most important effect modifiers. Our results show that elevated triglycerides have a larger harmful effect on people with obesity (Fig.9A, 9B). Another interesting finding is that direct bilirubin was also one of the top effect modifiers. Some studies show that elevated direct bilirubin may be related to reduced HRs of CHD[80]. Our results are in line with the above findings as we found that the treatment effect of elevated TG is smaller in people with higher direct bilirubin.

For the stroke causal model, we observe that HbA1c is the most important effect modifier. Interestingly, our results show that people with mid-range HbA1c will be affected less by the elevated TG, which is consistent with other studies pointing out that high and extremely low HbA1c levels are associated with a high risk of stroke[81] (Fig 9C/9D). Like CHD, we also identified lipoprotein A as the top effect modifier within the model. The result indicates that increasing TG may lower stroke morbidity in people with higher lipoprotein A. As lipoprotein A is also a causal risk factor for stroke, we postulate that higher lipoprotein A may dilute the effect of high TG on stroke and CHD. Another interesting finding is that the treatment effect of elevated TG is smaller in the elderly population, which is also reflected by the SHAP analysis results. In line with this finding, a higher level of triglycerides in middle-aged stroke patients compared with elders[82].

3.2.5 Total Cholesterol

3.2.5.1 Total Cholesterol increment homogeneously increases the risk of CHD and stroke

In addition to two lipoproteins, we are also interested in whether total cholesterol is a causal risk factor for coronary heart disease and stroke. High cholesterol within the blood can lead to atherosclerosis, which may increase the risk of CHD and total stroke[83]. Based on our ITE analysis, nearly all participants show an increased risk of CHD with the increment of total cholesterol (Fig.4A).

Similarly, we observe that the individualized treatment effects of total cholesterol on stroke are consistent with that on CHD, which may indicate that elevated total cholesterol plays a similar role in increasing the risk of CHD and stroke (Fig.4B). Our results are also consistent with several meta-analyses and prospective studies[84,85]. Furthermore, we did not detect heterogeneity in treatment effects, which indicates that the total cholesterol's effect on CHD and stroke is consistent across the population.

3.2.5.2 Variable Importance

We assess the top variables using SHAP as above (Fig.6D). In our CHD model, we also observe that testosterone is the top variable that contributes to treatment effect, similar to triglycerides. The results may indicate a close relationship between the model for total cholesterol and triglycerides, and people with higher testosterone may have smaller treatment effects. Considering the difference in testosterone levels between males and females, our results also indicate that males will have lower treatment effects than females for both TG and TC (Fig.10A, 10B Testosterone). Another observation is that creatinine plays an essential role in the CHD model, in which the people with higher creatinine show smaller treatment effects. Similar to TG, direct bilirubin is observed in the top variables. An interesting finding is that the Townsend deprivation index acts as the top covariates; the results show that the treatment effect of TC in people with a higher degree of deprivation is smaller than that in rich subjects. We postulate that this may be associated with the close relationship between deprivation and health inequality.

For stroke, urea is the most important variable contributing to the treatment effect. We find that people with a normal range (2.1-8.5 mmol/L) of urea nitrogen are more resistant to the harmful effect caused by elevated total cholesterol (Fig.10C, 10D). Another interesting finding is that vitamin D is the second top variable of the model, and a higher level of Vitamin D may help mitigate the harmful effect of elevated TC on stroke (Fig.10C, 10D Vitamin D).

4 Discussion

4.1 Overview

In this study, we extend the regular Mendelian randomization approach to infer the *individualized* causal effects of risk factors/treatments in observational studies. Classic MR only focuses on inferring the average causal effect[86], which may not be sufficient in the current age of precision medicine. Although the estimated average effect is still meaningful in designing policy/treatment for the population, it can mask the individual's response. To overcome the drawback of classic MR, we propose a novel framework incorporating Mendelian Randomization and machine learning methodologies to estimate the individualized treatment effect. We also presented two permutation-based tests to test the presence of effect heterogeneity in our framework. Through a simulation study, we demonstrate the applicability of our proposed MR framework under the scenario with an unobserved confounder. We also evaluate the performance of our proposed heterogeneity testing approaches via an extra simulation.

As a proof-of-concept example, we applied our framework to study the individualized causal effects of several lipid traits on CHD and stroke, including LDL, HDL, Triglycerides, and Total cholesterol. We indeed found evidence of heterogeneity especially for CHD. Through Shapley value plots, we also uncovered key clinical features which may modify the effects of lipids on CHD/stroke. This application illustrates the usefulness of our approach and may have important clinical implications. For example, subjects who may suffer from a larger effects of dyslipidaemia on CHD/stroke may benefit more intensive treatment programs, e.g. through diet/exercise and medications. For health policy planning, as medical resources are usually limited, targeting treatment to those who are predicted to show more marked adverse effects from risk factors may enable the most efficient use of resources. In addition, this study also revealed possible clinical factors that contribute to the heterogeneity of treatment effects.

4.2 Strengths and Limitations

There are several strengths of our study. To our best knowledge, this is the first study aiming to estimate individualized causal treatment effects leveraging principles of MR. Although it is possible for researchers to study ITE under an RCT setting, the difficulty and high cost in designing and implementing an RCT often makes such designs impractical. Our approach provides an alternative to infer the individualized treatment effect using genetic instruments, which is much less susceptible to unknown confounders and reverse causality. The ITE estimation strategies are also based on machine learning (ML) methods, which allows flexible modelling of complex relationships. In addition, the DRIV approach also allows virtually any ML models to be used, further enhancing the flexibility and applicability of our approach.

Our simulation results show that our proposed methods of testing heterogeneity performed well in most cases. Finally, incorporating SHAP analysis within the framework can help identify the top variables contributing to ITE, allowing more detailed model explanations and may help to subgroup/subtype patients in practice.

Our study also has several limitations. For example, our simulation results show that the permutation variance test does not perform well in some complicated cases; for example, there exists strong non-linear treatment effects or the interaction between different types of confounders. Another limitation is that we lack an external dataset to validate our results as it is relatively difficult to find a large-sample and phenotype-rich dataset like the UK-Biobank. Nevertheless, the estimated ITE are reasonable and the range is close to estimates from previous RCTs with lipid-lowering agents. Furthermore, we mainly consider a linear effect of exposure in our framework, while non-linear causal effects may be present in practice[87]. In addition, the power to detect significant ITEs will depend on the sample size; for example, the number of subjects with stroke is considerably smaller than the number of CHD cases, which may limit the power to detect significant heterogeneity.

4.3 Conclusions

In summary, we have designed a novel MR framework capable of estimating the individualized causal effect under observational study settings. We estimated the ITEs of lipid-trait on CHD and stroke, and revealed important clinical features that contribute to the effect heterogeneity using Shapley value analyses. We hope our work will open a new direction of research for MR studies, and provide a new way to identify ITEs that will ultimately be translated to clinical practice to design more personalized treatment plans for patients.

References

1 Reynolds, R. *et al.* A systematic review of chronic disease management interventions in primary care. *BMC Family Practice* **19**, 11, doi:10.1186/s12875-017-0692-3 (2018).

2 Kent, D. M., Steyerberg, E. & van Klaveren, D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Bmj* **363** (2018).

3 Sussman, J. B., Kent, D. M., Nelson, J. P. & Hayward, R. A. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. *Bmj* **350** (2015).

4 Akobeng, A. K. Understanding randomised controlled trials. *Archives of disease in childhood* **90**, 840-844 (2005).

5 Edwards, S. *et al.* Ethical issues in the design and conduct of randomised controlled trials. *Health Technology Assessment* **2**, 1-132 (1999).

6 Roessner, V. Large sample size in child and adolescent psychiatric research: the way of salvation? *European Child & Adolescent Psychiatry* **23**, 1003-1004, doi:10.1007/s00787-014-0635-7 (2014).

7 Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133-1163, doi:https://doi.org/10.1002/sim.3034 (2008).

8 Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26**, 2333-2355, doi:10.1177/0962280215597579 (2015).

9 Wald, A. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics* **11**, 284-300 (1940).

10 Imbens, G. W. & Angrist, J. D. Identification and Estimation of Local Average Treatment Effects. *Econometrica* **62**, 467-475, doi:10.2307/2951620 (1994).

11 Burgess, S. & Thompson, S. G. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* **42**, 1134-1144, doi:10.1093/ije/dyt093 (2013).

12 Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299-319, doi:10.1093/biomet/asaa076 (2021).

13 Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications* **11**, 376, doi:10.1038/s41467-019-14156-4 (2020).

14 Neyman, J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10**, 1-51 (1923).

15 Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**, 688 (1974).

16 Li, Y. *et al.* in *Proceedings of the 2020 KDD Workshop on Causal Discovery*. 4-22 (PMLR).

17 Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *Annals of Statistics* **47**, 1148-1178 (2019).

18 Tibshirani, R. & Hastie, T. Local Likelihood Estimation. *Journal of the American Statistical Association* **82**, 559-567, doi:10.2307/2289465 (1987).

19 Syrgkanis, V. *et al.* Machine learning estimation of heterogeneous treatment effects with instruments. *arXiv preprint arXiv:1905.10176* (2019).

20 Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1-C68, doi:10.1111/ectj.12097 (2018).

21 Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99**, 139-153, doi:https://doi.org/10.1016/j.ajhg.2016.05.013 (2016).

22 Zhu, Z. *et al.* Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics* **96**, 377-385, doi:https://doi.org/10.1016/j.ajhg.2015.01.001 (2015).
23 Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581-590, doi:10.1038/s41576-018-0018-x (2018).
24 Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, doi:10.1093/gigascience/giz082 (2019).
25 Breiman, L. Stacked regressions. *Machine learning* **24**, 49-64 (1996).
26 Varadhan, R. & Seeger, J. D. in *Developing a protocol for observational comparative effectiveness research: A user's guide* (Agency for Healthcare Research and Quality (US), 2013).
27 Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* **90**, 389-405 (2008).
28 Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353, doi:10.1073/pnas.1510489113 (2016).
29 Zhao, K. *Research on Machine Learning for Drug Discovery and Precision Medicine*, Hong Kong University of Science and Technology (Hong Kong), (2020).
30 Robinson, P. M. Root-N-Consistent Semiparametric Regression. *Econometrica* **56**, 931-954, doi:10.2307/1912705 (1988).
31 O'Neill, E. & Weeks, M. Causal Tree Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes. *arXiv preprint arXiv:1810.09179* (2018).
32 Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
33 Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and regression trees*. (CRC press, 1984).
34 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).
35 Lundberg, S. M. & Lee, S.-I. in *Advances in neural information processing systems.* 4765-4774.
36 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2**, 2522-5839 (2020).
37 Powers, S. *et al.* Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine* **37**, 1767-1787, doi:https://doi.org/10.1002/sim.7623 (2018).
38 Marees, A. T. *et al.* A tutorial on conducting genome‐wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* **27**, e1608 (2018).
39 Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15**, 2759-2772, doi:10.1038/s41596-020-0353-1 (2020).
40 Ho, D. E., Imai, K., King, G. & Stuart, E. A. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* **15**, 199-236, doi:10.1093/pan/mpl013 (2007).
41 Staiger, D. & Stock, J. H. Instrumental Variables Regression with Weak Instruments. *Econometrica* **65**, 557-586, doi:10.2307/2171753 (1997).
42 Li, M. *et al.* Statins for the Primary Prevention of Coronary Heart Disease. *Biomed Res Int* **2019**, 4870350-4870350, doi:10.1155/2019/4870350 (2019).
43 Toth, P. P. & Palmer, M. K. High risk US adults with untreated hyperlipidaemia in NHANES: estimated reduction in cardiovascular events with statin treatment. *European Heart Journal* **41**, ehaa946.2992, doi:10.1093/ehjci/ehaa946.2992 (2020).

44 Collaboration, P. S. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *The Lancet* **370**, 1829-1839 (2007).
45 Assessment, R. Major lipids, apolipoproteins, and risk of vascular disease. *Jama* **302**, 1993-2000 (2009).
46 O'Regan, C., Wu, P., Arora, P., Perri, D. & Mills, E. J. Statin therapy in stroke prevention: a meta-analysis involving 121,000 patients. *The American journal of medicine* **121**, 24-33 (2008).
47 Wang, W. & Zhang, B. Statins for the prevention of stroke: a meta-analysis of randomized controlled trials. *PLoS One* **9**, e92388 (2014).
48 Ní Chróinín, D. *et al.* Statin therapy and outcome after ischemic stroke: systematic review and meta-analysis of observational studies and randomized trials. *Stroke* **44**, 448-456 (2013).
49 Sacks, F. M. *et al.* The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine* **335**, 1001-1009 (1996).
50 Franco, O. H., Peeters, A., Looman, C. W. N. & Bonneux, L. Cost effectiveness of statins in coronary heart disease. *Journal of Epidemiology and Community Health* **59**, 927-933, doi:10.1136/jech.2005.034900 (2005).
51 Corvol, J.-C. *et al.* Differential effects of lipid-lowering therapies on stroke prevention: a meta-analysis of randomized trials. *Archives of internal medicine* **163**, 669-676 (2003).
52 Noda, H. *et al.* Gender difference of association between LDL cholesterol concentrations and mortality from coronary heart disease amongst Japanese: the Ibaraki Prefectural Health Study. *Journal of internal medicine* **267**, 576-587 (2010).
53 Cupido, A. J., Asselbergs, F. W., Schmidt, A. F. & Hovingh, G. K. Low-Density Lipoprotein Cholesterol Attributable Cardiovascular Disease Risk Is Sex Specific. *Journal of the American Heart Association*, e024248 (2022).
54 McGovern, A. P. *et al.* Serum phosphate as a risk factor for cardiovascular events in people with and without chronic kidney disease: a large community based cohort study. *PLos one* **8**, e74996 (2013).
55 Hozawa, A., Folsom, A. R., Sharrett, A. R., Payne, T. J. & Chambless, L. E. Does the impact of smoking on coronary heart disease differ by low-density lipoprotein cholesterol level?: the Atherosclerosis Risk in Communities (ARIC) Study. *Circ J* **70**, 1105-1110, doi:10.1253/circj.70.1105 (2006).
56 Deedwania, P. C. *et al.* Differing predictive relationships between baseline LDL-C, systolic blood pressure, and cardiovascular outcomes. *Int J Cardiol* **222**, 548-556, doi:10.1016/j.ijcard.2016.07.201 (2016).
57 Amarenco, P. *et al.* Relative and cumulative effects of lipid and blood pressure control in the Stroke Prevention by Aggressive Reduction in Cholesterol Levels trial. *Stroke* **40**, 2486-2492, doi:10.1161/strokeaha.108.546135 (2009).
58 März, W. *et al.* HDL cholesterol: reappraisal of its clinical relevance. *Clinical Research in Cardiology* **106**, 663-675 (2017).
59 Després, J.-P., Lemieux, I., Dagenais, G.-R., Cantin, B. & Lamarche, B. t. HDL-cholesterol as a marker of coronary heart disease risk: the Québec cardiovascular study. *Atherosclerosis* **153**, 263-272, doi:https://doi.org/10.1016/S0021-9150(00)00603-1 (2000).
60 Rader, D. J. & Hovingh, G. K. HDL and cardiovascular disease. *The Lancet* **384**, 618-625, doi:https://doi.org/10.1016/S0140-6736(14)61217-4 (2014).
61 Saha, S. A., Kizhakepunnur, L. G., Bahekar, A. & Arora, R. R. The role of fibrates in the prevention of cardiovascular disease—a pooled meta-analysis of long-term randomized placebo-controlled clinical trials. *American heart journal* **154**, 943-953 (2007).
62 Frick, M. H. *et al.* Helsinki Heart Study: primary-prevention trial with gemfibrozil in middle-aged men with dyslipidemia. *New England Journal of Medicine* **317**, 1237-1245 (1987).

63 Barter, P. J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. *New England journal of medicine* **357**, 2109-2122 (2007).
64 Feng, M. *et al.* Free cholesterol transfer to high-density lipoprotein (HDL) upon triglyceride lipolysis underlies the U-shape relationship between HDL-cholesterol and cardiovascular disease. *European journal of preventive cardiology* **27**, 1606-1616 (2020).
65 Soyama, Y. *et al.* High-density lipoprotein cholesterol and risk of stroke in Japanese men and women: the Oyabe Study. *Stroke* **34**, 863-868 (2003).
66 Polak, J. F., Pencina, M. J., O'Leary, D. H. & D'Agostino, R. B. Common carotid artery intima-media thickness progression as a predictor of stroke in multi-ethnic study of atherosclerosis. *Stroke* **42**, 3017-3021 (2011).
67 Brown, B. G. *et al.* Simvastatin and niacin, antioxidant vitamins, or the combination for the prevention of coronary disease. *New England Journal of Medicine* **345**, 1583-1592 (2001).
68 Pugeat, M. *et al.* Interrelations between sex hormone-binding globulin (SHBG), plasma lipoproteins and cardiovascular risk. *J Steroid Biochem Mol Biol* **53**, 567-572, doi:10.1016/0960-0760(95)00102-6 (1995).
69 Semmens, J., Rouse, I., Beilin, L. J. & Masarei, J. R. Relationship of plasma HDL-cholesterol to testosterone, estradiol, and sex-hormone-binding globulin levels in men and women. *Metabolism* **32**, 428-432 (1983).
70 Burchardt, P., Goździcka-Józefiak, A. & Siminiak, T. Review article IGF-1–a new risk factor for coronary atherosclerosis. *Kardiologia Polska (Polish Heart Journal)* **64**, 1297-1302 (2006).
71 Succurro, E. *et al.* Positive association between plasma IGF1 and high-density lipoprotein cholesterol levels in adult nondiabetic subjects. *European Journal of Endocrinology* **163**, 75-80, doi:10.1530/EJE-10-0113 (2010).
72 Ruban, A. *et al.* Liver Enzymes and Risk of Stroke: The Atherosclerosis Risk in Communities (ARIC) Study. *J Stroke* **22**, 357-368, doi:10.5853/jos.2020.00290 (2020).
73 Angulo, P. Nonalcoholic fatty liver disease. *N Engl J Med* **346**, 1221-1231, doi:10.1056/NEJMra011775 (2002).
74 Xu, J. *et al.* Severity of Nonalcoholic Fatty Liver Disease and Risk of Future Ischemic Stroke Events. *Stroke* **52**, 103-110, doi:10.1161/strokeaha.120.030433 (2021).
75 Laufs, U., Parhofer, K. G., Ginsberg, H. N. & Hegele, R. A. Clinical review on triglycerides. *European heart journal* **41**, 99-109c (2020).
76 Ginsberg, H. N., Zhang, Y.-L. & Hernandez-Ono, A. Regulation of plasma triglycerides in insulin resistance and diabetes. *Archives of medical research* **36**, 232-240 (2005).
77 Harchaoui, K., Visser, M., Kastelein, J., Stroes, E. & Dallinga-Thie, G. Triglycerides and cardiovascular risk. *Current cardiology reviews* **5**, 216-222 (2009).
78 Bhatt, D. L. *et al.* Cardiovascular risk reduction with icosapent ethyl for hypertriglyceridemia. *New England Journal of Medicine* **380**, 11-22 (2019).
79 Watts, G. F. Treating low HDL-cholesterol in normocholesterolaemic patients with coronary disease: statins, fibrates or horses for courses? *European Heart Journal* **25**, 716-719, doi:10.1016/j.ehj.2003.12.024 (2004).
80 Zhang, R. *et al.* Elevated serum bilirubin may significantly reduce coronary heart disease risk in females: A prospective cohort study. *Nutrition, Metabolism and Cardiovascular Diseases* **32**, 648-657, doi:https://doi.org/10.1016/j.numecd.2021.12.015 (2022).
81 Shen, Y. *et al.* Association between Hemoglobin A1c and Stroke Risk in Patients with Type 2 Diabetes. *J Stroke* **22**, 87-98, doi:10.5853/jos.2019.01704 (2020).
82 Seok, S. & Lee, J. Comparison of stroke risk factors between middle-aged and elderly patients. *Journal of Korean Gerontological Nursing* **19**, 13-27 (2017).
83 Peters, S. A., Singhateh, Y., Mackay, D., Huxley, R. R. & Woodward, M. Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. *Atherosclerosis* **248**, 123-131 (2016).

84 Sacco, R. L. *et al.* High-density lipoprotein cholesterol and ischemic stroke in the elderly: the Northern Manhattan Stroke Study. *Jama* **285**, 2729-2735 (2001).
85 Collaboration, P. S. Cholesterol, diastolic blood pressure, and stroke: 13 000 strokes in 450 000 people in 45 prospective cohorts. *The Lancet* **346**, 1647-1653 (1995).
86 Dixon, P., Hollingworth, W., Harrison, S., Davies, N. M. & Davey Smith, G. Mendelian Randomization analysis of the causal effect of adiposity on hospital costs. *J Health Econ* **70**, 102300-102300, doi:10.1016/j.jhealeco.2020.102300 (2020).
87 Staley, J. R. & Burgess, S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genetic Epidemiology* **41**, 341-352, doi:https://doi.org/10.1002/gepi.22041 (2017).

Table 1: Heterogeneity Testing Methods simulation results

Pleiotropy Scenario	Perm-Variance Test			Perm-τ-risk Test		
	1	2	3	1	2	3
Scenario 1	0	0	0	0.02	0.04	0.02
Scenario 2	1	1	1	1	1	1
Scenario 3	1	1	1	1	1	1
Scenario 4	1	1	1	1	1	1
Scenario 5	1	1	1	1	1	1
Scenario 6	0.58	0.76	0.64	1	1	1
Scenario 7	0	0.02	0.02	0.88	0.84	0.92
Scenario 8	1	1	1	1	1	1

Updated (Keep the mean and variance the same across scenarios)

Pleiotropy Scenario	Perm-Variance Test			Perm-τ-risk Test		
	1	2	3	1	2	3
Scenario 1	0	0	0	0.02	0.02	0.02
Scenario 2	1	1	1	1	1	1
Scenario 3	1	1	1	1	1	1
Scenario 4	0.98	0.98	1	1	1	1
Scenario 5	0.74	0.84	1	1	1	1
Scenario 6	0.26	0.30	0.64	0.98	1	1
Scenario 7	0	0.02	0.02	0.96	0.92	0.92
Scenario 8	0.18	0.36	1	1	0.98	1

Legend:

The table shows the simulation results of two heterogeneity test methods, the permutation-variance test, and the permutation-τ-risk test. Scenario 1 is the homogeneous treatment effect scenario, and we use type 1 error to describe whether the test will incorrectly detect the heterogeneity. The rest scenarios are heterogeneous treatment effect scenarios, and the power will be used to denote the performance of the test in detecting heterogeneity.

Table 2: Covariates included in Table

Traits	Shared Discrete Variables	Shared Continuous Variables	Lipid related variables
LDL-C	Gender, Diabetes diagnosed by doctor, British, Indian, Caribbean, African, Non-alcohol drinker, Previous alcohol drinker, Current alcohol drinker, Non-smoker, Previous smoker, Current smoker, Cholesterol_lowering_medication, Blood_pressure_medication	PC1-10, Townsend Deprivation Index, diastolic blood pressure, systolic blood pressure, BMI, waist-hip-ratio, Age, Creatinine, C-reactive protein, Cystatin C, Gamma glutamyl transferase, Glucose, HbA1c, Aspartate aminotransferase, Direct bilirubin, Urea, IGF-1, Phosphate, SHBG, Testosterone, Total protein, Urate, Vitamin D	HDL-C, Apolipoprotein A, Triglycerides
HDL-C			LDL-C, Apolipoprotein B, Lipoprotein A, Triglycerides,
Triglycerides			LDL-C, HDL-C, Apolipoprotein A, Apolipoprotein B, Lipoprotein A, Total Cholesterol
Total Cholesterol			Triglycerides

Table 3: Baseline characteristics of participants for continuous clinical covariates.

	Coronary Heart Disease (N = 22158)		Stroke (N = 10946)	
	Mean	Standard Deviation	Mean	Standard Deviation
Age	59.23	7.34	61.33	6.59
BMI	28.32	4.82	28.10	4.73
Waist-Hip-Ratio	0.91	0.09	0.91	0.09
diastolic blood pressure (mmHg)	84.35	10.73	83.74	10.92
systolic blood pressure (mmHg)	145.55	19.74	146.03	20.26
Triglycerides (mmol/L)	1.86	1.02	1.78	0.96
Total Cholesterol (mmol/L)	5.51	1.04	5.39	1.07
LDL-C (mmol/L)	3.50	0.81	3.378	0.82
HDL-C (mmol/L)	1.34	0.34	1.36	0.36
Apolipoprotein A (g/L)	1.48	0.25	1.49	0.26
Apolipoprotein B (g/L)	1.03	0.22	0.99	0.22
Creatinine (umol/L)	77.25	20.85	77.20	20.52
C-reactive protein (mg/L)	2.92	4.95	2.92	4.89
Cystatin C (mg/L)	0.96	0.20	0.98	0.22
Gamma glutamyltransferase (U/L)	42.46	44.68	43.33	52.78
Glucose (mmol/L)	5.26	1.47	5.32	1.55
HbA1c (mmol/mol)	37.18	8.00	37.76	8.28
Aspartate aminotransferase (U/L)	27.38	10.84	27.29	10.59
Direct bilirubin (umol/L)	1.88	0.85	1.90	0.95
Urea (mmol/L)	5.60	1.49	5.70	1.59
IGF-1 (nmol/L)	21.22	5.60	20.74	5.57
Phosphate (mmol/L)	1.14	0.16	1.14	0.16
SHBG (nmol/L)	46.13	22.78	48.60	23.80
Testosterone (nmol/L)	8.52	5.85	8.22	6.01
Total Protein (g/L)	72.52	4.10	72.29	4.07
Urate (umol/L)	336.10	80.18	332.72	81.17
Vitamin D (nmol/L)	49.22	21.12	49.66	20.91

Townsend Deprivation Index	-1.37	3.05	-1.32	3.06

Table 4: Baseline characteristics of participants for discrete dummy clinical covariates.

	Coronary Heart Disease (N = 22158)		Stroke (N = 10946)	
	0	1	0	1
Gender	6849 Female	15309 Male	3784 Female	7162 Male
Diabetes diagnosed by doctor	20120	2038	9761	1185
British	780	21378	333	10613
Indian	21698	460	10786	160
Caribbean	21972	186	10848	98
African	22024	134	10871	75
Non-alcohol drinker	21328	830	10492	454
Previous alcohol drinker	21282	876	10412	534
Current alcohol drinker	1706	20452	988	9958
Non-smoker	11374	10784	5683	5263
Previous smoker	13421	8737	6551	4395
Current smoker	19521	2637	9658	1288
Cholesterol_lowering_medication	16716	5442	7515	3431
Blood_pressure_medication	18670	3488	9227	1719

Table 5: Partial F-test to assess the weak instrument.

	Coronary Heart Disease		Stroke	
	Simple F-test	Heteroscedasticity-consistent F-test	Simple F-test	Heteroscedasticity-consistent F-test
LDL-C	1238.147	1233.461	440.202	446.6602
HDL-C	279.7468	274.749	528.3735	523.6446
Triglycerides	110.4805	106.8085	38.79924	37.7728
Total Cholesterol	717.6225	727.1178	347.7115	340.0772

Table 6: Permutation-based test to assess the presence of heterogeneity (Continuous Trait).

	Coronary Heart Disease		Stroke	
	Perm-Var test	Perm-Risk test	Perm-Var test	Perm-Risk test
LDL-C	**0.02**	**0.05**	0.9	0.7
HDL-C	0.12	**0.04**	0.9	0.4
Triglycerides	0.54	**0.02**	0.25	0.3
Total Cholesterol	1	0.6	0.4	0.6

Figure 1.

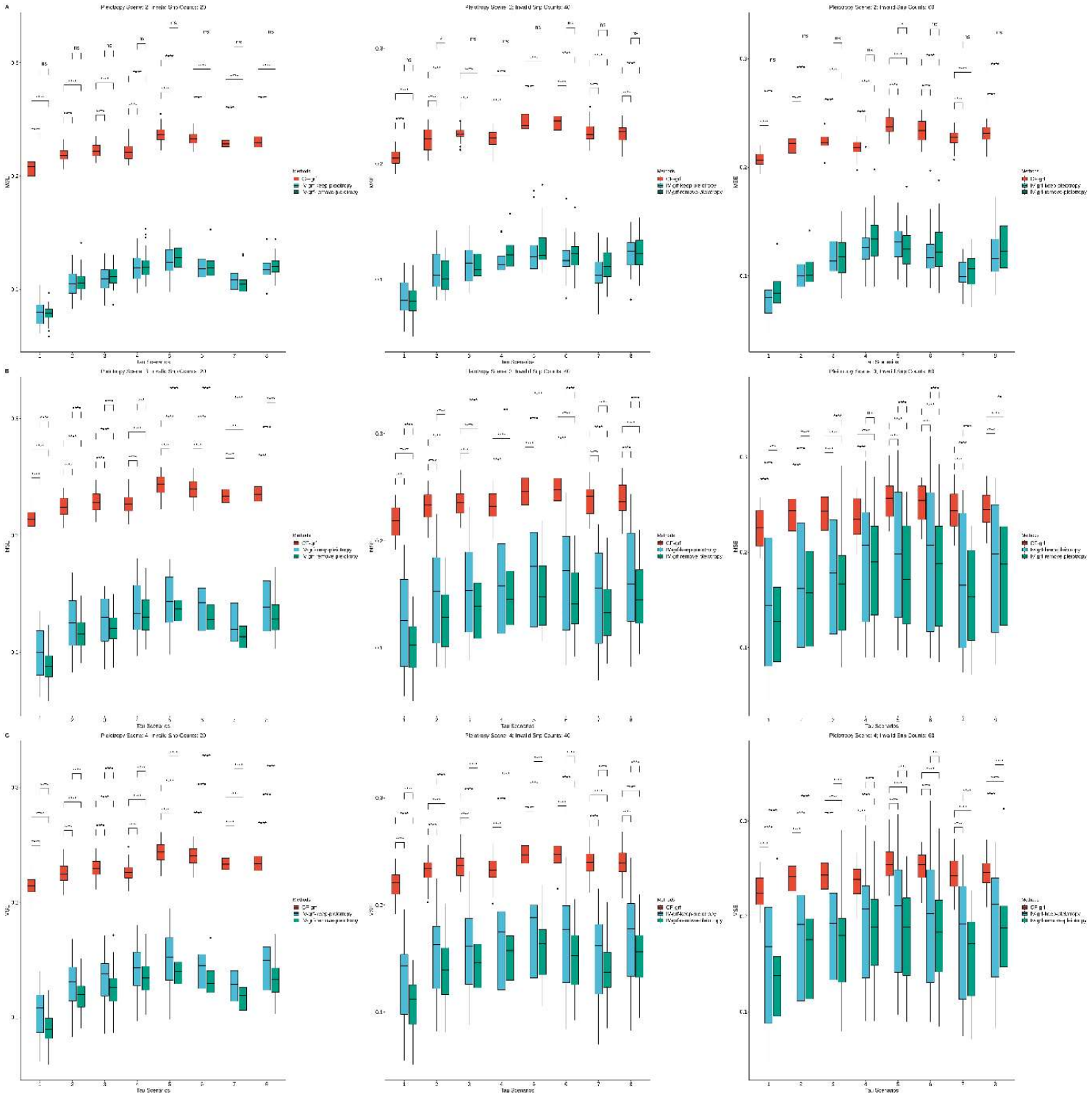


Figure 2 Predicted Treatment Effect of LDL on CHD and Stroke (Continuous trait).

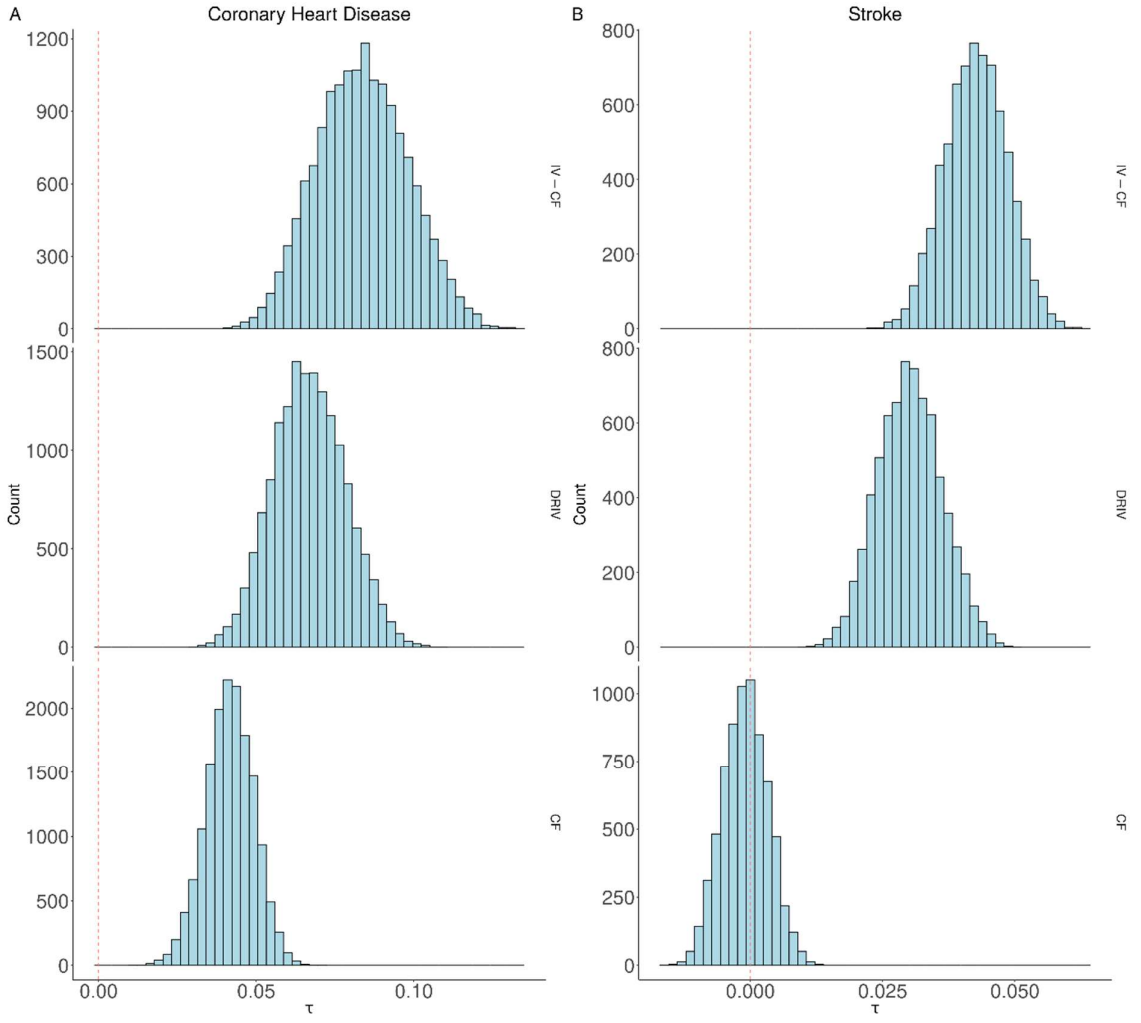


Figure 3 Predicted Treatment Effect of HDL on CHD and Stroke (Continuous trait).

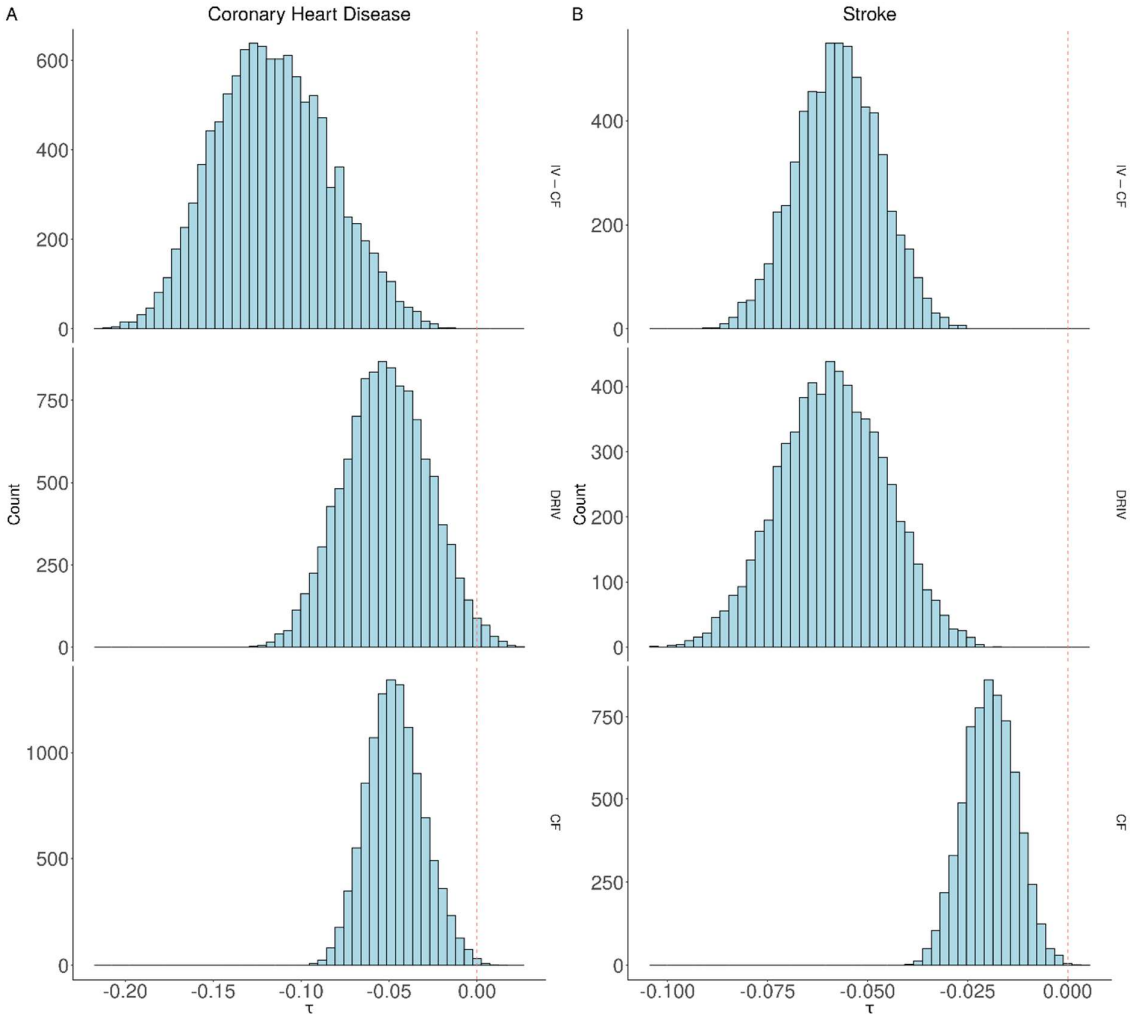


Figure 4 Predicted Treatment Effect of TC on CHD and Stroke (Continuous trait).

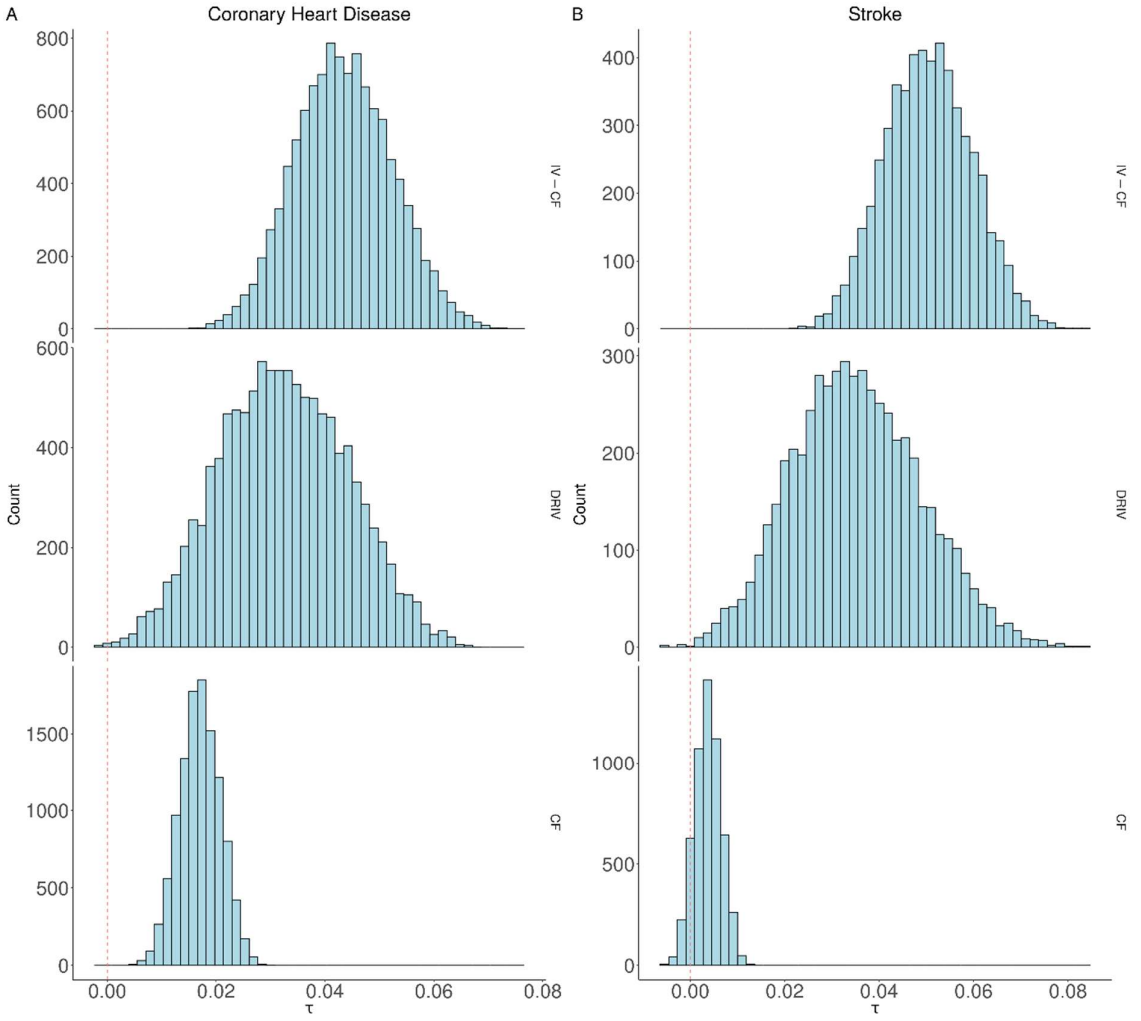


Figure 5 Predicted Treatment Effect of Triglycerides on CHD and Stroke (Continuous trait).

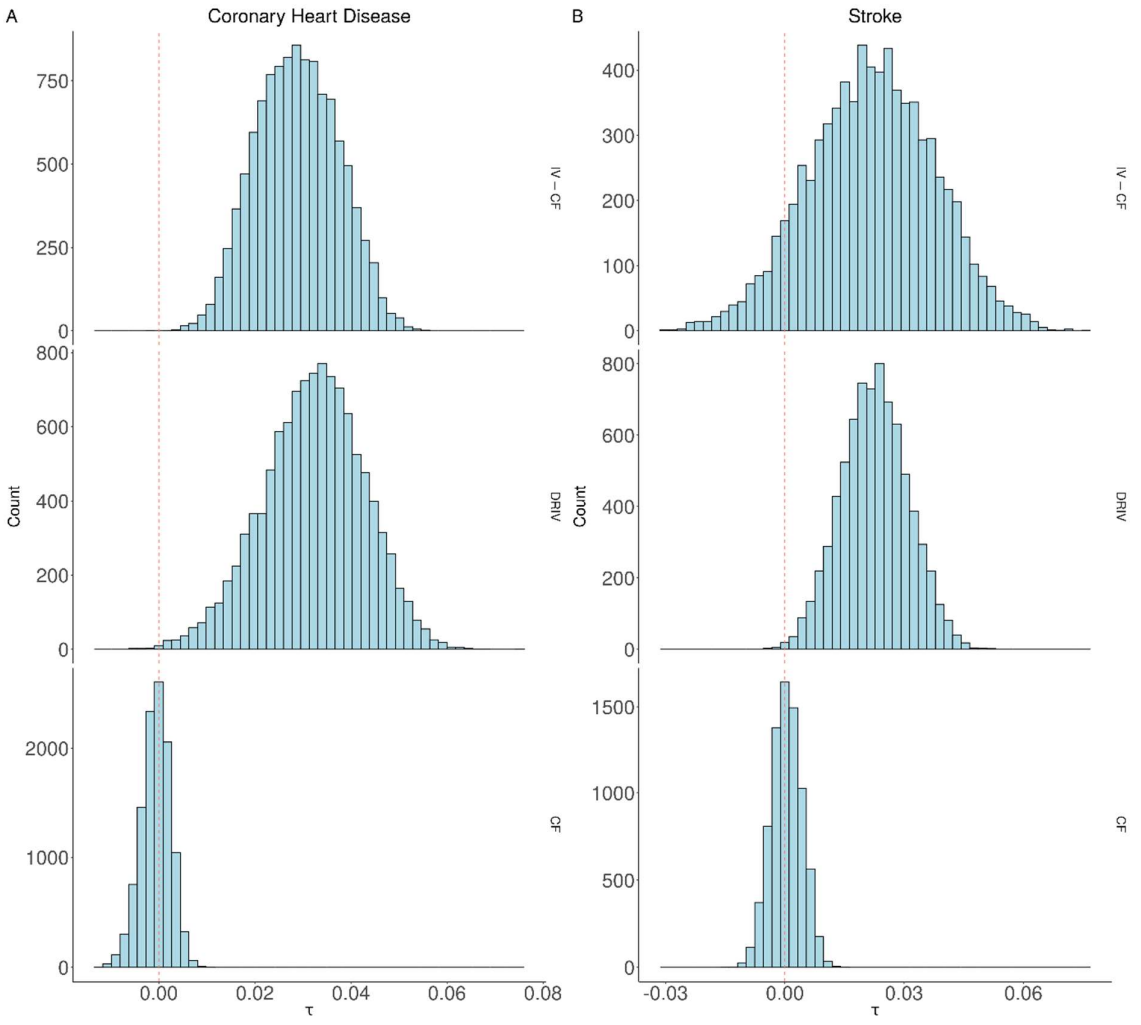
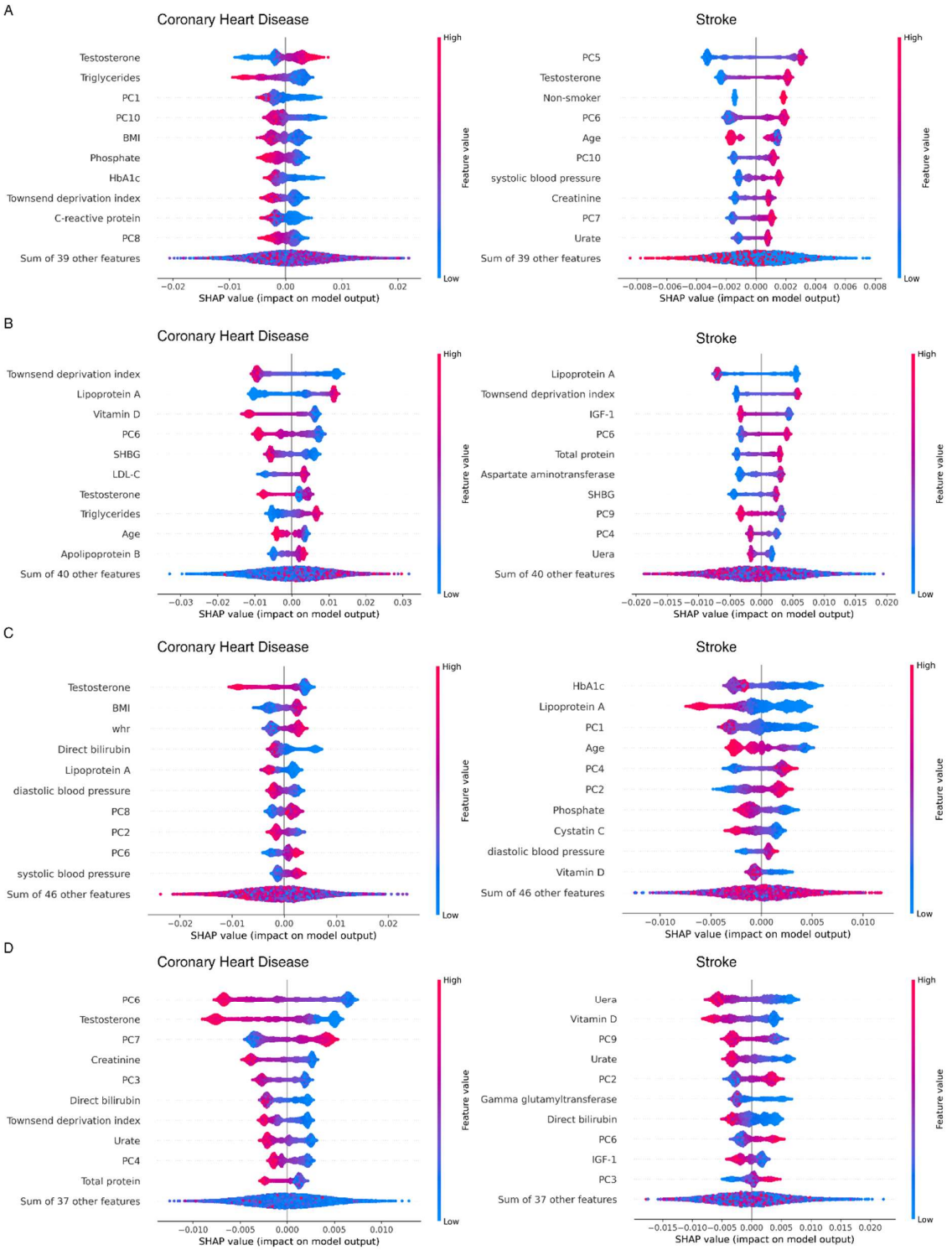


Figure 6 Overall Variable Importance of CHD A: LDL, B: HDL, C: Triglycerides D:Total Cholesterol



(In the Shapley value plots below, A and B refers to CHD as outcome; C and D refers to stroke as outcome. In plots A and C, the Shapley value of the variable was shown; in plots B and D, the estimated ITE (tau) was shown)

Fig 7 Shapley value plot with LDL-C as risk factor

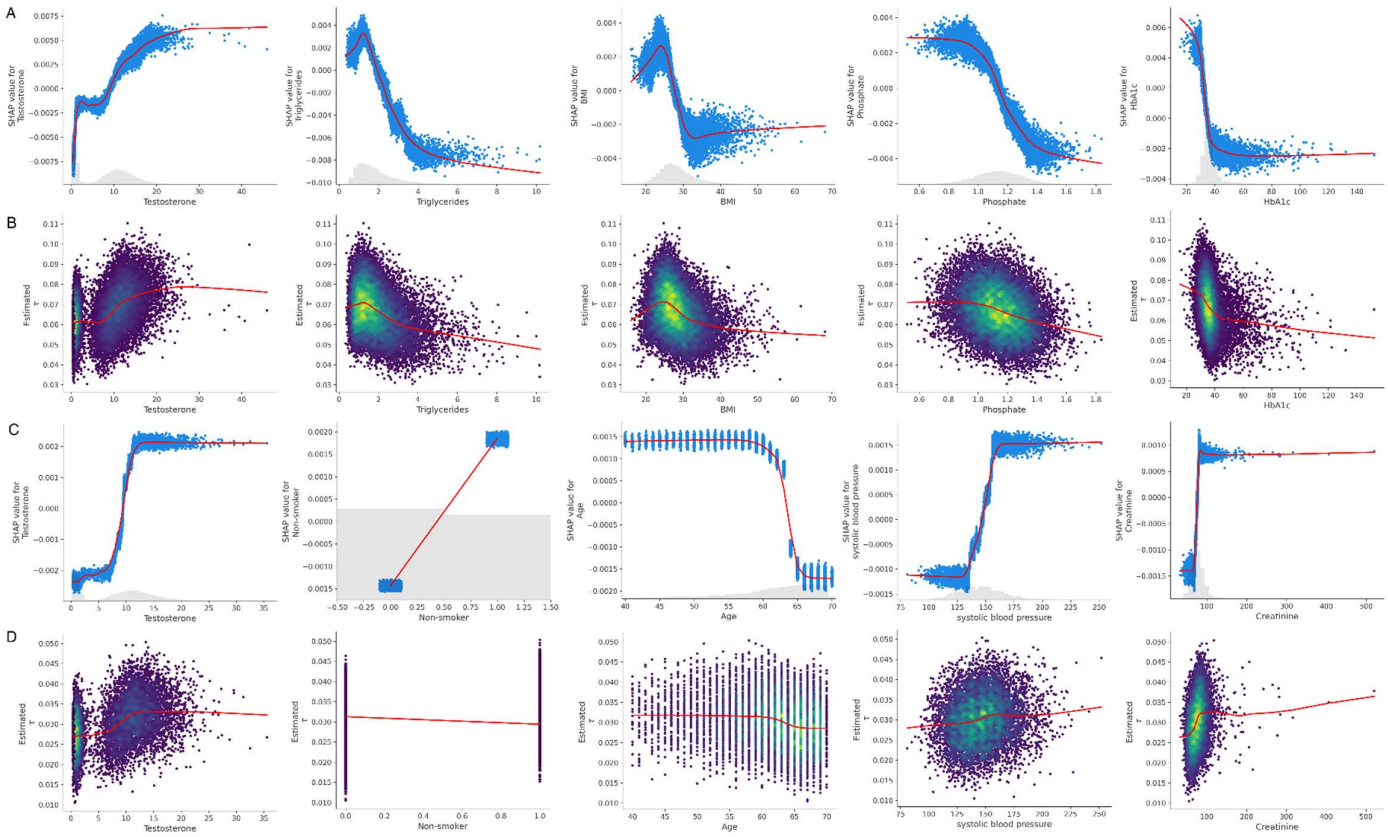


Fig 8 HDL variable importance

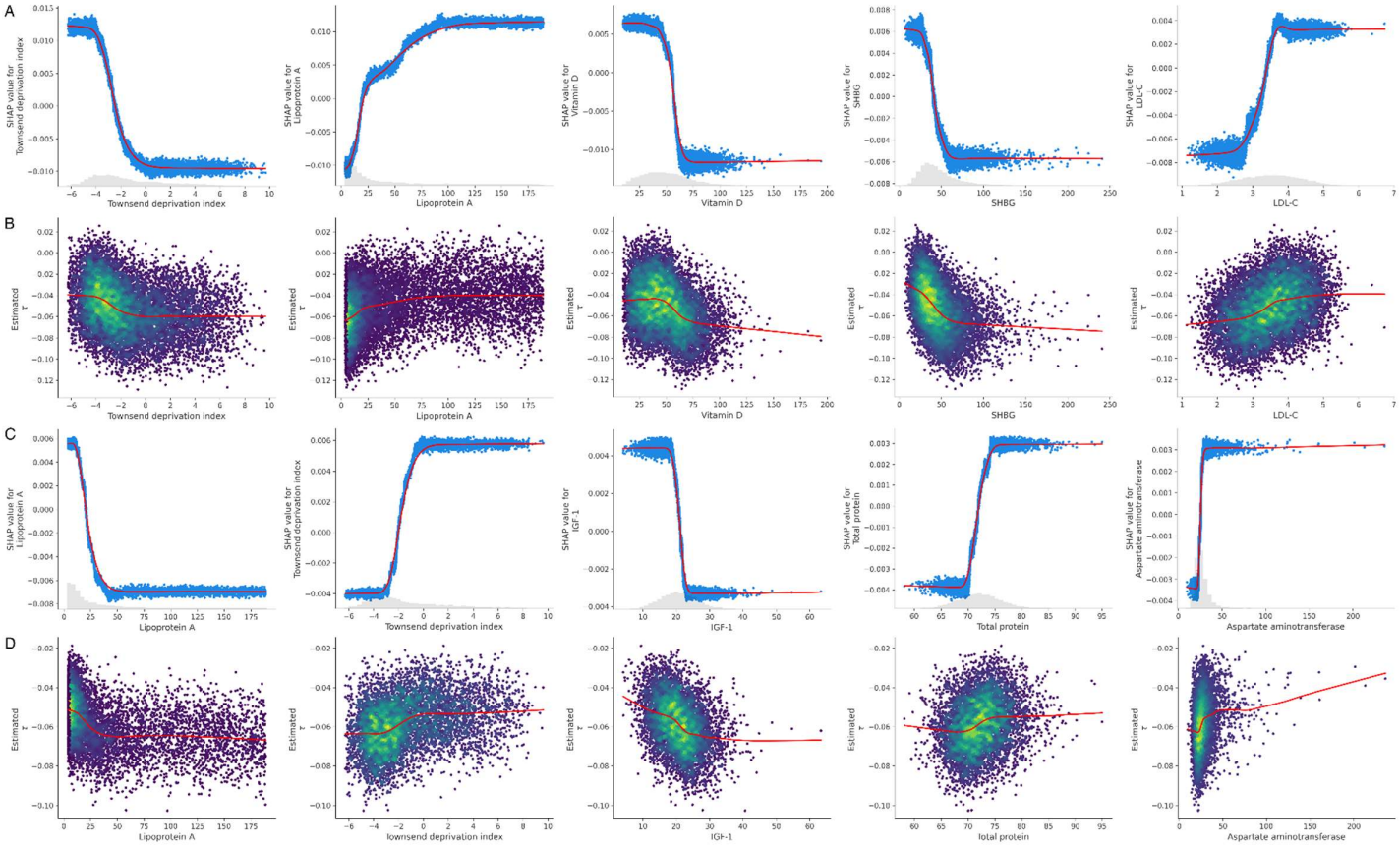


Fig 9: Triglycerides Variable Importance

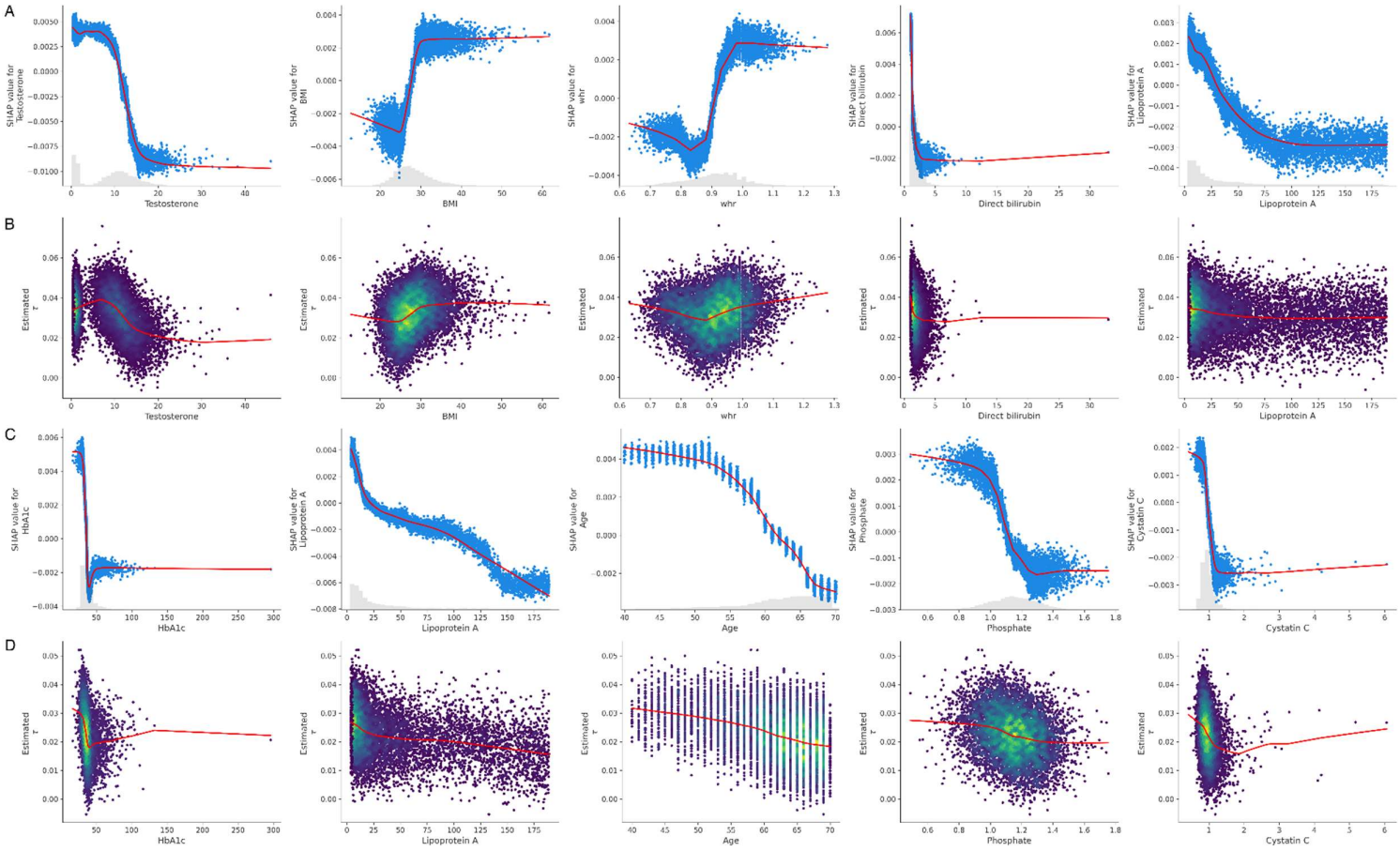
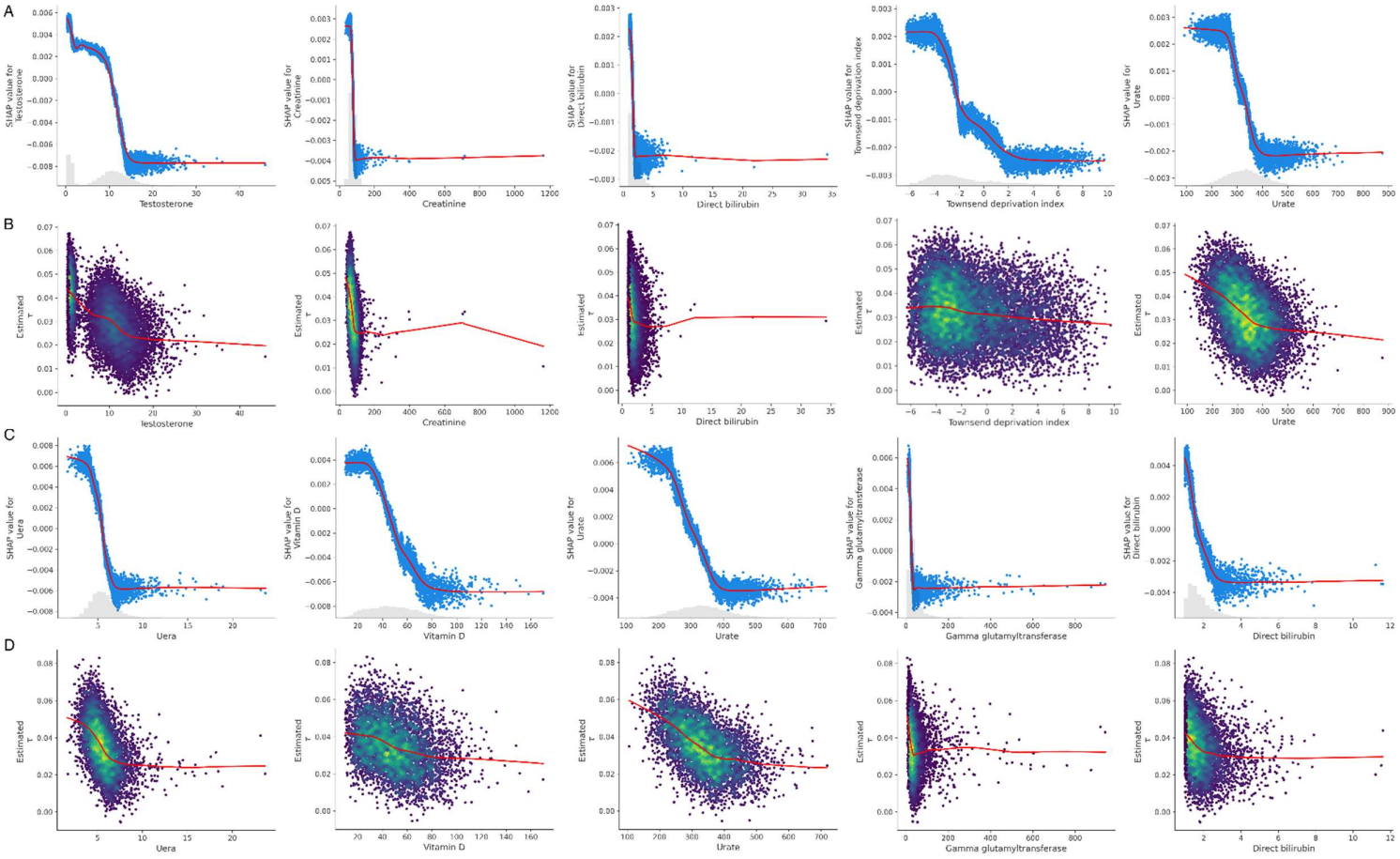
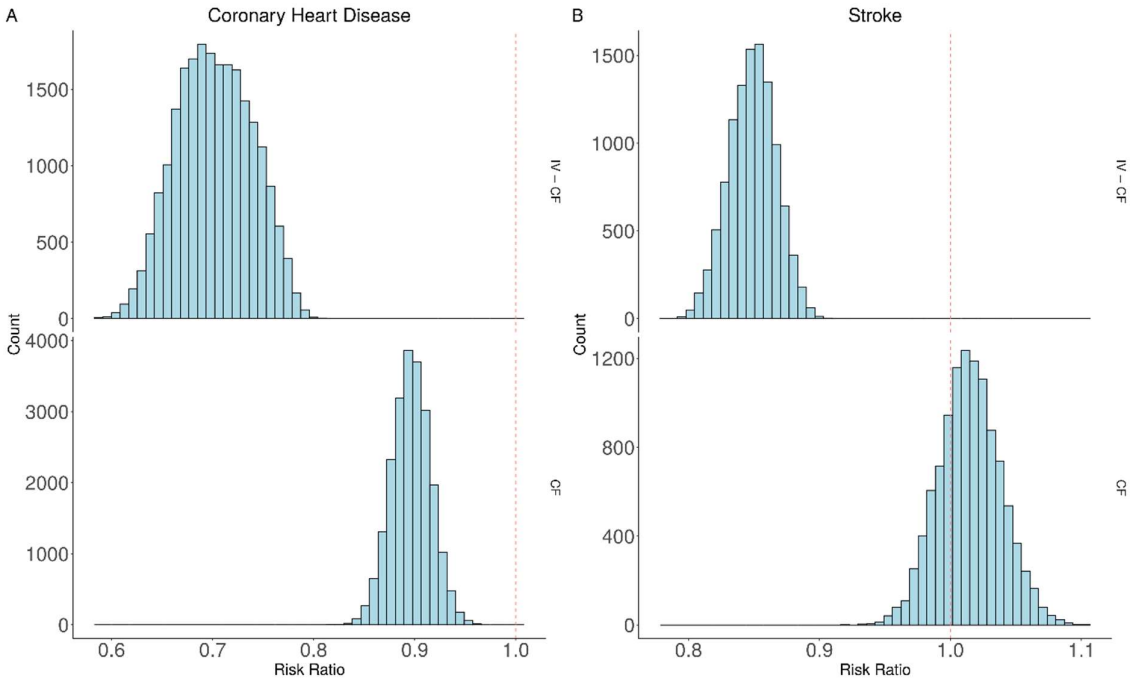


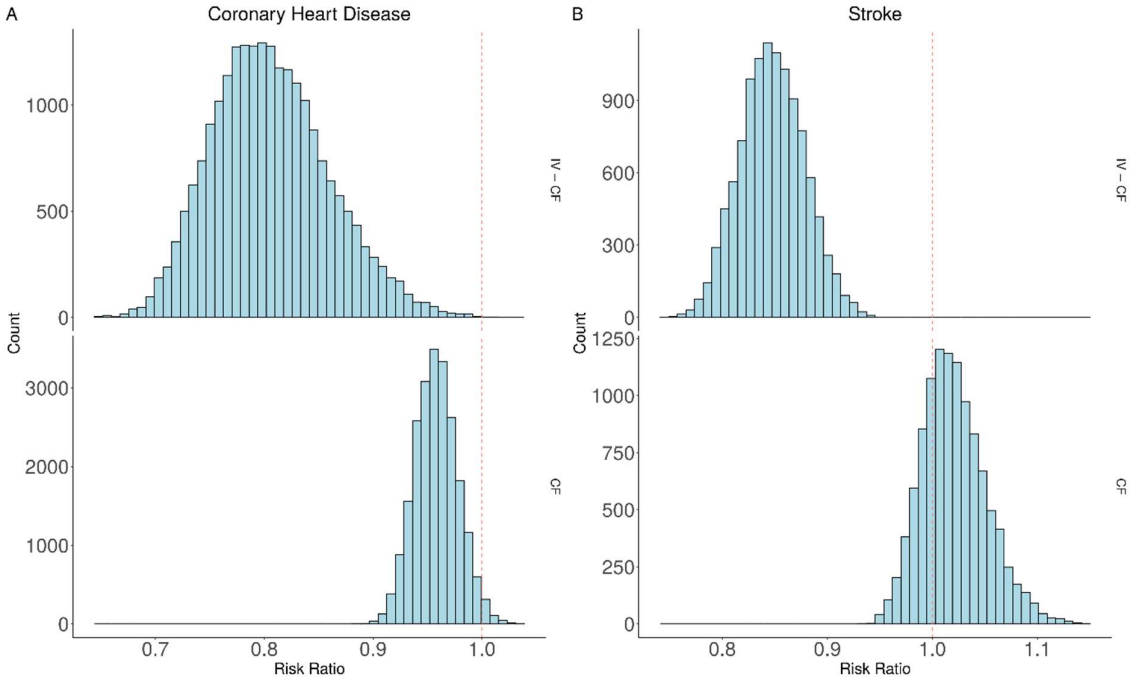
Fig 10: Total Cholesterol Variable Importance



Supp Fig 1: LDL



Supp Fig 2: HDL



Supp Fig 3