A framework for detecting causal effects of risk factors at an individual level based on principles of Mendelian randomisation: applications to modelling individualised effects of lipids on coronary artery disease

Yujia Shi,[a] Yong Xiang,[a] Yuxin Ye,[a] Tingwei He,[a] Pak-Chung Sham,[h] and Hon-Cheong So[a,b,c,d,e,f,g,*]

[a]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, China
[b]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and the Chinese University of Hong Kong, China
[c]Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong, China
[d]CUHK Shenzhen Research Institute, Shenzhen, China
[e]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong, China
[f]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China
[g]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
[h]Department of Psychiatry, University of Hong Kong, Hong Kong SAR, China

eBioMedicine
2025;113: 105616
Published Online 28 February 2025
https://doi.org/10.1016/j.ebiom.2025.105616

Summary

Background Mendelian Randomisation (MR) has been widely used to study the causal effects of risk factors. However, almost all MR studies concentrate on the population's average causal effects. With the advent of precision medicine, the individualised treatment effect (ITE) is often of greater interest. For instance, certain risk factors may pose a higher risk to some individuals than others, and the benefits of treatments may vary across individuals. This study proposes a framework for estimating *individualised* causal effects in large-scale observational studies where unobserved confounding factors may be present.

Methods We propose a framework (MR-ITE) that expands the scope of MR from estimating average causal effects to individualised causal effects. We present several approaches for estimating ITEs within this MR framework, primarily grounded on the principles of the "R-learner". To evaluate the presence of causal effect heterogeneity, we also proposed two permutation testing methods. We employed polygenic risk score (PRS) as instruments and proposed methods to improve the accuracy of ITE estimates by removal of potentially pleiotropic single nucleotide polymorphisms (SNPs). The validity of our approach was substantiated through comprehensive simulations. The proposed framework also allows the identification of important effect modifiers contributing to individualised differences in treatment effects. We applied our framework to study the individualised causal effects of various lipid traits, including low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC), on the risk of coronary artery disease (CAD) based on the UK-Biobank (UKBB). We also studied the ITE of C-reactive protein (CRP) and insulin-like growth factor 1 (IGF-1) on CAD as secondary analyses.

Findings Simulation studies demonstrated that MR-ITE outperformed traditional causal forest approaches in identifying ITEs when unobserved confounders were present. The integration of the contamination mixture (ConMix) approach to remove invalid pleiotropic SNPs further enhanced MR-ITE's performance. In real-world applications, we identified positive causal associations between CAD and several factors (LDL-C, Total Cholesterol, and IGF-1 levels). Our permutation tests revealed significant heterogeneity in these causal associations across individuals. Using Shapley value analysis, we identified the top effect modifiers contributing to this heterogeneity.

Interpretation We introduced a new framework, MR-ITE, capable of inferring individualised causal effects in observational studies based on the MR approach, utilizing PRS as instruments. MR-ITE extends the application of MR from estimating the average treatment effect to individualised treatment effects. Our real-world application of MR-ITE underscores the importance of identifying ITEs in the context of precision medicine.

*Corresponding author. School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, China.
E-mail address: hcso@cuhk.edu.hk (H.-C. So).

Funding This work was supported partially by a National Natural Science Foundation of China grant (NSFC; grant number 81971706), the KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China, and the Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong.

Keywords: Mendelian randomisation; Individualised treatment effect; Causal inference; Heterogeneity; Coronary artery disease

Research in context

Evidence before this study

Traditionally, epidemiologists have primarily focused on studying the average causal effect of interventions in populations, often overlooking the significance of population heterogeneity. However, mounting evidence suggests that treatment effects often vary among individuals, with the same intervention yielding different benefits across subpopulations. For example, metformin has been shown to have varying effects on diabetes mellitus (DM) among patients with impaired glucose metabolism. These findings underscore the importance of estimating individualised treatment effects (ITEs) for advancing precision medicine. There were limited methods available to infer ITEs in large-scale observational studies that account for both observed and unobserved confounding factors. In addition, no previous studies have integrated ITE estimation with Mendelian randomisation (MR) principles with real-world clinical applications.

Added value of this study

We developed MR-ITE, a framework that leverages Mendelian Randomisation with polygenic risk score (PRS) as instruments to infer individualised causal effects in observational studies. We also proposed employing a contamination mixture approach to exclude potentially pleiotropic single nucleotide polymorphisms (SNPs) before PRS calculation. We developed two permutation tests to rigorously evaluate the presence of effect heterogeneity within MR-ITE. Additionally, we proposed the use of Shapley value (SHapley Additive exPlanations) analysis to identify key effect modifiers contributing to this heterogeneity. These innovations make MR-ITE a powerful tool for uncovering treatment effect heterogeneity and the underlying mechanisms. The proposed approach may also have the potential to inform the design of personalized combination therapies to optimize clinical outcomes.

Implications of all the available evidence

This study presents a framework for estimating individualised treatment effects leveraging MR, offering a new avenue for exploring the causal influence of specific risk factors on disease outcomes in large-scale observational datasets. Crucially, this approach accommodates both observed and unobserved confounders, advancing the limitation of existing ITE estimation approaches. By facilitating a more precise understanding of treatment effect variability, MR-ITE could significantly contribute to the development of personalized therapies, ultimately improving patient outcomes.

Introduction

The rising incidence and mortality rates of chronic diseases have imposed a significant burden on numerous countries over the past decades.[1] Consequently, identifying potential causal risk factors and designing appropriate interventions have emerged as top priorities. In the past, epidemiologists focused primarily on studying the *average* causal effect of interventions in the population, thereby overlooking the importance of population heterogeneity. The presence of heterogeneity suggests that individuals may derive varying benefits from the same intervention. For instance, a randomized controlled trial (RCT) demonstrated that metformin could have a heterogeneous impact on diabetes mellitus (DM) prevention among patients with impaired glucose metabolism[2,3]; patients at a higher risk of diabetes might experience a more substantial absolute risk reduction than those at lower risk. This study underscores the importance of estimating individualised treatment effects (ITEs). To optimize intervention efficiency across the population and minimize costs, it is important to estimate the potential benefit a specific patient may gain from an intervention (or risk factor prevention). In this study, we aimed to estimate the individualised causal treatment effect of a given intervention to individual patients, leveraging the principles of Mendelian randomisation (MR).

We wish to highlight that our work is different from conventional statistical/machine learning (ML) prediction models, which are focused on predicting a clinical phenotype/outcome based on covariates. Our work, on the other hand, is designed for estimating/predicting *the causal and individualised treatment effect*. Briefly, we ask the following question: how would the outcome change if a person receives the treatment (or exposed to a risk factor), vs. the case that the person does not receive it?

This is also known as a 'counterfactual' argument. Our work falls under the "causal ML" area, as opposed to conventional ML models, as reviewed elsewhere.[4]

It is widely accepted that the most accurate approach to estimate the causal effect is via an RCT, in which both known and unknown confounding factors can be controlled for by treatment randomisation.[5] However, RCTs are often prohibitively expensive, limited by ethical considerations or logistical constraints, and may lack generalizability due to strict inclusion/exclusion criteria.[6,7] Consequently, researchers frequently resort to observational studies to estimate causal effects. Unlike RCTs, a major concern with observational studies is that unmeasured confounding may influence causal inference. Mendelian Randomisation (MR) serves as a valuable approach to mitigate the risk of unmeasured confounding and is largely immune to reverse causality. In MR, genetic variants are utilized as instruments to represent the exposure.[8]

Following years of development and innovation, a variety of statistical methods have been established for MR analyses, including the Wald ratio method, two-stage least squares, MR-IVW, MR-Egger, weighted median etc.[9] 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, and there is a lack of innovations regarding the estimation of *individualised* treatment (or risk factor) effects.

Our main contribution is the introduction of a framework, MR-ITE, capable of inferring individualised causal effects in observational studies based on the MR approach, utilizing the polygenic risk score (PRS) as an instrument. We proposed several ITE estimation methodologies within the MR framework, grounded on the principles of "R-learners".[10] These methods offer high flexibility as they leverage supervised machine learning (ML) approaches for modelling, imposing virtually no restrictions on the type of ML models employed. Our other contributions to the MR-ITE framework include: (1) To mitigate the risks of bias from invalid instruments, we proposed the use of the contamination mixture approach to eliminate potential pleiotropic single nucleotide polymorphisms (SNPs) prior to calculating the PRS[11]; (2) we presented permutation-based approaches to test for the presence of heterogeneity under MR-ITE; (3) we proposed methods to identify important *effect modifiers* contributing to effect heterogeneity, for instance by employing Shapley values. The identification of potential biomarkers modulating the effect of exposures is important as it sheds light on the mechanisms underlying effect heterogeneity, and practically, it may contribute to the design of combination therapies to improve individualised treatment effects. (4) we applied our proposed framework to study the individualised (causal) effects of lipids on risks of coronary artery disease (CAD). Our findings indicate that low-density lipoprotein cholesterol (LDL-C) and total cholesterol (TC) may exert heterogeneous causal effects on CAD risks, and we also uncovered the major effect modifiers; In addition to lipid traits, as an additional analysis, we also studied the effects of two other proteins, namely C-reactive protein (CRP) and insulin-like growth factor 1 (IGF-1), on CAD. Both have also been reported to be associated with obesity[12–14] but are less polygenic than lipid traits. We identified heterogeneous causal effects of IGF-1 on coronary artery disease (CAD) risk, while no significant causal effect was found with CRP. (5) Furthermore, we also discussed and presented potential clinical implications, such as disease subtyping or subgrouping patients with divergent treatment responses.

To summarize, our study represents a pioneering effort to expand the scope of MR from estimating *average* causal effects to *individualised* causal effects.

Methods

Set-up and notation

Rubin's causal model

A causal model needs to be formalized first. A well-established and popular choice is the Neyman-Rubin causal model, also called the potential outcome (counterfactual) framework.[15] We consider a dataset with N units, indexed by $i = 1, ..., N$. 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 $(Y_i^{Status}, W_i, X_i, Z_i)$ 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 may be difficult to define an arbitrary cutoff to partition the population into treatment and control groups. In this case, the unit-level causal effect is defined as the effect of unit increment of treatment on the outcome, $\tau_i = Y_i^{W+\Delta W} - Y_i^{W}$.

Assumptions of instrumental variables

As we regard our approach as an extension of the MR framework, we also require similar assumptions that the MR framework needs to achieve a consistent estimation of the causal effect. Theoretically, the instrument must satisfy three distinct assumptions to be valid: the relevance assumption, the exclusion restriction assumption, and the independence assumption.

The relevance assumption necessitates a genuine association between the instrument and the exposure, which is the only assumption that can be easily directly tested.[16] We evaluate the strength of the instrument using F-statistics; in general, an F-statistic greater than 10 indicates that the instrument meets the relevance criterion.

The exclusion restriction assumption stipulates that the instrument is independent of the outcome given the exposure and possible confounders. The independence assumption requires that the instrument is not correlated with any factors that may confound the exposure–outcome relationship.[17,18] However, unlike the relevance assumption, these two assumptions are difficult to be fully verified.[16,19,20] Given the challenges in directly testing the exclusion restriction and independence assumptions, our approach involves meticulous selection of SNPs that are likely to comply with these criteria. We employ a contamination mixture model[11] (ConMix) to identify SNPs that are unlikely to breach these assumptions. Only SNPs retained by the ConMix approach are included in our PRS calculations. Of note, the ConMix model has been widely employed in MR studies, and as shown in simulations, it is able to tackle both horizontal and correlated pleiotropy (i.e., pleiotropy via confounders) well with low type I error inflation and good power.[11]

Estimating the individualised treatment effect

We provide two different methods to estimating the individualised treatment effect, including Generalized Random Forest (GRF) and Double Robustness Instrument Variable estimator (DRIV). The GRF, introduced by Athey et al., extends Breiman's random forests to estimate any quantity $\theta(x)$ through local moment conditions.[21] This flexibility allows the GRF to adapt to various scenarios, ranging from basic regression problems to complex causal inference studies. In this framework, we employ an instrumental causal forest (IV-GRF), an important application of the GRF. Unlike the standard causal forest, IV-GRF modifies the gradient-based labeling formula to $\rho_i = (Z_i - \overline{Z}_p)((Y_i - \overline{Y}_p) - (W_i - \overline{W}_p)\hat{\tau}_p)$, where $\overline{Y}_p$, $\overline{Z}_p$ and $\overline{W}_p$ stand for the average of Y, Z and W over the parent node P. IV-GRF then performs a standard CART regression split, aiming to maximize the heterogeneity of the in-sample θ-estimates, using the criterion:

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

The second methodology, the double robustness instrumental variable estimator (DRIV), innovates by designing a loss function that enables the use of general machine learning methods for minimization, rather than modifying existing algorithms.[22] The DRIV process is bifurcated into a preliminary individualised treatment effect (ITE) estimation step, termed double machine learning IV (DMLIV), followed by a doubly robust estimation step named DRIV. Detailed comparisons between these steps and additional implementation details are available in the supplementary notes and the original DRIV paper.

Polygenic risk score construction

Polygenic risk score (PRS) as instrument

In our framework, we utilize polygenic risk score (PRS) as the instrument to perform the individualised MR analysis. The PRS summarizes the estimated effects of multiple trait-associated genetic variants on an individual's phenotype, typically defined as a weighted sum of trait-associated risk alleles across multiple genetic loci.[23] PRS has been commonly employed as an instrumental variable (IV) for MR analyses.[24,25]

There are several reasons for our choice of PRS as an instrumental variable within our analytical framework. Firstly, the ITE estimation method employed in our framework, including DRIV and GRF, are originally optimized for scenarios involving a single instrument. On the other hand, using a single genetic variant as instrument in MR could suffer from low statistical power and susceptibility to weak instrument bias. One may however employ PRS, which aggregates the effects of multiple genetic variants, to serve as the instrument to improve power and instrument strength.[26]

Secondly, Burgess et al. has shown that MR analyses using PRS or summary statistics (inverse-variance weighted, IVW) methods in general produce very similar results.[25] In addition, the above study showed that estimates obtained from the summary statistics method with external weights tend to be biased toward the null, when those weights are imprecisely estimated (e.g. when GWAS (Genome-wide association study) sample size of the exposure trait is not large). In contrast, allele score (i.e., PRS) estimates remain unbiased. In our study, we employed external weights to derive PRS of exposures; as such, the use of allelic scores as instruments might enjoy the advantage of producing less biased causal effect estimates. Burgess et al. also showed that when using equal or external weights, both methods provide valid tests of the null hypothesis of no causal effect, even when there are many potentially weak instruments. In other words, the type I error rates are controlled using either method.

Thirdly, we may theoretically employ a single genetic variant as an instrument, perform MR-ITE analysis with DRIV or instrumental GRF approach each time, and combine the causal estimates from each variant in a second step (similar to the IVW approach). However, since both DRIV or instrumental GRF are machine learning models, this approach would necessitate fitting multiple models independently, which could be exceedingly time-consuming especially if a large number of variants is involved.

Given these considerations, we have chosen to use the PRS of the exposure as an instrumental variable in our framework. This decision is grounded in both the established validity of PRS as a robust instrument in genetic epidemiology and the practical benefits in our analysis. However, a significant concern with using an allelic score as an IV in MR is the potential violation of IV assumptions due to the inclusion of pleiotropic SNPs. To address this, we apply the contamination mixture (ConMix) approach to pre-identify and exclude potential pleiotropic SNPs, thereby ensuring the validity of the PRS as an instrument.

The contamination mixture (ConMix) approach

The ConMix method is based on a likelihood function tailored to variant-specific causal estimates. It classifies SNPs into two categories, each following a distinct normal distribution: valid SNPs are normally distributed around the true causal effect value, while invalid SNPs follow a normal distribution centred around zero with a larger standard deviation. This differentiation allows the likelihood function to incorporate a two-component mixture distribution for each variant. By maximizing this function, we can discern each genetic variant's contribution to the likelihood, and classify each variant as either 'valid' or 'invalid'. This facilitates the exclusion of invalid SNPs in subsequent PRS calculations. Detailed descriptions of the algorithm are available in the supplementary notes.

Methods for calculating PRS

With the validated SNPs, we employed two methods to calculate PRS for further analysis: PRSice-2 (as the primary approach) and LDPred2 (as an additional method). PRSice-2 is an efficient PRS calculation software that automates PRS analyses.[27] LDPred2, an advanced version of LDPred, is a Bayesian approach that estimates posterior mean causal effect sizes using GWAS summary statistics, assuming a prior for genetic architecture and leveraging linkage disequilibrium (LD) information from a reference panel.[28] LDPred2 addresses several limitations inherent in the original LDPred, offering an alternative to PRSice-2 for estimating PRS in our study. Please also refer to the supplementary text for further details.

Assessing the presence of treatment effect heterogeneity

In addition to estimating the individualised treatment effect (ITE), we introduce two permutation-based methods to evaluate heterogeneity among the estimated treatment effects. Heterogeneity of treatment effect typically refers to non-random, explainable variability in ITE.[29] Another perspective views heterogeneity as whether the predicted treatment effects deviate significantly from the average effect, beyond what would be expected by chance.[30] These conceptualizations guide the development of methods both to estimate ITE and to assess heterogeneity. For example, within each split, a causal tree aims to maximize the variance of the estimated treatment effect across its leaves while also penalizing the uncertainty of these estimates.[31] If covariates do not contribute to heterogeneity, the variance of the predicted ITE across leaves will be smaller compared to when splitting on covariates that do contribute to heterogeneity. This observation inspired the development of our heterogeneity testing methods based on covariate permutation.

We present two such permutation-based methods: the permutation-variance test and the permutation-τ-risk test, to determine if heterogeneity is statistically significant. While these principles are also applicable to standard Heterogeneity of Treatment Effects (HTE) models (ref[32], Chapter 4), our tests are specifically tailored for our framework involving instrumental variables, particularly for the modification in the τ-risk test. The primary distinction between these methods lies in the target function of interest we used to compare models trained on raw covariates against those trained on permuted covariates. Specifically, one method focuses on the variance of the estimated individualised treatment effect, while the other assesses the improvement in τ-risk. The τ-risk improvement is defined as follows:

$$\widehat{L}_{improve} = \frac{1}{n}\sum_{i}([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\bar{\tau}(X_i)](Z_i - E[Z_i|X_i]))^2 - \frac{1}{n}\sum_{i}([(Y_i - E[Y_i|X_i]) - (W_i - E[W_i|X_i])\hat{\tau}(X_i)](Z_i - E[Z_i|X_i]))^2$$

where $\bar{\tau}(X_i)$ represents the average treatment effect (i.e., assume no heterogeneity in treatment effects), and $\hat{\tau}\{X_i\}$ represents the individualised treatment effect. The details of the two permutation methods can be found in the supplement notes.

Measuring variable importance

In addition to identifying the ITE, we also wish to identify which covariate may contribute to the heterogeneity. In other words, we identify important effect modifiers that lead to differential treatment effects across individuals.

We mainly used two approaches to achieve the goal. The first approach is the split-frequency based approach, which is mainly designed for the generalized random forest algorithm since it calculates the variable importance based on the split frequencies. The other approach is the SHAP (SHapley Additive exPlanations) approach, which allows us to capture the features' contribution for each individual's ITE predictions or for the whole sample with considering the rest of the features.

Summary of the individualised treatment effect estimation framework

To summarize the above, we present a causal analytic framework, MR-ITE, to study the causal effect of risk factors at an individual level. The workflow for MR-ITE

is summarized in Fig. 1. The proposed framework integrates the idea of Mendelian Randomisation (MR) to identify individualised treatment effects (ITEs), which reduces the risks of unmeasured confounding and reverse causality.

The MR-ITE framework comprises several main steps. First, we identify valid SNPs associated with the exposure of interest, while minimizing the risk of pleiotropic effects. Second, using the identified SNPs, we estimate a polygenic risk score (PRS) that serves as an instrumental variable. Third, we employ two approaches, including GRF and DRIV, to estimate the ITEs. These methods can potentially handle nonlinear relationships and high-dimensional data, making them suitable for estimating the ITEs in large-scale datasets. Finally, we propose permutation-based methods to test for the presence of heterogeneity in treatment effects across individuals.

Simulation study

We conducted two simulations to assess the performance of our proposed framework in estimating ITEs and the power of our proposed heterogeneity testing methods.

We set up a simulation study with different pleiotropic scenarios to compare our proposed framework's performance with the regular causal forest (which does not employ instruments). We also compared individualised vs. constant treatment effects to demonstrate the importance of inferring individualised effects when heterogeneity is present. Overall, three different pleiotropic scenarios were included in our simulations, similar to ref[11]:

1. Balanced pleiotropy: some genetic variants directly affect the outcome, with pleiotropic effects that are a mixture of positive and negative effects averaging to zero.
2. Directional pleiotropy: some genetic variants directly affect the outcome, with all pleiotropic effects being positive.
3. Correlated Pleiotropy: some genetic variants affect the outcome via a confounder. In this case, the Instrument Strength Independent of Direct Effect (InSIDE) assumption is violated.

The simulation is set up following a similar idea from Burgess et al.[11] and the data is generated as follows:

$$U_i = \sum_{j=1}^{J} \xi_j G_{ij} + \epsilon_{Ui}$$

$$W_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_{Xi}$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \tau(X_i) W_i + U_i + \epsilon_{Yi}$$

where,

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

Here, U represents the confounders that contribute to both the treatment W and outcome Y. We simulated X as potential effect modifiers and it only contributes to the ITE ($\tau(X_i)$). We incorporated eight different treatment effect functions $\tau(X_i)$ to simulate the treatment effect τ; details can be found in Section 4, Appendix, of the Supplementary Methods.[33] We simulated 100 genetic variants $G_j, j = 1, \cdots, 100$ in each scenario and

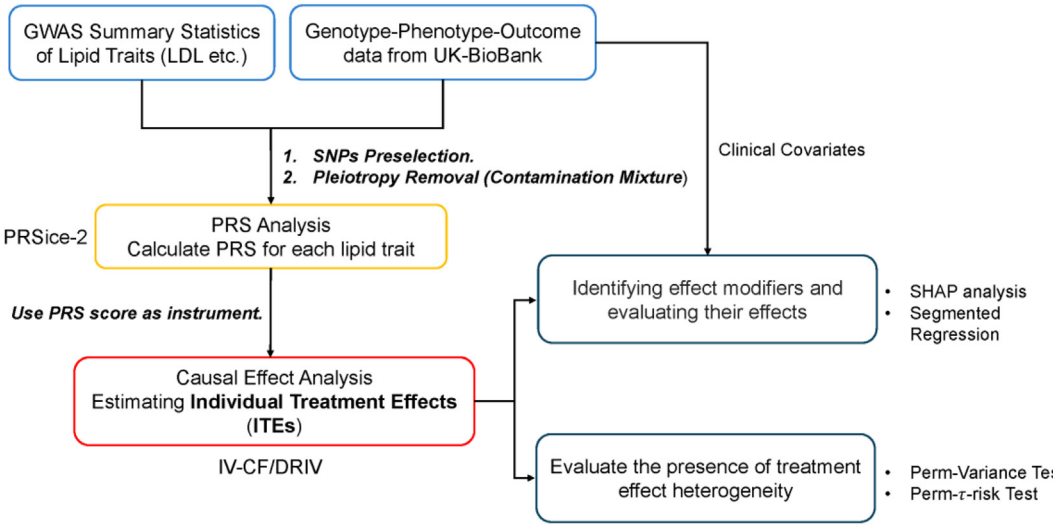

Fig. 1: **Workflow for the MR-ITE framework**. The figure presents the workflow for the MR-ITE framework, and the details of the method can be found in section 2.6 of main text and the supplement notes.

considered three cases with 20, 40, and 60 invalid instruments. We simulated two types of effect modifier X, including continuous (standard normal) and binary (binomial with probability 0.5) variables. Following the setting from Powers et al.,[33] we simulated 50 X_s for scenario 1 and 2, 40 X_s for scenarios 3 and 4, 30 X_s for scenarios 5–6 and 20 X_s for scenarios 7–8. Among these X_s, half were simulated as continuous variables, and another half as binary variables.

We set α_j and ξ_j to 0 for valid instruments. For invalid instruments, α_j and ξ_j were set differently for different scenarios. In the balanced pleiotropy scenario (scenario 1), α_j was simulated from uniform (–0.1, 0.1), and the ξ_j was set to 0. For the directional pleiotropy scenario (scenario 2), α_j was simulated from uniform (0, 0.1), and we set ξ_j to 0. In scenario 3 (Correlated pleiotropy), α_j were set to 0 and ξ_j were drawn from uniform (–0.1, 0.1). We fit a regression forest to model the relationship between exposure and instruments, and used the prediction from the regression forest model as the instrument in the simulation.[34] The simulation was repeated 50 times in each scenario. The sample size was set at 10,000. We assume that the SNP-exposure associations were derived from an independent dataset from the SNP-outcome data. In addition, the performance of our proposed HTE-testing methods was evaluated through simulations. We set up the simulation data following similar approaches discussed above, but only 40 invalid SNPs were included. Similarly, we repeated the simulation 50 times for each scenario. Furthermore, we conducted two additional simulations with varying configurations. First, we increased the sample size to 50,000. Next, we conducted simulations with the same sample size of 50,000 but included a total of 500 variants to assess our framework's ability to manage both larger sample sizes and a higher number of variants.

To evaluate the performance of our proposed framework in ITE estimation, we compared the estimated ITE with the true ITE and computed the mean squared error (MSE). Lower MSE indicates better performance. In addition, we examined the bias of the proposed approaches in estimating the ITE for individuals with true ITE ranked at the top 10% (considering the absolute value). We did not consider the bias considering the whole population, as treatment effects for those with large ITE tends to be underestimated, and vice versa; as such the positive and negative bias may cancel out each other. In addition, it is often more clinically relevant to focus on subjects with more extreme ITEs.

Applications to real data: heterogeneous effects of lipid traits on coronary artery disease risk

Overall analytic strategy

Using data from the UK-Biobank (UKBB) study, we applied our framework to study the heterogeneous treatment effect for several lipid-related risk factors on coronary artery disease (CAD). UK-Biobank is a large-scale cohort consisting of genetic and clinical data from ~500,000 participants. We selected white participants with data available for principal component analysis, to minimize risks of population stratification.

Exposure

The main exposure is lipid levels including LDL-C, HDL-C, triglyceride, and total cholesterol. They were extracted from the UKBB, detailed can be found in the GitHub repository.

Outcome and covariates

CAD diagnosis was determined by International Classification of Diseases, Tenth Revision (ICD-10) code I25 in field 41202–0.0 and date in field 41262–0.0. We only considered those CAD patients with CAD diagnosis after the date of the biomarker assessment.

For covariates, we selected clinical variables likely influencing both outcomes and exposure, which can be roughly classified into three groups: biomarkers, medical history and lifestyle history (detailed in Table S1). We converted discrete variables to dummy variables, and missing data was imputed by the missRanger package.[35]

We trained two models with different covariates sets. For covariate model 1, we only included age and sex as covariates; this mimics practical applications when there is only limited covariate information. For covariate model 2, we additionally adjusted for multiple biomarkers and socio-demographic covariates (Table S1). With the incorporation of a larger set of covariates, we hope to identify covariates contributing to potential heterogeneity of the effect of lipids on CAD risks.

Genetic instruments

The GWAS summary statistics for lipid traits was obtained from the Global Lipids Genetics Consortium.[36] We also obtained CAD summary statistics dataset from CARDIoGRAMplusC4D Consortium.[37] In addition, we also checked that the summary statistics dataset used for PRS calculations had no overlap with the UKBB cohort.

Considering that we are using polygenic risk score as an instrument, additional quality control of genetic data is required. We followed the recommended quality control pipeline of PRSice-2 to ensure target data meets GWAS standards. Specifically, we removed SNPs with low genotyping rate (-geno 0.01), low minor allele frequency (-maf 0.001), and individuals with the low genotyping rate (-mind 0.01) following the default settings.[38,39] Only variants strongly associated with the exposure were included for subsequent analysis (P-value < 5e-8).

ITE analysis

Overall, the main study included 276,054 subjects of European ancestry, among whom 13,010 were identified as having coronary artery disease (CAD) which occurred after biomarker measurements. Additionally, we

included 2559 African and 6254 South Asian subjects in a supplementary analysis to explore the impact of ancestry on ITE estimates.

The main outcome is the development of coronary artery disease (CAD) after the measurement of lipid levels. In our application, we also compared IV-GRF and DRIV approaches with a more standard (non-instrumental) approach, causal forest (CF, implemented in the R package GRF), in which the risk factor was directly modelled without genetic instruments.

Since so far ML-based ITE models are mainly developed for linear outcomes, we model the outcome (CAD) also as a continuous outcome, hence the treatment effects are on a linear probability scale (i.e., it reflects the changes in absolute risk or incidence of CAD per unit change of the exposure/treatment). In fact, it is not uncommon to employ linear models for binary outcomes in GWAS studies,[40] and such use may be justified by the observation that linear model is a first order Taylor approximation to a generalized linear model.[41]

As for the "treatment" variable, we considered two cases: lipid levels as a continuous and a binary treatment. In the former case of a continuous 'treatment', the ITE reflects the change in the absolute risk of CAD per unit *increase* of lipid level; whereas for a binary treatment, the ITE is the change in absolute risk of CAD for a change from dyslipidaemia (LDL-C>130 mg/dL[42]; TC > 220 mg/dL[43]; HDL-C <46 mg/dL[44]; TG > 150 mg/dL[42,45]) to normal levels.

Additional analysis with CRP and IGF-1 as exposures

We also investigated C-reactive protein (CRP) and Insulin-like growth factor-1 (IGF-1) as potential causal risk factors for coronary artery disease (CAD), beyond lipid traits. These inflammatory and metabolic markers were selected due to their associations with CAD risk factors and lower risk of pleiotropy in MR analysis. Their causal roles and heterogeneous treatment effects on CAD remain unclear. Please refer to the supplementary text (Supplementary Notes, Section 2.2) for details.

Subgroup analysis and clinical implications for patient/disease subtyping

Due to the lack of a comparable external dataset, we propose a subgroup analysis to provide support to our heterogeneity findings. We first trained an ITE estimation model using a generalized random forest, and selected the 'best representative tree' (the tree with the lowest R-loss) as the final model to partition people into different subgroups. This model, trained on a fraction of the data, was applied to the remaining samples to identify subgroups with significantly different local average causal treatment effects (LATE). This analysis also suggests an important clinical application of the MR-ITE framework, namely identifying distinct subgroups of individuals with diverse responses to treatment or risk factors. For details, please also refer to the supplementary text (Supplementary Notes, Section 1.8).

Role of funders

The study sponsors did not play a role in manuscript design, data collection, data analysis, interpretation, or in the writing of the manuscript.

Results

Simulation study

Fig. 2 presents the simulation results for various treatment effect scenarios, considering different counts of invalid SNPs. In the balanced pleiotropy scenario, the Instrumental Causal Forest (IV-CF) significantly outperformed the regular Causal Forest (CF), as evidenced by a substantially lower Mean Squared Error (MSE) (Fig. 2a, b, c). Under balanced pleiotropy, the removal of invalid pleiotropic SNPs did not appear to influence the performance of the estimator; there was no significant difference between methods that keep or remove the pleiotropic SNPs.

We now turned to the scenarios of directional pleiotropy and correlated pleiotropy. As expected, IV-CF outperformed ordinary CF across all treatment effect scenarios, although the degree of improvement gradually diminished with an increase in the count of invalid SNPs. Contrary to the balanced pleiotropy scenarios, the removal of invalid pleiotropic SNPs significantly enhanced the performance of IV-CF (Fig. 2d–i) under directional pleiotropy or correlated pleiotropy. These simulation results underscored the importance of incorporating an appropriate step for the removal of invalid pleiotropic SNPs within our framework. We further evaluated the efficacy of the ConMix approach in eliminating invalid SNPs. We found that the ConMix approach could detect invalid SNPs with an accuracy of approximately 80% in our simulations (Table S2).

To highlight the importance of inferring individualised treatment effects, we also compared the performance of the ATE and ITE estimators. ATE represents the average treatment effect, which assumes a constant treatment effect across all individuals, while ITE allows the treatment effects to differ by individual.

Under heterogeneous treatment effect scenarios (scenarios 2–8), the MSE of ATE are notably higher (worse) than those of ITE, in comparisons of ITE vs. ATE or MR-ITE (pleiotropy removed) vs. MR-ATE (pleiotropy removed). On the contrary, under the homogeneous treatment effect scenario (scenario 1), no significant difference is observed when ATE was compared to ITE. These findings emphasize the importance of inferring individualised treatment effects in the presence of heterogeneity.

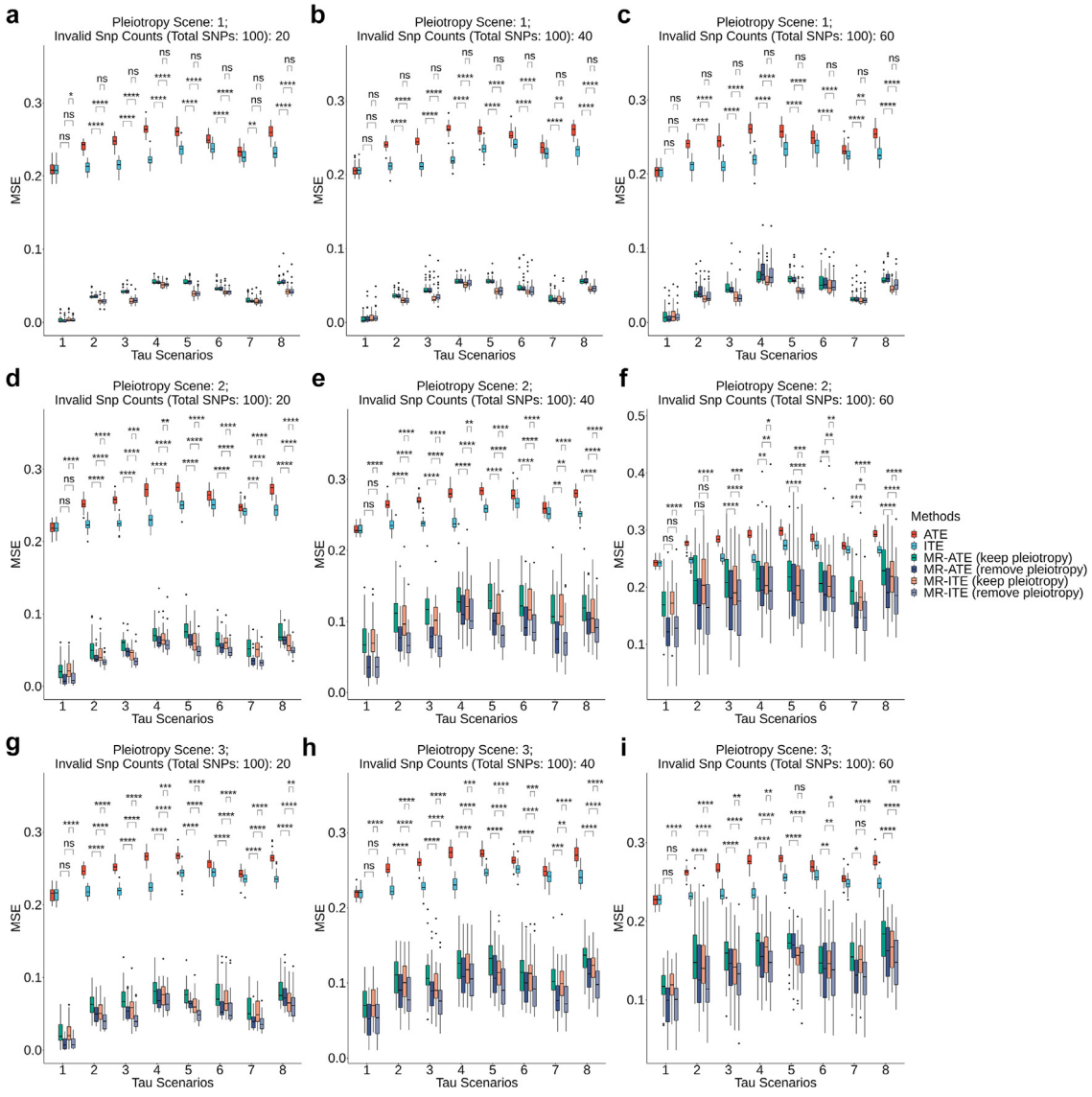

Fig. 2:* Simulation results across different methods to evaluate causal treatment effects**. The figure presents the simulation results across 8 different treatment effect scenarios, 3 pleiotropy scenarios and 3 invalid SNPs scenarios. For details of the generating distribution and scenarios, see section 2.7 of main text. The 6 estimators being evaluated as follows: ATE = average treatment effect (without using instruments); ITE = individualised treatment effect (without using instruments); MR-ATE (keep pleiotropy) = MR-based average treatment effect with the presence of pleiotropy; MR-ITE (keep pleiotropy) = MR-based individualised treatment effect with the presence of pleiotropy; MR-ATE (remove pleiotropy) = MR-based average treatment effect with pleiotropic instruments removal; MR-ITE (remove pleiotropy) = MR-based individualised treatment effect with pleiotropic instruments removal. All results presented in the figure are estimated using the generalized random forest (i.e., causal/instrumental forest) approach. We presented paired one-tailed t-test results on the following 3 comparison sets (from the lowest position to the highest position above each scenario): (1) MR-ATE (keep pleiotropy) vs. MR-ITE (keep pleiotropy); (2) MR-ATE (remove pleiotropy) vs. MR-ITE (remove pleiotropy); (3) MR-ITE (keep pleiotropy) vs. MR-ITE (remove pleiotropy). We hypothesized that MR-ITE methods performed better than MR-ATE (that ignores treatment effect heterogeneity), and that methods removing pleiotropic variants performed better. ("*": P-value < 0.0001; "***": 0.0001 < P-value < 0.001; "**": 0.001 < P-value < 0.01; "*": 0.01 < P-value < 0.05; "ns": P-value > 0.05; Paired T-test).

We also plotted the bias, and observed that our proposed MR-ITE approach exhibits superior bias control compared to conventional methods in our simulations (Fig. S1). Here we focused on the bias of subjects having the highest 10% of true ITEs, as the positive and negative bias of ITE may cancel out if we calculate the

average bias from all subjects. Also note that our simulations yielded predominantly negative ITEs, such that a positive bias is expected when comparing the true ITEs of those having the strongest 10% ITE to the ATE (which would be closer to zero). We also performed a comparison of the DRIV and IV-CF MR-ITE approaches (Figs. S2 and S3); there is no single method that uniformly dominated the other in all scenarios, suggesting that it may be useful to present both approaches in real data applications.

Furthermore, we conducted two additional simulations with different settings. First, we increased the sample size to 50,000 to explore our proposed framework's performance in handling larger datasets. We found that MR-ITE with pleiotropy removal achieved significantly better performance than its competitors (Figs. S4 and S5). Under larger sample sizes, the MSE and bias in general were lower for all approaches including MR-ITE. Additionally, we performed simulations with a sample size of 50,000 and a total of 500 variants to explore our framework's capability in handling situations with both larger numbers of variants and increased sample size (Figs. S6 and S7). Note that the number of invalid SNPs was also increased to 100, 200 and 300 correspondingly. The results demonstrated that our proposed framework is applicable under these more complex scenarios. We found that MR-ITE with pleiotropy removal still achieved the smallest MSE and bias compared to other methods. As a whole, these additional simulations demonstrate the robustness of our proposed approach in handling larger and more complex datasets.

In addition to benchmarking the performance of applying PRS as an instrument in inferring heterogeneous treatment effects, we conducted a simulation to validate our proposed heterogeneity-detecting methods. The results, summarized in Table 1, reveal that both methods maintain good type 1 error control in a scenario with no heterogeneity (scenario 1). They also demonstrate good power in several scenarios (scenarios 2, 3, 5), where the ITE functions ($\tau(\cdot)$) are simple linear combinations of the same types of covariates or exhibit weak nonlinear effects without interaction between different types of covariates. However, the permutation τ-risk test outperforms the permutation-variance test when an interaction between different types of covariates is present (Scenarios 4, 6). Interestingly, the permutation-variance test exhibits low power in scenario 7, where a strong nonlinear effect exists in the $\tau(\cdot)$, while the permutation τ-risk test maintains relatively good power in this scenario.

We also recorded the time and memory requirements for each simulation replicate under the simulation scenarios with 50,000 samples and 500 variants (Table S3), which provide a reference for estimating the running time and memory requirement in real applications.

Treatment effect of lipid-related traits on coronary artery disease

Baseline characteristics of included participants

The baseline characteristics of the study's continuous and categorical covariates are summarized in the Supplementary Materials. We conducted a partial F-test to evaluate the strength of the polygenic risk score as an instrument.[46] The F-statistics significantly exceeded 10 across all models (Table S4), indicating that the polygenic risk score can be considered a strong instrument. The covariates included for various lipid-related traits models are also outlined in the Supplemental Materials (Table S1). We also compared the estimate of overall treatment effect based on a standard regression against that from an instrumental regression, using the Wu–Hausman test as implemented in IVreg.[47–49] If the null hypothesis is rejected, it indicates that the explanatory variable is endogenous. In this case, the IV estimator is consistent, while the standard regression estimator is not. Conversely, if the null hypothesis is not rejected,

Pleiotropy scenario	Perm–Variance test			Perm-τ-risk test		
	Balanced pleiotropy	Directional pleiotropy	Correlated pleiotropy	Balanced pleiotropy	Directional pleiotropy	Correlated pleiotropy
Scenario 1	0.00	0.00	0.00	0.04	0.04	0.04
Scenario 2	0.28	0.28	0.56	0.88	0.92	0.98
Scenario 3	0.64	0.68	0.88	0.98	1.00	1.00
Scenario 4	0.12	0.18	0.42	0.92	0.98	0.98
Scenario 5	0.44	0.52	0.86	1.00	0.98	1.00
Scenario 6	0.12	0.28	0.32	0.88	0.86	0.96
Scenario 7	0.02	0.04	0.00	0.42	0.54	0.50
Scenario 8	0.36	0.54	0.68	1.00	1.00	1.00

The table shows the simulation results of two permutation-based heterogeneity testing methods, the permutation-variance test, and the permutation-τ-risk test. Scenario 1 is a scenario with homogeneous treatment effect, and the results refer to the type I error of the method. The rest of the scenarios show heterogeneous treatment effect, and the results reflect the power of the test in detecting the presence of heterogeneity.

Table 1: **Simulation results for two permutation-based heterogeneity testing methods.**

the IV and the ordinary regression estimator are considered to be both consistent, although the IV estimator has a larger variance. The original regression estimator is preferred in this case. Our results suggest that the IV estimator is preferred in the studies of LDL-C, total cholesterol and IGF-1 (Table S4). Consequently, we focus our discussion primarily on the findings from these three instrumental variable models.

LDL-C

LDL-C imposes heterogeneous effects on CAD. We utilized our MR-ITE framework to investigate the causal association between LDL-C and CAD risk under both continuous and binary exposure scenarios. The findings reaffirmed that elevated LDL-C level is causally linked to an increased risk of CAD. This association was consistently observed across two distinct covariate models (Fig. 3a and b, and Fig. S8a and b).

In our study on CAD, we observed that IV-GRF, DRIV, and causal forest (CF) predicted positive treatment effects for all participants. Notably, IV-GRF and DRIV yielded higher predictions of treatment effects than CF in Model 2. Additionally, DRIV detected a less significant treatment effect compared to IV-GRF. Note that for the CF approach, we simply modelled the original risk factors without using instruments, as previously described.

In covariate Model 1 ('reduced' model; only age/sex included), the CF approach failed to detect a significant treatment effect for most patients (especially under binary treatment setting), with ITE centred around zero (Table S5). We hypothesize that this may be due to the inclusion of only two covariates, leading to a failure in controlling for potential confounders (model 1 is designed to mimic the case when only very limited information on confounders is available). This contrasts with the established finding that higher LDL-C increases CAD risk. We note that the IV-based methods reported substantially larger treatment effects under model 1, suggesting better ability to handle unmeasured confounding.

In Model 2 (full covariate model), analyses using IV-GRF and DRIV indicate that a per unit (1 mg/dL) increment of LDL-C increases the individual absolute CAD risk by approximately 0.03%. In contrast, for CF, the average risk increase was less than 0.02% (Fig. 3a LDL-C and Table 2). Under a 'binary treatment' scenario (normal lipid levels vs. dyslipidaemia; modelling drug treatment effects), IV-GRF/DRIV also yielded higher treatment effect predictions. Reducing LDL-C levels to a normal range below the optimal cutoff (130 mg/dL) was estimated to reduce CAD incidence by ~3% based on the MR-ITE model (Table 3).

Intriguingly, we found that the CF only detects an average CAD risk reduction of approximately 0.5% under the 'binary treatment' setting. This is significantly less than what is detected in the MR-ITE framework (Fig. 3b and Table 3), probably owing to unmodelled confounding factors. The Wu–Hausman test also indicated significant differences between the non-instrument and the IV estimates. Our findings are largely consistent with previous studies. For instance, Brian et al. reported an absolute risk reduction in Atherosclerotic Cardiovascular Disease (ASCVD) ranging from 2.1% to 8.6% for patients whose LDL-C levels were controlled under 100 mg/dL following LDL-C reduction therapy.[50] Notably, compared to ordinary MR methods, our proposed MR-ITE framework enables the estimation of a causal effect for *each individual*.

We further evaluated the *heterogeneity* of treatment effects using our proposed permutation-based tests. Our findings indicate that LDL-C modification results in heterogeneous effects on CAD in both continuous and binary treatment settings (Table 4). We also observed a considerable variability in the ITE estimates, as indicated by a large (absolute) percentage difference between different quantiles. Specifically, the percentage difference was >50% when comparing the 5th vs. the 95th, and the 10th vs. the 90th percentiles under a continuous treatment setting (Table 5) and was >30% under a binary treatment setting (Table 6).

Additionally, our findings of heterogeneity were supported by the subgroup analysis. We observed significant differences in the mean treatment effects across the derived subgroups, in both binary and continuous treatment settings (Continuous Model: ANOVA P-value = 0.0096; Binary Model: ANOVA P-value = 0.0138) (Tables S14 and S16).

Another issue addressed in our study is the applicability of PRS developed for ancestries other than European. To explore this, we conducted additional analyses on participants of African and South Asian descent from the UK Biobank (UKBB) cohorts. Our results indicate that the results for African ancestry are comparable to those for European ancestry (Fig. S15). However, due to much smaller sample sizes of both ethnic groups in the UKBB compared to Europeans, and the smaller sample sizes used to derive the GWAS summary statistics for LDL-C,[51] we did not observe significant ITE ($P < 0.05$) for either ethnic group. For Africans, we observed positive ITEs for the majority of subjects (Fig. S15), while the pattern was less clear for South Asians.

In addition to ancestry, we further explored the application of other PRS construction algorithms within our framework. Specifically, we additionally employed LDPred2 to calculate the PRS for LDL-C and conducted ITE estimation using the LDPred2 PRS as the instrument. Our results indicated that the ITEs estimated from LDPred2-derived PRS and PRSice-2-derived PRS were comparable (Fig. S14), adding to the robustness of our findings. However, we observed that the number of observations with significant ITEs identified using

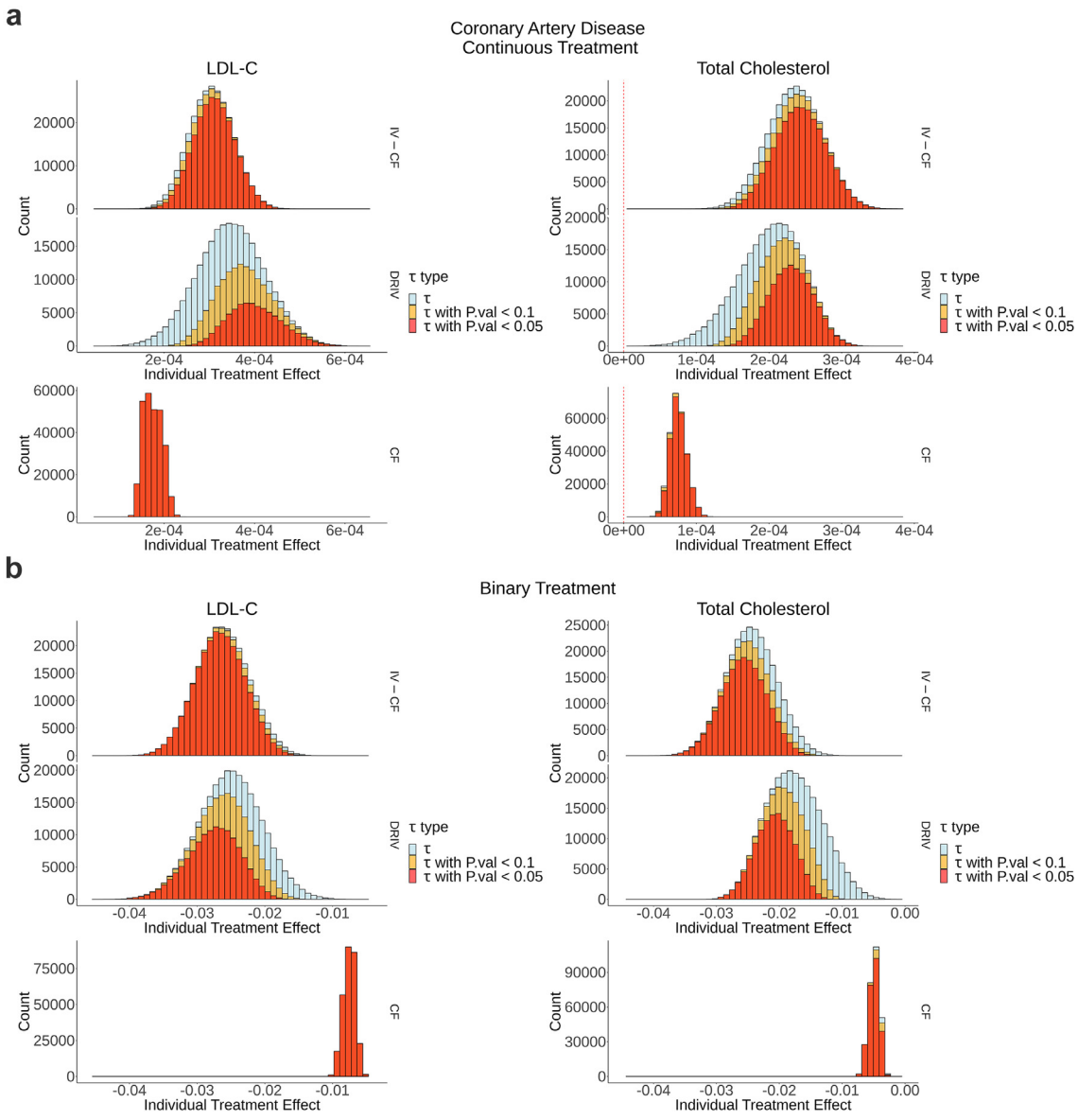

***Fig.* 3: Predicted treatment effect of LDL-C and total cholesterol on CAD (Model 2).** The figure presents the histogram results of the individualised causal effect estimation of LDL-C/Total Cholesterol on CAD using DRIV, IVCF, and CF methods, incorporating covariate set 2 (full model). a: individualised causal effect estimation in the continuous treatment setting. b: individualised causal effect estimation in the binary treatment setting. For continuous treatment settings, the ITE reflects changes in absolute risk of CAD per unit increase in lipids; for binary treatment settings, the ITE reflects the effect of having normal lipid levels vs. dyslipidaemia (the latter as baseline), analogous to the effect of receiving a treatment for dyslipidaemia (the same principles apply to figures below).

LDPred2-derived PRS is substantially smaller than those identified using PRSice-derived PRS. This discrepancy may possibly be due to the large set of instrument SNPs with linkage disequilibrium (LD) used for LDpred2 input. It should be noted that the ConMix approach was developed to handle independent SNPs. Consequently, the inclusion of pleiotropic SNPs may compromise the validity of the PRS, introducing additional bias into the ITE estimation process. Please also refer to the supplementary text (Supplementary Notes, Section 1.2.2.2) for further discussions.

Clinical features that contribute to the heterogeneity of LDL-C on CAD. Beyond the identification of heterogeneity concerning the impact of LDL-C on CAD, our interest extends to the covariates that contribute to this heterogeneity. Fig. 4a and l depict the Shapley value (SHAP) patterns of the top 10 most important clinical features,

Traits	Methods	Positive Tau	Tau with P-value <0.1	Tau with P-value <0.05	Min	5%	10%	25%	Median	75%	90%	95%	Max	Mean
LDL-C	CF	100.00%	100.00%	100.00%	1.190E-04	1.454E-04	1.497E-04	1.586E-04	1.737E-04	1.909E-04	2.024E-04	2.081E-04	2.419E-04	1.750E-04
	DRIV	100.00%	55.94%	27.99%	3.867E-05	2.241E-04	2.515E-04	2.977E-04	3.483E-04	3.990E-04	4.458E-04	4.728E-04	6.622E-04	3.484E-04
	IVCF	100.00%	94.86%	86.69%	1.118E-04	2.255E-04	2.424E-04	2.715E-04	3.047E-04	3.378E-04	3.672E-04	3.845E-04	5.060E-04	3.047E-04
HDL-C	CF	100.00%	99.98%	99.81%	-4.262E-04	-3.471E-04	-3.316E-04	-3.017E-04	-2.628E-04	-2.189E-04	-1.818E-04	-1.646E-04	-9.956E-05	-2.598E-04
	DRIV	99.96%	14.03%	3.35%	-7.429E-04	-4.795E-04	-4.367E-04	-3.617E-04	-2.803E-04	-2.078E-04	-1.515E-04	-1.217E-04	6.381E-05	-2.881E-04
	IVCF	97.34%	3.69%	1.09%	-1.227E-03	-6.854E-04	-6.125E-04	-4.938E-04	-3.643E-04	-2.368E-04	-1.226E-04	-5.452E-05	5.590E-04	-3.661E-04
Total Cholesterol	CF	100.00%	99.11%	96.12%	3.167E-05	5.698E-05	6.060E-05	6.681E-05	7.415E-05	8.233E-05	9.032E-05	9.508E-05	1.280E-04	7.486E-05
	DRIV	100.00%	69.62%	46.00%	-2.107E-06	1.210E-04	1.403E-04	1.717E-04	2.045E-04	2.338E-04	2.572E-04	2.697E-04	3.456E-04	2.012E-04
	IVCF	100.00%	89.78%	78.33%	8.537E-05	1.735E-04	1.872E-04	2.104E-04	2.363E-04	2.621E-04	2.845E-04	2.974E-04	3.863E-04	2.360E-04
Triglyceride	CF	100.00%	80.98%	65.32%	5.382E-06	1.734E-05	1.897E-05	2.193E-05	2.553E-05	2.915E-05	3.226E-05	3.407E-05	4.666E-05	2.558E-05
	DRIV	75.86%	0.00%	0.00%	-1.593E-04	-3.960E-05	-2.401E-05	1.117E-06	2.826E-05	5.489E-05	7.895E-05	9.342E-05	2.224E-04	2.775E-05
	IVCF	21.99%	0.58%	0.13%	-1.590E-04	-7.762E-05	-6.616E-05	-4.677E-05	-2.502E-05	-3.096E-06	1.658E-05	2.835E-05	1.097E-04	-2.490E-05

The table summarizes the estimated individualised treatment effects (ITE) for different lipid traits to CAD risk, treating lipid as a continuous predictor. The model (referred to as 'Model 2' in the main text) was trained using all available information, including sex, age and other clinical measured covariates. Tau indicates the change in the probability of outcome (i.e., absolute risk of CAD) for every one-unit (10 mg/dL) increase of the lipid level.

Table 2: **Summary of the individualised treatment effect (tau), where lipid is treated as a continuous trait.**

as identified through the DRIV model of LDL-C's influence on CAD. These patterns are presented under both continuous and binary treatment settings, utilizing a beeswarm plot for visualization. Furthermore, we segmented the population into deciles based on the corresponding feature values, enabling the visualization of potential effect modifiers.

We discovered that the body fat percentage was the most significant variable under both settings (Fig. 4a, l). Patients with higher body fat showed weaker CAD protection from LDL-C reduction vs. those with lower fat. This pattern was also seen in other obesity indicators such as BMI (Fig. S9e, j). These findings align with several studies that highlight the robust relationship between obesity and CAD. For example, Sandfort et al. demonstrated that obese patients with hyperlipidaemia experience more severe atheroma progression despite optimized statin therapy.[52] This is largely consistent with our observation that obese patients may derive less benefit from LDL-C reduction. Our results suggest that a combination treatment of both obesity and elevated LDL-C may achieve a more substantial protective effect against CAD.

Additionally, systolic blood pressure (SBP) emerged as an important variable, akin to the obesity-related covariates previously mentioned, in both continuous and binary treatment settings (Fig. 4n and s, Fig. S9a, f). Numerous studies have established hypertension as one of the most potent risk factors for cardiovascular diseases, including CAD.[53,54] Our research suggests that hypertension may act as an effect modifier of LDL-C's impact on CAD. We note a reduced protective treatment effect in the population with SBP in the top 10% (Fig. S9f), and the SHAP analysis yields a positive SHAP estimation (negative SHAP indicates protective effects) for SBP exceeding 150 in the binary model. These findings imply that hypertension could significantly weaken the protective effect of LDL-C lowering against CAD. Our results also corroborate with a study finding that combination therapy of LDL-C and blood pressure-lowering agents was associated with a lower risk of CAD compared to monotherapy.[55]

Testosterone exhibits a similar pattern to age (Fig. 4d and i), suggesting males may receive a larger protective effect against CAD with the decrease of LDL-C to a normal range. Of note, Petretta et al. reported that statin therapy significantly reduced the risk of CHD events in men without prior cardiovascular disease, while its effect on women was less significant.[56] In addition, another large-scale meta-analysis of RCTs reported that the absolute risk reduction of cardiovascular events was larger in men than women (absolute number of vascular events reduced per 1000 treated was 12 in men vs. 9 in women, for those with low baseline risks).[57]

Traits	Methods	Negative Tau	Tau with P-value <0.1	Tau with P-value <0.05	Min	5%	10%	25%	Median	75%	90%	95%	Max	Mean
LDL-C	CF	100.00%	100.00%	99.97%	−1.076E-02	−9.056E-03	−8.748E-03	−8.156E-03	−7.464E-03	−6.879E-03	−6.430E-03	−6.173E-03	−3.996E-03	−7.529E-03
	DRIV	100.00%	73.81%	47.48%	−4.568E-02	−3.304E-02	−3.126E-02	−2.829E-02	−2.508E-02	−2.192E-02	−1.910E-02	−1.743E-02	−4.850E-03	−2.513E-02
	IVCF	100.00%	96.63%	91.32%	−4.283E-02	−3.265E-02	−3.124E-02	−2.891E-02	−2.627E-02	−2.358E-02	−2.117E-02	−1.979E-02	−1.082E-02	−2.625E-02
HDL-C	CF	0.00%	100.00%	99.98%	5.405E-03	6.774E-03	7.040E-03	7.568E-03	8.186E-03	8.759E-03	9.232E-03	9.495E-03	1.120E-02	8.162E-03
	DRIV	13.16%	0.88%	0.07%	−1.595E-02	−2.571E-03	−8.170E-04	2.292E-03	6.029E-03	1.000E-02	1.363E-02	1.572E-02	2.929E-02	6.226E-03
	IVCF	10.19%	0.86%	0.21%	−2.978E-02	−3.843E-03	−1.080E-04	6.368E-03	1.400E-02	2.228E-02	3.018E-02	3.512E-02	7.274E-02	1.462E-02
Total Cholesterol	CF	100.00%	97.14%	90.91%	−8.263E-03	−6.291E-03	−5.992E-03	−5.439E-03	−4.809E-03	−4.259E-03	−3.849E-03	−3.628E-03	−2.170E-03	−4.869E-03
	DRIV	100.00%	68.25%	45.63%	−3.338E-02	−2.473E-02	−2.339E-02	−2.091E-02	−1.783E-02	−1.455E-02	−1.167E-02	−1.003E-02	−1.001E-04	−1.766E-02
	IVCF	100.00%	82.87%	68.88%	−4.498E-02	−3.111E-02	−2.968E-02	−2.726E-02	−2.453E-02	−2.177E-02	−1.931E-02	−1.783E-02	−8.020E-03	−2.451E-02
Triglyceride	CF	100.00%	70.43%	53.96%	−6.121E-03	−4.517E-03	−4.279E-03	−3.865E-03	−3.386E-03	−2.883E-03	−2.421E-03	−2.148E-03	−2.828E-04	−3.365E-03
	DRIV	40.30%	0.87%	0.13%	−2.541E-02	−8.080E-03	−5.941E-03	−2.438E-03	1.408E-03	5.478E-03	9.837E-03	1.298E-02	3.648E-02	1.750E-03
	IVCF	11.76%	0.38%	0.07%	−1.815E-02	−2.860E-03	−5.994E-04	3.219E-03	7.529E-03	1.183E-02	1.562E-02	1.780E-02	3.223E-02	7.513E-03

The table summarizes the estimated individualised treatment effects for different lipid-trait to disease models under binary treatment setting in model 2. Model 2 is trained using available information, including sex, age, and other clinical measured covariates. Tau indicates the increase in probability of outcome for controlling the lipid level under the optimal threshold, in other words, we compare a favourable lipid profile (coded as 1) vs. an unfavourable lipid profile (coded as 0), analogous to having received a treatment to improve dyslipidaemia. Therefore, the direction of tau is reversed compared to the continuous treatment setting. Tau indicates the change in the probability of outcome (i.e., absolute risk of CAD) comparing normal vs. abnormal lipid profiles.

***Table 3*: Summary of the individualised treatment effect (tau), where lipid is treated as a binary trait (normal lipid levels vs. dyslipidaemia with the latter as baseline, analogous to the effect of receiving a treatment for dyslipidaemia).**

	Coronary artery disease	
	Perm-Var test	Perm-Risk test
Continuous trait		
LDL-C	0.01	0.01
Total cholesterol	0.01	0.00
IGF-1	0.01	0.00
Binary trait		
LDL-C	0.02	0.00
Total cholesterol	0.00	0.01

The table summarizes two permutation-based test results for assessing the presence of heterogeneity in different lipid-trait to disease setting. The number of permutation times was set to 50 in our application.

***Table 4*: Permutation-based test to assess the presence of heterogeneity (P-values from permutations shown).**

We also observed gamma glutamyl transferase (GGT) as a top-ranked feature (Fig. 4a), highlighting its role in the development of CAD. Previous studies have demonstrated that elevated levels of serum GGT are associated with the pathogenesis of CAD.[58,59] Consistent with these findings, our results suggest that a high level of GGT may attenuate the treatment effect of lowering LDL-C on CAD.

Serum calcium level was also identified as one of the top-ranked variables, in both continuous and binary treatment models. Our results indicate that a lower level of calcium corresponds to a stronger protective effect (negative SHAP value) under the binary model (Fig. 4f, k, p, u). This aligns with several studies that have highlighted the association between genetically elevated serum calcium and increased odds of CAD and myocardial infarction.[60,61] Interestingly, we also observed vitamin D as another top-ranked variable. We found that elevated vitamin D levels corresponded to stronger protective effects under the binary treatment model, consistent with studies suggesting potential beneficial effects of vitamin D on CAD.[62,63]

In addition to LDL-C, we also studied the ITEs of total cholesterol (TC), C-reactive protein (CRP) and insulin-like growth factor 1 (IGF-1) on the risk of CAD. We additionally explored the top variables that may modify the treatment effects using Shapley values. For detailed results, please refer to the supplementary text (Supplementary Notes, Section 2).

Discussion

Overview

In this study, we extend the regular MR approach to infer the *individualised* causal effects of risk factors/treatments in observational studies. Traditional MR primarily focuses on inferring the average causal effect,[64] which may not suffice in the era of precision medicine. Although the estimated average effect is still meaningful in designing policies or treatments for the population, it may obscure individual responses. To

Traits	Methods	50% Inter-quartile range (25% vs. 75%)	80% Inter-percentile range (10% vs. 90%)	90% Inter-percentile range (5% vs. 95%)	Absolute % difference (25% vs. 75%)	Absolute % difference (10% vs. 90%)	Absolute % difference (5% vs. 95%)
LDL-C	CF	3.226E-04	5.267E-04	6.275E-04	20.34%	35.18%	43.17%
	DRIV	1.013E-03	1.943E-03	2.486E-03	34.01%	77.25%	110.93%
	IVCF	6.634E-04	1.248E-03	1.590E-03	24.44%	51.50%	70.51%
Total Cholesterol	CF	1.552E-04	2.973E-04	3.810E-04	23.23%	49.05%	66.85%
	DRIV	6.207E-04	1.168E-03	1.487E-03	36.15%	83.24%	122.86%
	IVCF	5.168E-04	9.728E-04	1.239E-03	24.57%	51.97%	71.41%

The table summarizes the inter-percentile ranges of several selected percentiles, and the absolute percentage differences. The inter-percentile range is defined as $|\tau_{A\%} - \tau_{B\%}|$. The absolute percentage difference is defined as $\frac{|\tau_{A\%} - \tau_{B\%}|}{\tau_{B\%}} \times 100\%$. *A* and *B* denote a specific quantile (percentile).

Table 5: **Inter-percentile ranges and (absolute) percentage differences of ITE across percentiles, where lipid is treated as a continuous trait.**

address this limitation, we introduce a framework, MR-ITE, that integrates MR and machine learning methodologies to estimate individualised causal treatment effects. Under this framework, we also present two permutation-based tests to assess the presence of effect heterogeneity. The validity of the MR-ITE framework is supported by extensive simulations. Importantly, we also demonstrate the applicability of our proposed MR-ITE framework in realistic scenarios likely involving unobserved confounders.

As a proof-of-concept example, we applied our framework to study the individualised causal effects of several lipid-related traits on CAD, including LDL-C, HDL-C, triglyceride, and TC using the UKBB cohort, one of the largest biobank cohorts of the world, consisting of approximately 500,000 participants. We conducted rigorous data cleaning and limited the analysis to individuals of European ancestry, retaining ~300,000 subjects for formal analysis. This substantial sample size enhances the robustness and generalizability of our findings within European populations. Additionally, prior studies support the broader generalizability of exposure-outcome associations detected in the UKBB across other cohorts. For instance, Batty et al. demonstrated strong consistency in associations between risk factors and mortality endpoints when comparing the UKBB cohort with pooled data from the Health Surveys for England (HSE) and the Scottish Health Surveys (SHS),[65] highlighting the generalizability of UKBB findings to other cohorts. Similarly, Lin et al. evaluated the predictive power of combining biomarker-based PRS with standard PRS for CAD across two nationwide cohorts, UKBB and FinnGen, and observed consistent effect directions despite minor differences in effect sizes.[66] These findings further support the transferability of risk factor associations identified in the UKBB to other European cohorts.

Our analysis revealed evidence of effect heterogeneity, particularly for LDL-C and TC's effect on CAD. Using Shapley value analysis, we identified key clinical features that may modify the effects of LDL-C and TC on CAD.

On the other hand, our MR-ITE model did not detect significant causal effects of HDL-C and triglycerides, aligning with previous findings. For instance, Holmes et al. conducted an MR analysis using PRS as instruments, similar to our approach. They found no significant causal relationships between HDL-C or triglycerides and CAD,[67] particularly in results based on a more rigorous 'restricted' PRS (which filtered out SNPs associated with other lipid traits), or those based on the

Traits	Methods	50% Inter-quartile range (25% vs. 75%)	80% Inter-percentile range (10% vs. 90%)	90% Inter-percentile range (5% vs. 95%)	Absolute % difference (25% vs. 75%)	Absolute % difference (10% vs. 90%)	Absolute % difference (5% vs. 95%)
LDL-C	CF	1.277E-03	2.318E-03	2.882E-03	15.66%	26.50%	31.83%
	DRIV	6.372E-03	1.215E-02	1.561E-02	22.52%	38.88%	47.25%
	IVCF	5.334E-03	1.007E-02	1.285E-02	18.45%	32.23%	39.37%
Total cholesterol	CF	1.180E-03	2.142E-03	2.664E-03	21.69%	35.76%	42.34%
	DRIV	6.366E-03	1.172E-02	1.470E-02	30.44%	50.10%	59.43%
	IVCF	5.489E-03	1.037E-02	1.328E-02	20.14%	34.94%	42.68%

The table summarizes the inter-percentile ranges of several selected percentiles, and the absolute percentage differences. The inter-percentile range is defined as $|\tau_{A\%} - \tau_{B\%}|$. The absolute percentage difference is defined as $\frac{|\tau_{A\%} - \tau_{B\%}|}{\tau_{B\%}} \times 100\%$. *A* and *B* denote a specific quantile (percentile).

Table 6: **Inter-percentile ranges and (absolute) percentage differences of ITE across percentiles, where lipid is treated as a binary trait (normal lipid levels vs. dyslipidaemia with the latter as baseline, analogous to receiving a treatment for dyslipidaemia).**

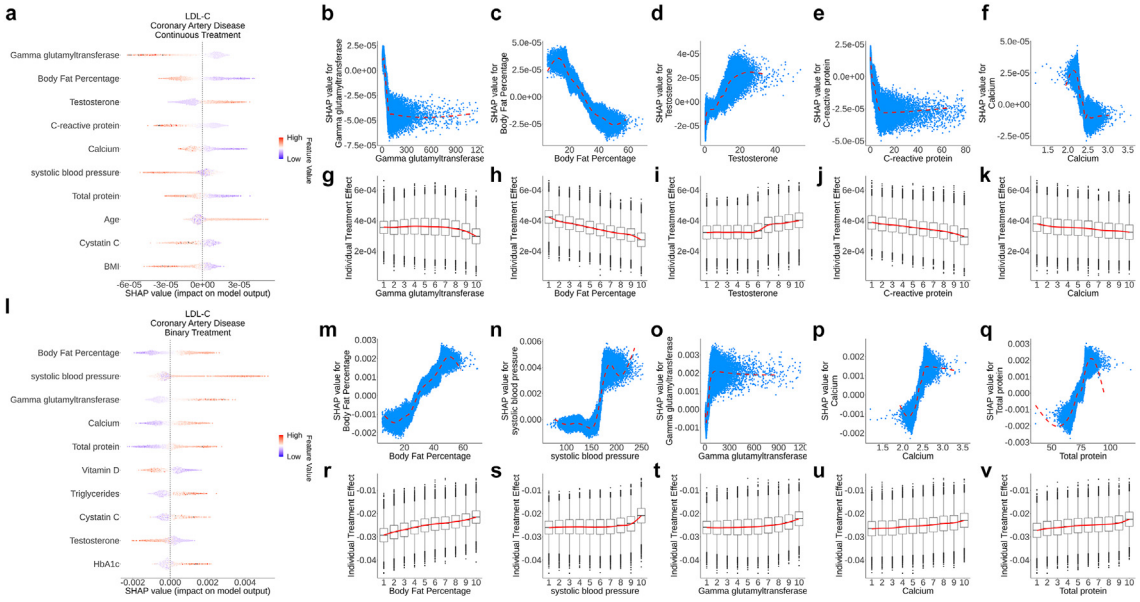

Fig. 4: **Top important variables identified with LDL-C as risk factor on CAD in continuous/binary treatment setting**. a, l: Beeswarmplot of top 10 important covariates identified under continuous trait and binary trait scenarios with SHAP analysis b, c, d, e, f, m, n, o, p, q: Scatterplots of SHAP value (y-axis) vs. observed value (x-axis) of top 5 important covariates identified under continuous/binary treatment scenario (LDL-C as risk factor and CAD as the outcome of interest). (b, c, d, e, f: Continuous Trait; m, n, o, p, q: Binary Trait) g, h, i, j, k, r, s, t, u, v: Boxplots of estimated individualised treatment effects (y-axis) vs. observed value (x-axis) of top 5 important covariates identified under continuous/binary treatment scenario where LDL-C serves as risk factor and CAD as the outcome of interest. In the boxplots, the x-axis represents the deciles of the feature. (g, h, i, j, k: Continuous Trait; r, s, t, u, v: Binary Trait).

unrestricted score with adjustment of statin use and other lipid traits. Similarly, Uribe et al. reported no significant association between genetically determined HDL-C and CAD.[68] Another MR study based on a Korean population also failed to find a clear significant causal link between HDL-C or TG with CAD.[69]

In addition to lipid traits, we analysed two proteins potentially linked to cardiometabolic disorders, namely CRP and IGF-1. We showed that the CRP is not causally associated with the risk of CAD, which agrees with findings from previous studies. For example, the C-Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) conducted an MR analysis of individual-level data, which indicated that CRP concentrations are unlikely to be a significant causal factor in CAD.[70] Similarly, Kuppa et al. reported no causal relationships between CRP and various cardiovascular diseases through a two-sample bidirectional MR study.[71]

In contrast, we found that higher IGF-1 is associated with increased CAD risks, and the effect showed significant heterogeneity across individuals. Notably, Larsson et al. performed an MR analysis that explored the causal role of IGF-1 in cardiometabolic diseases, and found that elevated serum IGF-1 levels were associated with higher risks of CAD, although this effect was attenuated after adjustment for diabetes.[72] In a meta-analysis of observational studies, Jing et al. reported that both low and high IGF-I levels were associated with elevated risks of cardiovascular disease (CVD) (when compared to the middle category of IGF-1 levels), especially in males.[73] The exact relationship between IGF-1 and CVD and the underlying mechanisms may warrant further studies.

Our application of the MR-ITE framework to UKBB underscores the utility of our approach, and the findings may have important clinical implications. Firstly, our study uncovered important insights that could help optimize the management of dyslipidemia and reduce CAD risk. By characterizing ITE and deriving an ML model to predict ITE, we could identify patients who may experience a more pronounced adverse impact of dyslipidemia on CAD risk. These high-risk individuals may be prioritized for more intensive lifestyle and pharmacological interventions aimed at treating dyslipidemia. Targeting treatments to those predicted to derive the greatest risk reduction from lipid control may maximize the efficiency of limited healthcare resources.

Secondly, our analysis revealed several clinical factors (effect modifiers) associated with differential responses to lipid-modifying therapies. Understanding such sources of treatment heterogeneity can guide clinical decision-making and more personalized prescription. For instance, we observed that obese patients may benefit less from lipid control; weight control in

addition to lipid-modifying drugs may lead to more pronounced benefit in terms of CAD prevention.

Moreover, we also uncovered several potential biomarkers (e.g. cystatin C, SHBG, GGT etc.) that may help differentiate patients with varying responses to lipid-modifying therapies. The identification of such effect modifiers may help in the development of combination therapies. For example, one may combine lipid-lowering drugs with another medication that target the effect modifier(s).

Finally, as described earlier, our proposed MR-ITE may also be employed for identifying patient subgroups with differential responses to treatment or risk factors.

Our study possesses several notable strengths. To the best of our knowledge, this is the first study aiming to estimate *individualised* causal treatment effects leveraging the principles of MR. Although it is possible for researchers to study ITE under an RCT setting, the inherent difficulties and substantial costs associated with RCTs often make such designs impractical. Our approach offers an alternative, enabling the inference of individualised treatment effects using genetic instruments. This method is considerably less susceptible to unknown confounders and reverse causality compared to observational studies. This key advantage could further expedite the progress of precision medicine, as interventions on risk factors can be customized for each individual based on the predicted ITE. Furthermore, our ITE estimation strategies are predicated on machine learning (ML) methods, which allow flexible modelling of complex relationships. In addition, the DRIV approach also allows virtually any ML model to be used, thereby enhancing the flexibility and applicability of our approach.

Moreover, our simulation results demonstrate that the proposed heterogeneity testing methods exhibited reasonable performance in the majority of scenarios. In addition, the integration of SHAP analysis within our framework helps to identify the primary variables contributing to ITEs. This not only facilitates more comprehensive model explanations but also potentially assists in patient sub-grouping or disease subtyping in practical applications.

Our study has several limitations that provide opportunities for future work. For instance, our simulation results show that the permutation variance test may not perform optimally in complex scenarios, such as those involving strong nonlinear treatment effects or interactions between different types of confounders. Additionally, unlike most MR studies which leveraged GWAS summary statistics only, MR-ITE requires individual-level data; however, the PRS may be derived from external summary data to minimize overfitting. Our study primarily focuses on estimating (individualised) absolute risk reduction (ARR) or changes as the target estimate; the study of ITE in terms of risk ratio (RR) will be considered as a topic for future studies. Another limitation is the absence of an independent external dataset for validating our results. It is relatively challenging to find a large-sample, phenotype-rich dataset with genotype information akin to the UK Biobank. Apart from the above, general limitations of MR may also apply[74,75]; for example, genetically predicted lipid levels may reflect long-term effects of lipid changes, and may not fully mimic the short-term effects of statins or other lipid-modifying drugs. Further replications and studies are also required to validate our findings regarding the effects of lipids on CAD risk.

Despite these limitations, the estimated ITEs are reasonable, and their range aligns with estimates from previous RCTs of lipid-lowering agents. We also note that our framework primarily considers a linear effect of the exposure, though nonlinear causal effects may be present in practical scenarios.[76]

Regarding the two MR-ITE approaches, we observed that the DRIV method did not yield as many statistically significant ITEs as IV-GRF, although the ITEs estimated from both methods are significantly correlated (Table S23). We speculate that one potential reason is that DRIV requires modelling the covariance between the exposure and the instrument conditional on the covariates, and this term needs to be included in the denominator (please refer to supplementary text for details and relevant formulas); this may lead to higher variance of DRIV estimates compared to IV-GRF. In practice, to determine which approach is preferred, we recommend comparing the modified R-loss (as defined in formula (6) in Supplementary Notes, Section 1.4.2) between DRIV and IV-GRF. If the difference is significant, the method with the smaller R-loss is considered more reliable. Otherwise, IV-GRF may be preferred due to its tendency to produce estimates with lower variance compared to DRIV. Future research may focus on further enhancing the performance of DRIV, particularly in scenarios where the instrument exhibits weak influence on the exposure within certain covariate regions.

We also acknowledge that the limited sample sizes (both for the exposure and outcome) of other ancestry populations, such as South Asians and Africans, raise concerns about the statistical power. In addition, South Asian and African participants were recruited in England rather than their regions of origin. Consequently, differences in environmental exposures and lifestyles, as well as variations in genetic backgrounds driven by factors such as the founder effect and gene flow, make it challenging to generalize the findings to populations in their ancestral homelands. As such, the results may not be as reliable as those estimated from European subjects. We hope to further investigate this issue with larger and more diverse datasets in the future.

In conclusion, we have developed a MR framework capable of estimating individualised causal effects in

observational studies. We have estimated the ITEs of lipid traits on CAD, and unveiled important clinical features that contribute to effect heterogeneity. It is our hope that our work will pioneer a new direction and paradigm for MR studies by providing a new method for identifying ITEs. We hope these insights will ultimately translate to clinical practice, informing more personalized treatment plans.

Contributors

Conceptualisation, Methodology, Writing—original draft: Yujia SHI, Hon-Cheong SO.

Data curation and investigation: Yujia SHI, Yong XIANG, Yuxin YE, Tingwei HE.

Formal analysis: Yujia SHI.

Writing—review & editing: Yujia SHI, Pak-Chung SHAM, Hon-Cheong SO.

Analysis, including accessing and verifying the underlying data, or Interpretation: Yujia SHI, Hon-Cheong SO.

All authors read and approved the final version.

Data sharing statement

UKBB data is available to any researcher who formally applies for the data. However, the data is not publicly available due to privacy concerns. All analysis code including simulation and formal analysis on UKBB data can be found in the following GitHub link: https://github.com/yujias424/MR-ITE.

Declaration of interests

The authors declare no relevant conflicts of interest.

Acknowledgements

This work was supported partially by a National Natural Science Foundation of China grant (NSFC; grant number 81971706), the KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China, and the Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong. An earlier version of this manuscript was posted on ResearchGate in August 2022 (https://doi.org/10.13140/RG.2.2.30569.77928/1) and MedRxiv in January 2024 (https://doi.org/10.1101/2024.01.18.24301507). During the preparation of this work, we have used GPT-4o in order to correct grammatical errors and improve the overall readability. We have reviewed and confirmed the validity of the text and take full responsibility for the content of the publication. We would also like to thank Dr Kai Zhao, Ms. Alexandria Lau, Prof. Stephen Tsui and Prof. Cara Cao for useful discussions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105616.

References

1 Reynolds R, Dennis S, Hasan I, et al. A systematic review of chronic disease management interventions in primary care. *BMC Fam Pract*. 2018;19:1–13.
2 Kent DM, Steyerberg E, Van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*. 2018;363:k4245.
3 Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of diabetes prevention program. *BMJ*. 2015;350:h454.
4 Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. *Nat Med*. 2024;30(4):958–968.
5 Akobeng AK. Understanding randomised controlled trials. *Arch Dis Child*. 2005;90(8):840–844.
6 Lilford RJ, Edwards S, Braunholtz DA, Jackson J, Thornton J, Hewison J. *Ethical issues in the design and conduct of randomised controlled trials. advanced handbook of methods in evidenced based healthcare*. 2001:11–24.
7 Roessner V. Large sample size in child and adolescent psychiatric research: the way of salvation? *Eur Child Adolesc Psychiatry*. 2014;23:1003–1004.
8 Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27(8):1133–1163.
9 Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*. 2017;26(5):2333–2355.
10 Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2021;108(2):299–319.
11 Burgess S, Foley CN, Allara E, Staley JR, Howson JM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat Commun*. 2020;11(1):376.
12 Spielman LJ, Little JP, Klegeris A. Inflammation and insulin/IGF-1 resistance as the possible link between obesity and neurodegeneration. *J Neuroimmunol*. 2014;273(1-2):8–21.
13 Franco C, Brandberg J, Lönn L, Andersson B, Bengtsson B, Johannsson G. Growth hormone treatment reduces abdominal visceral fat in postmenopausal women with abdominal obesity: a 12-month placebo-controlled trial. *J Clin Endocrinol Metab*. 2005;90(3):1466–1474.
14 Visser M, Bouter LM, McQuillan GM, Wener MH, Harris TB. Elevated C-reactive protein levels in overweight and obese adults. *JAMA*. 1999;282(22):2131–2135.
15 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.
16 Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Smith GD. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr*. 2016;103(4):965–978.
17 Teumer A. Common methods for performing Mendelian randomization. *Front Cardiovasc Med*. 2018;5:51.
18 KolesÃr M. *Estimation in an instrumental variables model with treatment effect heterogeneity*. 2013.
19 Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. *Nat Rev Methods Primers*. 2022;2(1):6.
20 Boehm FJ, Zhou X. Statistical methods for Mendelian randomization in genome-wide association studies: a review. *Comput Struct Biotechnol J*. 2022;20:2338–2351.
21 Athey S, Tibshirani J, Wager S. *Generalized random forests*. 2019.
22 Syrgkanis V, Lei V, Oprescu M, Hei M, Battocchi K, Lewis G. Machine learning estimation of heterogeneous treatment effects with instruments. *Adv Neural Inf Process Syst*. 2019;32. https://proceedings.neurips.cc/paper/2019/hash/3b2acfe2e38102074656ed938abf4ac3-Abstract.html.
23 Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581–590.
24 Dudbridge F. Polygenic Mendelian randomization. *Cold Spring Harb Perspect Med*. 2021;11(2):a039586.
25 Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med*. 2016;35(11):1880–1906.
26 Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol*. 2013;42(4):1134–1144.
27 Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience*. 2019;8(7):giz082.
28 Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. 2020;36(22-23):5424–5431.
29 Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. *Developing a protocol for observational comparative effectiveness research: a user's guide*. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.
30 Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Nonparametric tests for treatment effect heterogeneity. *Rev Econ Stat*. 2008;90(3):389–405.
31 Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113(27):7353–7360.
32 Zhao K, So H. *Chinese university of Hong Kong. Graduate School. Division of Biomedical Sciences, degree granting institution. Research on machine learning for drug discovery and precision medicine [dissertation]*. Chinese University of Hong Kong; 2020.
33 Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med*. 2018;37(11):1767–1787.

34 Chen J, Chen DL, Lewis G. Mostly harmless machine learning: learning optimal instruments in linear IV models. *arXiv preprint*. 2020. arXiv:2011.06158.
35 Mayer M, Mayer MM. *Package 'missRanger'*. R package; 2019.
36 Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45(11):1274–1283.
37 A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47(10):1121–1130.
38 Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):e1608.
39 Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759–2772.
40 Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio. *Genetics*. 2018;208(4):1397–1408.
41 Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet*. 2013;9(2):e1003264.
42 Jellinger PS. American association of clinical endocrinologists/american college of endocrinology management of dyslipidemia and prevention of cardiovascular disease clinical practice guidelines. *Diabetes Spectr*. 2018;31(3):234.
43 Okamura T, Tanaka H, Miyamatsu N, et al. The relationship between serum total cholesterol and all-cause or cause-specific mortality in a 17.3-year study of a Japanese cohort. *Atherosclerosis*. 2007;190(1):216–223.
44 Richter M. *Associations between plasma fatty acids, dietary fatty acids and cardiovascular risk factors: the PURE study*. North-West University, Potchefstroom Campus; 2014.
45 Wakabayashi I, Daimon T. Comparison of discrimination for cardio-metabolic risk by different cut-off values of the ratio of triglycerides to HDL cholesterol. *Lipids Health Dis*. 2019;18:1–10.
46 Staiger D, Stock JH. *Instrumental variables regression with weak instruments*. National Bureau of Economic Research Technical Working Paper Series; 1994:151.
47 Hausman JA. Specification tests in econometrics. *J Econom Soc*. 1978;46:1251–1271.
48 Zeileis A, Fox J, Kleiber C. *ivreg: two-stage least-squares regression with diagnostics*. R Package; 2021.
49 Bascle G. Controlling for endogeneity with instrumental variables in strategic management research. *Strateg Organ*. 2008;6(3):285–327.
50 Ference BA, Ginsberg HN, Graham I, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur Heart J*. 2017;38(32):2459–2472.
51 Graham SE, Clarke SL, Wu KH, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. 2021;600(7890):675–679.
52 Sandfort V, Lai S, Ahlman MA, et al. Obesity is associated with progression of atherosclerosis during statin treatment. *J Am Heart Assoc*. 2016;5(7):e003621.
53 Kjeldsen SE. Hypertension and cardiovascular risk: general aspects. *Pharmacol Res*. 2018;129:95–99.
54 Kannel WB. Blood pressure as a cardiovascular risk factor: prevention and treatment. *JAMA*. 1996;275(20):1571–1576.
55 Lafeber M, Spiering W, van der Graaf Y, et al. The combined use of aspirin, a statin, and blood pressure–lowering agents (polypill components) and the risk of vascular morbidity and mortality in patients with coronary artery disease. *Am Heart J*. 2013;166(2):282–289.e1.
56 Costanzo P, Perrone Filardi P, Petretta M, et al. Impact of gender in primary prevention of coronary heart disease with statin therapy: a meta-analysis. *J Am Coll Cardiol*. 2009;53:A210.
57 Cholesterol Treatment Trialists' (CTT) Collaboration. Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174 000 participants in 27 randomised trials. *Lancet*. 2015;385(9976):1397–1405.
58 Jiang S, Jiang D, Tao Y. Role of gamma-glutamyltransferase in cardiovascular diseases. *Exp Clin Cardiol*. 2013;18(1):53.
59 Ndrepepa G, Kastrati A. Gamma-glutamyl transferase and cardiovascular disease. *Ann Transl Med*. 2016;4(24):481–495.
60 Larsson SC, Burgess S, Michaëlsson K. Association of genetic variants related to serum calcium levels with coronary artery disease and myocardial infarction. *JAMA*. 2017;318(4):371–380.
61 Rohrmann S, Garmo H, Malmström H, et al. Association between serum calcium concentration and risk of incident and fatal cardiovascular disease in the prospective AMORIS study. *Atherosclerosis*. 2016;251:85–93.
62 Legarth C, Grimm D, Krueger M, Infanger M, Wehland M. Potential beneficial effects of vitamin d in coronary artery disease. *Nutrients*. 2019;12(1):99.
63 Bahrami LS, Ranjbar G, Norouzy A, Arabi SM. Vitamin D supplementation effects on the clinical outcomes of patients with coronary artery disease: a systematic review and meta-analysis. *Sci Rep*. 2020;10(1):12923.
64 Dixon P, Hollingworth W, Harrison S, Davies NM, Smith GD. Mendelian randomization analysis of the causal effect of adiposity on hospital costs. *J Health Econ*. 2020;70:102300.
65 Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ*. 2020;368:m131.
66 Lin J, Mars N, Fu Y, et al. Integration of biomarker polygenic risk score improves prediction of coronary heart disease. *JACC Basic Transl Sci*. 2023;8(12):1489–1499.
67 Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J*. 2015;36(9):539–550.
68 Prats-Uribe A, Sayols-Baixeras S, Fernández-Sanlés A, et al. High-density lipoprotein characteristics and coronary artery disease: a Mendelian randomization study. *Metab Clin Exp*. 2020;112:154351.
69 Lee SH, Lee J, hui Kim G, et al. Two-sample mendelian randomization study of lipid levels and ischemic heart disease. *Korean Circ J*. 2020;50(10):940–948.
70 C Reactive Protein Coronary Heart Disease Genetics Collaboration, (CCGC). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*. 2011;342:d548.
71 Kuppa A, Tripathi H, Al-Darraji A, Tarhuni WM, Abdel-Latif A. C-reactive protein levels and risk of cardiovascular diseases: a two-sample bidirectional Mendelian randomization study. *Int J Mol Sci*. 2023;24(11):9129.
72 Larsson SC, Michaëlsson K, Burgess S. IGF-1 and cardiometabolic diseases: a Mendelian randomisation study. *Diabetologia*. 2020;63:1775–1782.
73 Jing Z, Hou X, Wang Y, et al. Association between insulin-like growth factor-1 and cardiovascular disease risk: evidence from a meta-analysis. *Int J Cardiol*. 2015;198:1–5.
74 Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol*. 2004;33(1):30–42.
75 Davies NM, Holmes MV, Smith GD. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018;362:k601.
76 Staley JR, Burgess S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genet Epidemiol*. 2017;41(4):341–352.