OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf214
https://doi.org/10.1093/bib/bbaf214
Problem Solving Protocol

Accounting for the impact of rare variants on causal inference with RARE: a novel multivariable Mendelian randomization method

Yu Cheng[1,2,‡], Xinjia Ruan[1,‡], Xiaofan Lu[3], Yuqing Yang[1], Yuhang Wang[1], Shangjin Yan[4], Yuzhe Sun[5], Fangrong Yan[1,*], Liyun Jiang[1,*], Tiantian Liu[1,*]

[1]Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, #639 Longmian Ave, Jiangning District, Nanjing 211100, Jiangsu, China
[2]Department of Bioinformatics and Computational Biology, The University of Texas, M.D. Anderson Cancer Center, #7007 Bertner Ave, Texas Medical Center, Houston 77030, TX, United States
[3]Department of Cancer and Functional Genomics, Institute of Genetics and Molecular and Cellular Biology, CNRS/INSERM/UNISTRA, #10142 BP, Illkirch 67400, Bas-Rhin, France
[4]High School Affiliated to Nanjing Normal University, #37 Chahar Road, Gulou District, Nanjing 210003, Jiangsu, China
[5]Department of Biochemistry, Vassar college, Poughkeepsie, NY 12604, United States

*Corresponding authors. Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, #639 Longmian Ave, Jiangning Districst, Nanjing 211100, Jiangsu, China.
E-mail: 1020222675@cpu.edu.cn (Tiantian Liu), ljiang.cpu@foxmail.com (Liyun Jiang), f.r.yan@163.com (Fangrong Yan)
‡Yu Cheng and Xinjia Ruan contributed equally to this work.

Abstract

Mendelian randomization (MR) method utilizes genetic variants as instrumental variables to infer the causal effect of an exposure on an outcome. However, the impact of rare variants on traits is often neglected, and traditional MR assumptions can be violated by correlated horizontal pleiotropy (CHP) and uncorrelated horizontal pleiotropy (UHP). To address these issues, we propose a multivariable MR approach, an extension of the standard MR framework: MVMR incorporating **R**are variants **A**ccounting for multiple **R**isk factors and shared horizontal pl**E**iotropy (**RARE**). In the simulation studies, we demonstrate that RARE effectively detects the causal effects of exposures on outcome with accounting for the impact of rare variants on causal inference. Additionally, we apply RARE to study the effects of high density lipoprotein and low density lipoprotein on type 2 diabetes and coronary atherosclerosis, respectively, thereby illustrating its robustness and effectiveness in real data analysis.

Keywords: MVMR; horizontal pleiotropy; rare variants; polygenic risk scores; coronary atherosclerosis

Introduction

Traditional Mendelian randomization (MR), typically referred to as univariable MR, is an epidemiological approach that utilizes genetic variants as instrumental variables (IVs) to infer the causal effect of an exposure on an outcome [1, 2], provided that fundamental assumptions are satisfied (Fig. 1A) [3]. Recently, MR has become a powerful tool in epidemiology research, leveraging numerous publicly available genome-wide association studies (GWASs) summary-level data from large-scale cohort studies or prospective studies [4–6].

In practical MR analysis, researchers rigorously screen IVs to ensure compliance with fundamental assumptions, a process that frequently entails the exclusion of rare variants [7]. Although rare variants typically account for only a fraction of the heritability of traits, their modest effects still play an important role in trait expression [8–10]. However, there is no existing MR method accounts for the effects of rare SNPs, which can introduce estimation bias in causal effect analysis [11–13]. This challenge arises because low allele frequencies make it difficult for studies with conventional sample sizes to provide effective estimates, resulting in greater uncertainty. These increased variances reduce statistical power and can compromise the precision and reliability of causal effect estimates when rare variants are directly utilized in MR analyses. Moreover, the fundamental assumptions may not always hold in real-data cases and the wide-spread pleiotropy, including correlated horizontal pleiotropy (CHP) and uncorrelated horizontal pleiotropy (UHP), poses significant challenges [14, 15]. Specifically, UHP is a phenomenon where variants affect outcome via other pathways not through the exposure, while CHP describes scenarios where variants affect both exposure and outcome through a heritable shared factor, commonly to be unmeasured confounders (Fig. 1A) [15].

With the explosion in the size and scale of data sets from high-throughput experiments and concomitant genetic data [16–18], UHP can be partly addressed by extending univariable MR method to multivariable Mendelian randomization (MVMR) framework [19]. MVMR incorporates multiple exposures into the model, which is useful when a set of related risk factors for the outcome of interest share many commonly associated genetic

Received: September 11, 2024. **Revised:** April 10, 2025. **Accepted:** April 20, 2025

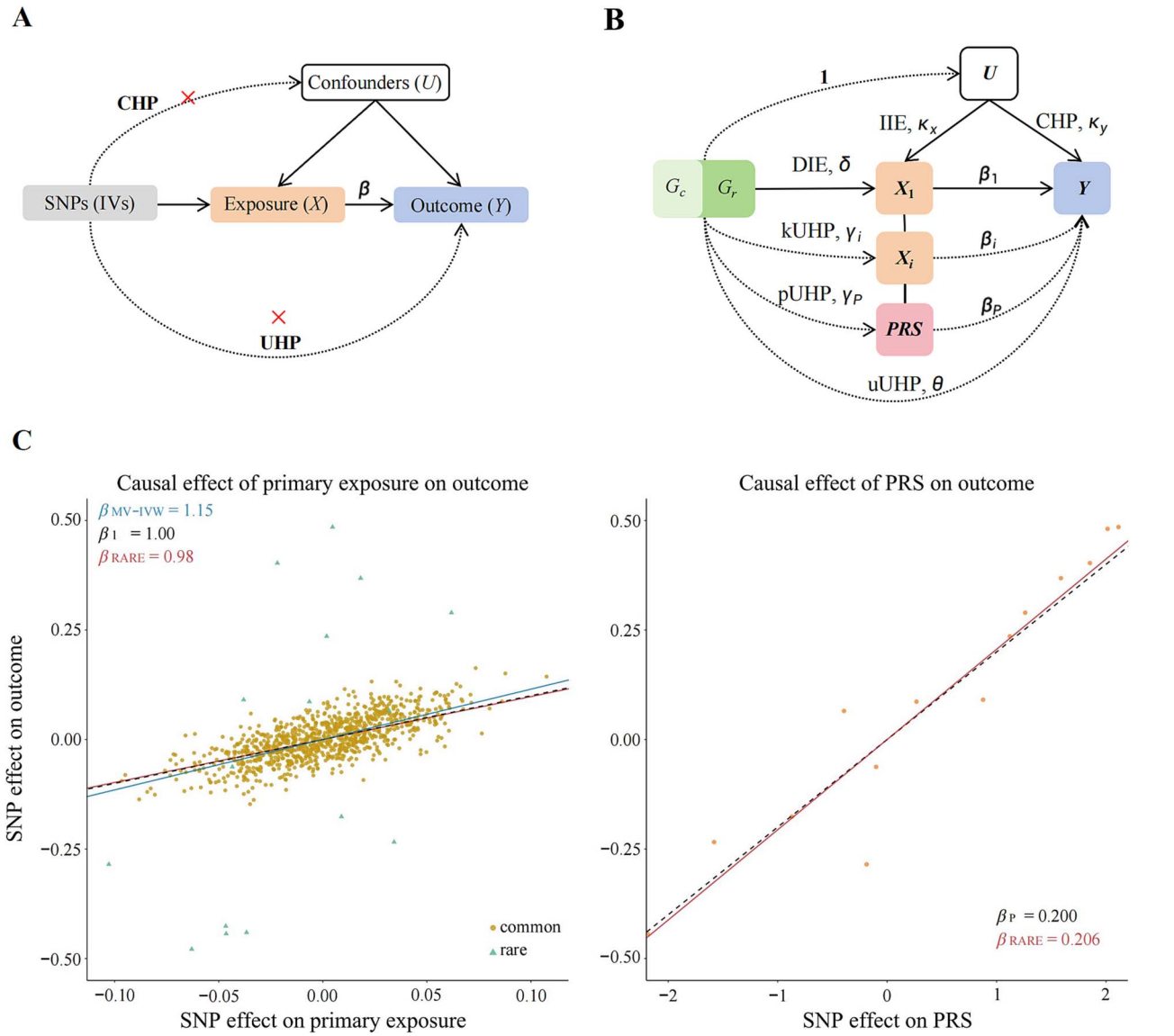

Figure 1. Causal diagram of classic MR and RARE, with an illustrative example. (A) Causal diagram assumed by the classic MR methods. The fundamental assumptions are: IVs are associated with the exposure (relevance); IVs affect the outcome through the exposure only (exclusion restriction); the exposure and the outcome are not confounded by unmeasured confounders (exchangeability). (B) An illustration of RARE model. RARE allows the IVs to be directly associated with the primary exposure directly (DIE) and indirectly through the unmeasured confounders (IIE), as well as to affect the outcome via the unmeasured confounders (CHP). RARE decomposes the UHP into three parts: Know pathways (kUHP), unknown pathways (uUHP), and pathways mediated by the rare variants (pUHP). The primary goal of RARE is to estimate the causal effect of the primary exposure on the outcome (β_1). (C) A simulated example illustrates the causal effect of the primary exposure on the outcome, estimated by MV-IVW and RARE, with the true causal effect $\beta_1 = 1$ (left panel), and the impact of the rare variants on outcome, with the true effect $\beta_P = 0.2$ (right panel).

variants [20]. For example, when considering the causal effect of high-density lipoprotein (HDL) on cardiovascular disease (CAD), it is natural to include low-density lipoprotein (LDL) and triglyceride (TG) to account for part of their UHP. MVMR can also relax the fundamental assumptions by allowing IVs to be associated with unmeasured confounders, thereby addressing the CHP [21]. Despite the proposal of several MVMR methods, such as MV-IVW [19], MV-Egger [22] and MV-LASSO [23], they can only address certain aspects of the challenges posed by MVMR. To be specific, type I error rate of MV-IVW inflates when IVs exit CHP and UHP; MV-Egger and MV-LASSO can only provide reliable estimation for moderate to high levels of UHP, with MV-LASSO particularly effective at selecting valid IVs by applying a penalty to the model. Additionally, a recently developed method, GRAPPLE [24], which models both weak IVs and strong IVs, has well expanded the applicability of MVMR. However, like the previously mentioned methods, GRAPPLE does not consider the potential impact of rare variants on causal estimation either.

With these limitations in mind, we propose a novel MVMR method, called RARE (MVMR incorporating **R**are variants **A**ccounting for multiple **R**isk factors and shared horizontal pl**E**iotopy). RARE focuses on one exposure of interest (the primary exposure) and considers other exposures (the secondary exposures) as partial potential UHP pathways. To effectively estimate the causal effects of the primary exposures on the outcome, RARE then integrates shared eligible rare variants and common variants. Considering the potential low power of the rare variants, we combine them into polygenic risk scores (PRS) [25, 26] and treat it as a quasi-exposure. Furthermore, RARE models the complex IV-to-primary exposure structure to account for CHP for more accurate estimation. Specifically, RARE allows each IV to directly affect the primary exposure (direct IV-to-exposure

effect, DIE) and/or indirectly affect the primary exposure via unmeasured confounders (indirect IV-to-exposure effect, IIE), and IV to affect outcome via unmeasured confounders (CHP). Moreover, RARE accommodates linkage disequilibrium among IVs, which helps mitigate potential bias in the presence of weak instruments and potentially boost statistical power by increasing the number of eligible IVs.

To demonstrate the advantages of RARE, we conduct experiments using both simulated data across various common scenarios and real data, comparing it with existing approaches. The results illustrate the superior performance of RARE over other existing MVMR methods.

Results

Overview of RARE

RARE is a novel MVMR method that simultaneously accounts for DIE, IIE, CHP, and UHP. In this method, we consider one exposure as the primary exposure and treat all other exposures as the secondary exposures, partially explaining UHP. As in standard MR, the UHP can be decomposed into UHP via known pathways (kUHP) and UHP via unknown pathways (uUHP). For example, when investigating causal associations between HDL and cardiovascular diseases, LDL can be seen as the kUHP, and other potential heritable risk factors can be regarded as the uUHP [9, 20]. To further account for the impact of rare variants on causal effect estimation, we integrate them into a quasi-variable [27] (PRS) that acts as an secondary exposure to address the potential issue of insufficient statistical power associated with rare variants [26].

As depicted in Fig. 1B, we assume that we have obtained a set of IVs denoted as $\boldsymbol{G} = \{\mathbf{IV}_1, \ldots, \mathbf{IV}_p\}$, containing p eligible IVs, of which n_c are common SNPs and n_r are rare SNPs, simultaneously correlated with the K exposures under the preset P value threshold. For the j-th IV, we allow the presence of DIE (δ_j), IIE (κ_{x_j}) and CHP (κ_{y_j}) for the primary exposure, as well as the potential uUHP (θ_j) [21, 28]. Based on the above definitions, the IV-to-outcome effect can be written as:

$$\Gamma_j = \beta_1 \left(\delta_j + \kappa_{xj}\right) + \sum_{i=2}^{K} \beta_i \gamma_{ij} + \beta_P \gamma_{Pj} + \kappa_{yj} + \theta_j, \tag{1}$$

where β_1 is the causal effect of the primary exposure on the outcome, and β_2 to β_K are the causal effects of the secondary exposures on outcome, respectively; β_P is the potential effect of PRS on outcome. For definitions of γ_{ij} and γ_{Pj} see the **Methods** section. In accordance with the previous literature [14, 15], we define $\gamma_{1j} = \delta_j + \kappa_{xj}$ to capture both DIE and IIE for the primary exposure. Thus, Eq. (1) can be expressed as follows:

$$\Gamma_j = \sum_{i=1}^{K} \beta_i \gamma_{ij} + \beta_P \gamma_{Pj} + \kappa_{yj} + \theta_j. \tag{2}$$

Finally, RARE applies Gibbs sampling algorithm to estimate the parameters based on a Bayesian hierarchical model. The details are provided in **Methods** and Supplementary Note 1.

In Fig. 1C, we illustrate RARE and MV-IVW using a simulated data example to assess the causal effect of the primary exposure on the outcome, and we further evaluate the effect of rare SNPs on the outcome. We set the true effect of the primary exposure and the PRS on the outcome at $\beta_1 = 1$ and $\beta_P = 0.2$, respectively; with the CHP and UHP explaining 5% of the outcome Y heritability, respectively. RARE ultimately yields an estimate of the primary exposure on the outcome very close to the true value, which is 0.98, while the MV-IVW method produces an unacceptably biased estimation of 1.15. Additionally, the RARE method also estimates the impact of the variants on the outcome, with an effect size of 0.206. All results indicate that RARE effectively models the complex IV-to-exposure and the IV-to-outcome structures, utilizing both common SNPs and rare SNPs, and clarifies the impact of the rare SNPs in causal inference.

RARE estimates the causal effects and reduces false positives

We conduct simulation studies to evaluate and compare the performance of RARE with existing MVMR methods across a variety of scenarios. In all main scenarios, the number of exposures in Eq. (2) is set to be two, that is $K = 2$, where one exposure acts as the primary exposure and the other as the secondary exposure. Moreover, the exposures and the outcome are generated based on Eq. (7) and Eq. (8). To ensure consistency across methods, rare variants were incorporated into the analysis via the polygenic risk score (PRS) for all methods, including RARE.

We compare RARE with GRAPPLE, MV-IVW, MV-Egger, and MV-LASSO. In existing MVMR methods, MV-IVW and MV-LASSO are not designed to address UHP and CHP. Additionally, MV-LASSO involves the selection of IVs, which might neglect the interdependence among these IVs, thereby potentially compromising the precision of causal effect estimation. MV-Egger assumes that no IV is affected by CHP but allows IVs to have UHP effects, while GRAPPLE considers the pleiotropy effects from unknown pathway. Moreover, existing methods are unable to handle the numerical characteristics (such as mean, variance, and distribution) of the rare variants [29, 30]. The proposed RARE, however, allows all IVs to have both UHP and CHP, and accounts for the characteristics of rare variants into account. It is worth noting that to better evaluate the effect of introducing rare variants, we compare the performance of all methods both with and without the inclusion of rare variants. Specifically, methods that did not include rare variants were labeled as 'N-'.

We first evaluate the performance of type I error rate control (Fig. 2A) for all competing methods under different conditions of linkage disequilibrium (r_{LD}) and correlation between IIE and CHP (ρ_κ). We control the combinatorial values for heritability due to UHP and CHP, denoted as h^2_θ and $h^2_{\kappa_y}$, and we let $h^2_\theta = h^2_{\kappa_y} = h^2$. The results show that, with the increment of h^2, RARE demonstrates a substantial improvement in type I error rate controlling when rare variants are incorporated (RARE v.s. N-RARE). This highlights the necessity of considering rare variants in analysis. Additionally, RARE is not sensitive to low-to-moderate LD ($r_{LD} = 0.2$ v.s. $r_{LD} = 0.6$), which suggests that RARE can eliminate the need for IV selection in most applications, thereby enhancing the statistical power by increasing the number of eligible IVs. However, researchers should perform LD clumping to minimize potential bias when severe linkage disequilibrium is observed in the LD matrix calculated using the R package *LDlinkR* [31]. Moreover, RARE may suffer from increased levels of type I error rates when the correlation between IIE and CHP becomes stronger ($\rho_\kappa = 0.2$ v.s. $\rho_\kappa = 0.5$). In comparison, MV-Egger fails to account for CHP, leading to inflated type I error rates. However, type I error rates of MV-Egger are consistently lower than those of MV-IVW and MV-LASSO, as these methods account for neither UHP nor CHP. Additionally, since GRAPPLE, MV-IVW, MV-Egger and MV-LASSO do not consider the characteristics of rare variants, their performance remains unaffected by the inclusion of these variants. We also perform simulations of r_{LD} and ρ_κ over a broader range of variation. The results were largely similar that RARE can sharply control the type

I error rates when rare variants are incorporated into the analysis, and additional details can be found in Supplementary Note 2.1 and Supplementary Note 2.2.

We then compare the power of each method by varying $h^2_{\gamma_1}$, here $h^2_{\gamma_1} = h^2_\delta + h^2_{\kappa_x}$ and $h^2_\delta = 4h^2_{\kappa_x}$, while fixing $h^2_\theta = 0.02$, $h^2_{\kappa_y} = 0.05$, $h^2_{\gamma_2} = 0.02$, $r_{LD} = \rho_\kappa = 0.2$ (Fig. 2B). RARE achieves comparable statistical power with a low heritability of $h^2_{\gamma_1}$ while effectively controlling the type I error rate. Additionally, we set $\beta_1 = 1$ and calculate the estimation bias. As shown in Fig. 2C, all methods except RARE obtain unacceptable bias, which could partly explain their type I error inflation and high statistical power.

The conclusions remain qualitatively unchanged across various considerations, including the number of rare variants, the sample size of individual-level data, the correlation between exposures, the heritability magnitude of IIE and CHP, the PRS effect, the scale of exposures, and the different definitions of rare and common variants. Additional details are provided in Supplementary Notes 2.3 – 2.9. Furthermore, to assist researchers in better utilizing RARE, we provided a reference figure detailing its computational performance under different numbers of SNPs and exposures (Supplementary Note 2.10).

Two real-data examples

We illustrate the performance of RARE and other methods utilizing two real data examples. Genetic associations with the exposures and the outcomes are all obtained from large-scale GWAS studies. For each method, we consider the SNPs associated with both exposures at the genome-wide level of significance (P **value** $< 5 \times 10^{-8}$) as eligible IVs; SNPs with minor allele frequency (MAF) over 0.05 are defined as common variants, and those with MAF less than 0.01 are defined as rare variants [29]. Data for common variants are from GWAS summary statistics. To calculate the PRS using Eq. (3), we utilize genotype data simulated from summary statistics of GWAS studies and individual-level genotype data from the UK Biobank, respectively; and then we calculate the associations between PRS and eligible IVs by performing univariable regression.

HDL and LDL play significant roles in the pathogenesis of T2D and CA. High level of HDL is associated with improved insulin sensitivity and helps reduce atherosclerosis risk. However, LDL plays the opposite roles. Here, we select HDL and LDL as exposures, and T2D and CA as outcomes, respectively, and apply RARE and other methods to evaluate the causal effects.

We first set HDL and LDL as the primary exposure and the secondary exposure, respectively, to estimate the causal effects of them on T2D with the mentioned methods (Fig. 3A). As shown in Table 1, RARE identifies a significant putative negative causal effect of HDL on T2D ($\beta_{\text{RARE,HDL}} = -0.219, P$ **value** $= 0.003$), while MV-Egger obtains an opposite sign compared with other methods. Notably, only RARE reports a significant putative positive causal effect of LDL on T2D ($\beta_{\text{RARE,LDL}} = 0.264, P$ **value** $=$ 0.007), and all methods declare that rare variants significantly contribute to inferring the causal effect of the primary exposure on the outcome. Additionally, we illustrate the associations between HDL and LDL with T2D. When plotting IV-to-HDL effects against IV-to-T2D effects (Fig. 3B left panel), there is a negative causal relationship for the common variants (yellow points and yellow line) while rare variants entail a different pattern with an opposite slope (green points and green line). RARE then integrates the effects of common and rare variants to ultimately provide the causal effect of HDL on T2D (black line) and declares the impact of rare variants on the causal effect ($\beta_{\text{RARE,PRS}} = 0.313, P$ **value** < 0.001). Moreover, the right panel of Fig. 3B illustrates the causal relationship between LDL and T2D. We then reset LDL as the primary exposure with HDL as the secondary exposure, RARE is the only method that reports a putative significant positive causal association between LDL and T2D, without indicating any additive effect of the rare variants (Supplementary Table 1, Supplementary Fig. 11A). Similarly, when evaluating the causal effects between HDL and LDL on CA, RARE indicates that both common and rare variants exhibit a consistent pattern in inferring the causal effect of the primary exposure on CA. This indicates that rare variants do not introduce additional impact to the inference, regardless of whether HDL is the primary exposure (Table 2, Fig. 3A, Fig. 3C) or the secondary exposure (Supplementary Table 2, Supplementary Fig. 11B). We further compare the performance of RARE using both simulated and real genotype data, finding that both approaches yield comparable results. Therefore, when real genotype data is unavailable, simulated genotype data can be effectively used to calculate the associations between PRS and eligible SNPs (Supplementary Fig. 12).

Discussion

The central focus of this study is to address and mitigate the impact of rare variants on causal inference. To this end, we propose RARE, a novel MVMR method. For the first time, RARE effectively tackles potential estimation bias introduced by rare variants while simultaneously accounting for the pervasive presence of horizontal pleiotropy.

By consolidating the effects of individual rare variants into a single, more robust metric—the polygenic risk score—RARE enhances the identifiability of rare variants. The PRS captures the collective influence of rare variants, serving as a powerful tool to better understand how these less common genetic variations contribute to trait variability, which might be overlooked in traditional analyses. Furthermore, RARE effectively captures the complex IV-to-primary exposure structure, as well as CHP effect, which can be impacted by both genetic and environmental factors, thereby providing a more accurate estimate of causal effect. RARE also addresses partial UHP through other exposures, enhancing its robustness. This is particularly important when the primary exposure shares some genetic associations with other secondary exposures, such as various lipids like HDL with LDL and TG. Moreover, RARE does not require the pruning of IVs, thus allowing for the presence of linkage disequilibrium among them and indirectly increasing the number of eligible IVs. All these features contribute to reducing the estimation bias of primary exposure-outcome causal effect, thereby controlling the type I error rate and boosting the statistical power.

We employ a series of simulated studies mimicking common scenarios in genetic epidemiology to demonstrate the advantages of RARE. Results from the simulation studies indicate that, compared with the existing MVMR methods (GRAPPLE [24], MVMR-IVW [19], MVMR-Egger [22], and MVMR-LASSO [23]), RARE consistently achieves superior control in type I error rate, since the other methods can only partially address issues related to horizontal pleiotropy and linkage disequilibrium among IVs. Additionally, although we primarily present scenarios with two exposures in the main manuscript, we have also extended our model to accommodate multiple exposures (up to 10), demonstrating RARE's robustness across different number of exposures.

We subsequently apply RARE to identify the causal effects of HDL and LDL on T2D [32–35] and CA [36–39], respectively. Given that the causal relationships between these traits are

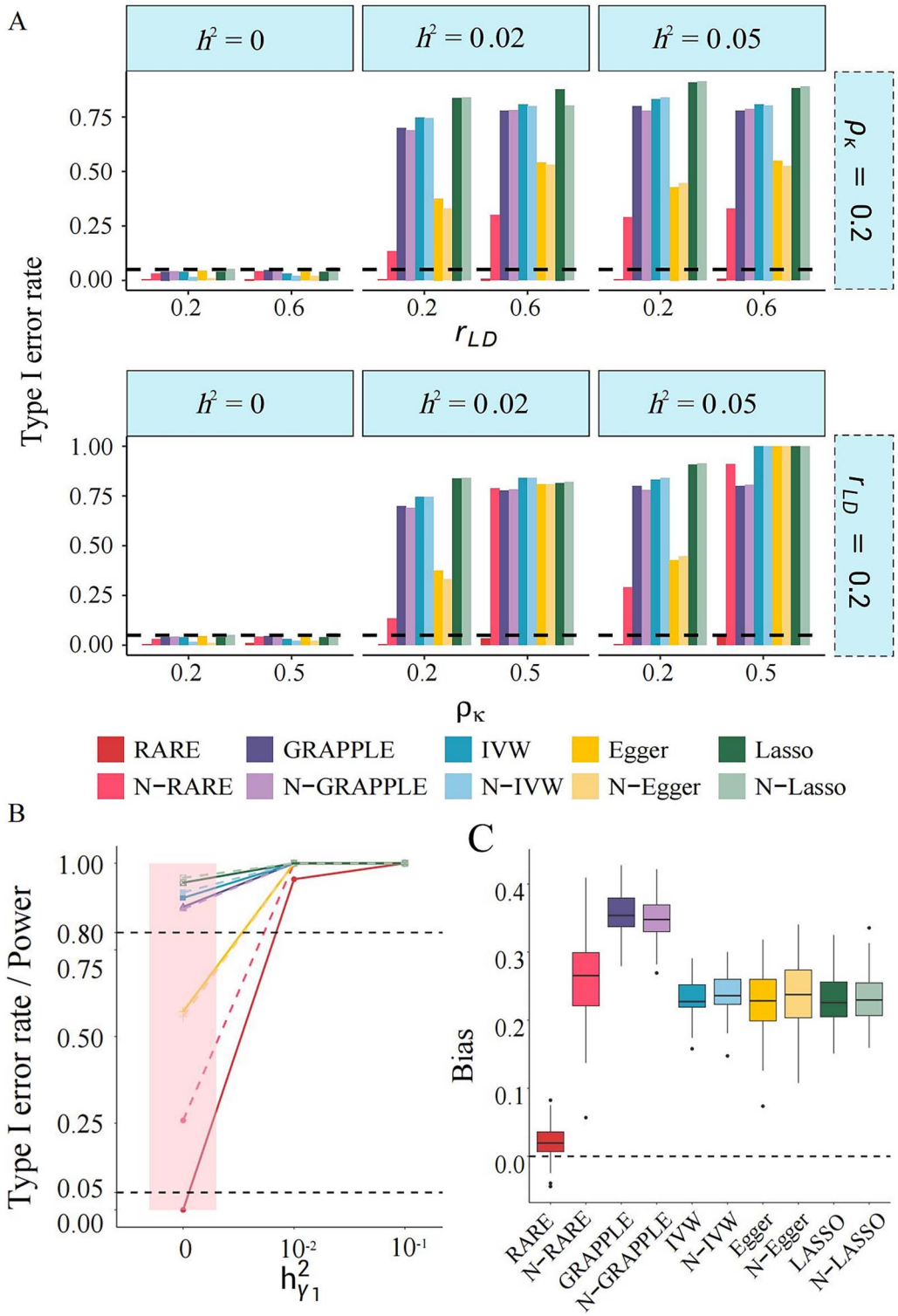

Figure 2. Comparison of RARE and other methods in simulation studies. (A) Comparison of type I error rates for RARE and other methods under combinatorial settings for r_{LD} and h^2 with $\rho_\kappa = 0.2$ (top pattern); and combinatorial settings for ρ_κ and h^2 with $r_{LD} = 0.2$ (bottom pattern). (B) Comparison of powers for RARE and other methods under the settings $r_{LD} = 0.2$, $\rho_\kappa = 0.2$, $h^2_\theta = 0.02$, $h^2_{\kappa_y} = 0.05$ and $h^2_{\gamma_2} = 0.02$. (C) Boxplots of point estimates over 100 replications for competing methods. The bounds of the boxes represent the 25th and 75th quantiles, with the center lines indicating the median values.

well-established, they provide a valuable benchmark to illustrate RARE's performance in real-world data scenarios. Specifically, HDL plays a crucial role in reverse cholesterol transport, facilitating the removal of cholesterol from arterial walls and mitigating plaque formation, thereby protecting against CA [36]. Additionally, HDL exhibits anti-inflammatory and antioxidant properties, which improve insulin sensitivity and lower the risk of T2D [33]. In contrast, elevated LDL levels contribute to the development of atherosclerotic plaques, leading to arterial narrowing and a higher risk of cardiovascular events [38]. Although the relationship

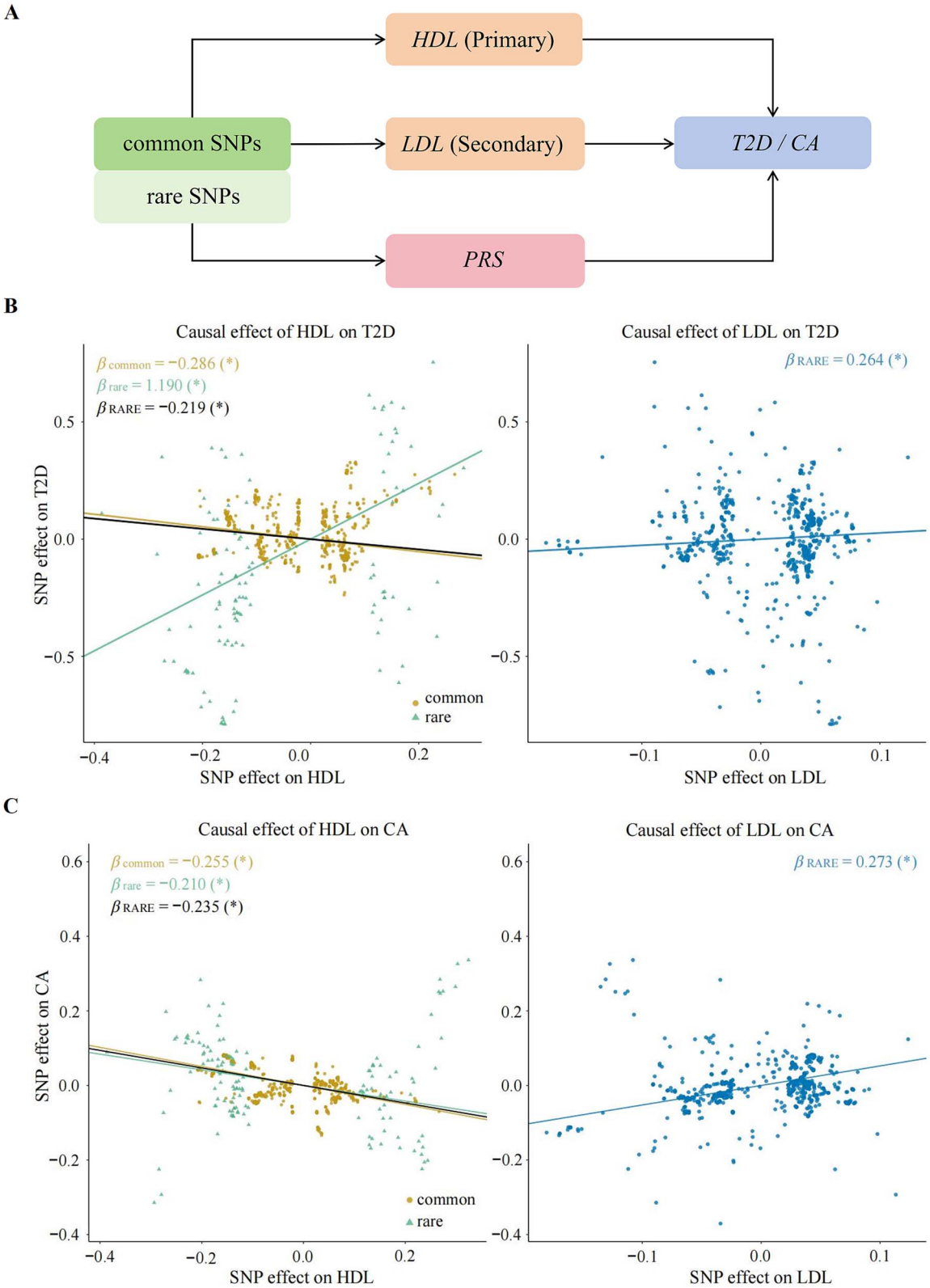

Figure 3. Two real data examples. (A) Data analysis diagram. HDL is set as the primary exposure, LDL as the secondary exposure, with T2D and CA as the respective outcomes. PRS is calculated utilizing the rare SNPs of HDL. (B), (C) Illustrations of the causal effect of HDL (primary exposure) and LDL (secondary exposure) on T2D and CA, respectively. Left panels are the causal effects of the HDL on T2D and CA, respectively; and the right panels are the causal effects of LDL on T2D and CA, respectively.

Table 1. Summary results for evaluating the causal effects of HDL (primary exposure) and LDL (secondary exposure) on T2D, considering the impact of rare variants.

Method	HDL (Primary)	LDL (Secondary)	PRS
RARE	−0.219 (−0.379, −0.061); **0.003**	0.264 (0.052, 0.476); **0.007**	0.313 (0.163, 0.462); < **0.001**
GRAPPLE	−0.327 (−0.426, −0.227); < **0.001**	0.186 (−0.015, 0.387); 0.069	0.377 (0.256, 0.497); < **0.001**
MV-IVW	−0.260 (−0.353, −0.167); < **0.001**	0.162 (−0.024, 0.348); 0.088	0.349 (0.217, 0.480); < **0.001**
MV-Egger	0.017 (−0.145, 0.180); 0.836	0.140 (−0.044, 0.325); 0.137	0.334 (0.203, 0.466); < **0.001**
MV-LASSO	−0.483 (−0.571, −0.396); < **0.001**	0.149 (−0.-14, 0.311); 0.072	0.427 (0.313, 0.540); < **0.001**
N-RARE	−0.115 (−0.267, 0.036); 0.136	0.260 (−0.045, 0.566); 0.095	–
N-GRAPPLE	−0.257 (−0.357, −0.156); < **0.001**	0.193 (−0.013, 0.400); 0.066	–
N-MVIVW	−0.223 (−0.317, −0.129); < **0.001**	0.162 (−0.027, 0.352); 0.093	–
N-MVEgger	0.086 (−0.077, 0.248); 0.300	0.136 (−0.051, 0.323); 0.155	–
N-LASSO	−0.370 (−0.456, −0.288); < **0.001**	0.142 (−0.020, 0.305); 0.085	–

Results are represented as 'point estimation of β (95% confidence interval); P value'.

Table 2. Summary results for evaluating the causal effects of HDL (primary exposure) and LDL (secondary exposure) on CA, considering the impact of rare variants.

Method	HDL (Primary)	LDL (Secondary)	PRS
RARE	−0.235 (−0.330, −0.140); < **0.001**	0.273 (0.109, 0.437); < **0.001**	0.026 (−0.011, 0.063); 0.161
GRAPPLE	−0.229 (−0.261, −0.198); < **0.001**	0.435 (0.372, 0.498); < **0.001**	0.016 (−0.005, 0.037) 0.132
MV-IVW	−0.193 (−0.220, −0.166); < **0.001**	0.405 (0.352, 0.458); < **0.001**	0.009 (−0.028, 0.046); 0.636
MV-Egger	−0.253 (−0.299, −0.206); < **0.001**	0.410 (0.357, 0.463); < **0.001**	0.017 (−0.021, 0.054); 0.374
MV-LASSO	−0.326 (−0.343, −0.310); < **0.001**	0.493 (0.452, 0.534); < **0.001**	0.035 (0.015, 0.055); **0.001**
N-RARE	−0.196 (−0.276, −0.116); < **0.001**	0.287 (0.098, 0.476); **0.001**	–
N-GRAPPLE	−0.224 (−0.253, −0.194); < **0.001**	0.435 (0.372, 0.498); < **0.001**	–
N-MVIVW	−0.192 (−0.218, −0.166); < **0.001**	0.405 (0.352, 0.458); < **0.001**	–
N-MVEgger	−0.249 (−0.295, −0.203); < **0.001**	0.409 (0.356, 0.462); < **0.001**	–
N-LASSO	−0.317 (−0.334, −0.301); < **0.001**	0.486 (0.445, 0.527); < **0.001**	–

Results are represented as 'point estimation of β (95% confidence interval); P value'.

between LDL and T2D is complex and bidirectional, high LDL levels can exacerbate insulin resistance [34, 35]. RARE utilizes both simulated genotype and genotype data from the UK Biobank to effectively identify the causal effects of HDL and LDL on T2D, as well as CA, and highlights the additional role of rare variants in estimating causal effects. Compared to previous studies, although the RARE method can identify some novel causal effects, the causal effects between HDL and CAD as well as LDL and T2D still remain somewhat controversial. Possible reasons include that previous studies have typically not considered the impact of horizontal pleiotropy or adequately addressed confounding factors, nor have they accounted for rare variants, and all these factors can lead to biased conclusions. Additionally, like all MR studies, RARE's finding should be regarded as a more reliable approach than association studies for causal effect estimation, but not the golden standard of causal inference. To fully validate these results, large-scale randomized controlled trials are still necessary, with careful handling of rare variants as well as the elimination of confounding factors.

Despite the innovations introduced by RARE, some limitations warrant further investigation. First, while we assume that all IVs could have potential UHP effect and CHP effect, in reality, only a subset of IVs have CHP effect, which may lead to a biased estimation [14, 15]. Second, RARE requires multiple IVs including at least dozens of rare variants to identify the causal effect between primary exposure and outcome. In practical applications, obtaining a sufficient number of IVs may be challenging, thereby may limit the applicability of RARE. Third, RARE assumes a linear relationship between each exposure and the outcome, indicating that the absence of a linear effect does not imply a null causal effect [23, 40]. Non-linear relationships, such as U-shaped or J-shaped effects, are commonly observed among traits [41, 42]. We are working on improving the automation of this step. Finally, as our model does not account for sample overlapping between datasets, researchers should avoid selecting GWAS data from the same study as much as possible to mitigate the potential impact of sample overlap on causal effect estimation. For instance, if HDL GWAS data is sourced from the Global Lipids Genetics Consortium database [43], LDL GWAS data should be obtained from a different source.

Though emerging as a powerful tool for causal inference, it is essential to exercise caution. Various MR methods rely on their underlying assumptions and the quality of the genetic variants, as well as, more broadly, the quality of the GWAS data being utilized. While MR cannot entirely substitute traditional experimental studies, it is crucial for more accurately identifying pathogenic or protective factors for diseases, and employ various MR methods to estimate the same causal effect is worthwhile and warranted.

Methods

Data input for RARE

Suppose we have K exposures of interest, denoted as $X_1, X_2, \ldots, X_K$, and one outcome Y. We let X_1 to be the primary exposure, and $X_2, \ldots, X_K$ are considered as the secondary exposures. Our primary objective is to estimate the causal effect between the primary exposure (X_1) and the outcome (Y). Suppose there are p shared SNPs, among which n_c are common variants and n_r are rare variants, that are simultaneously associated with these K exposures, with P values below a given threshold from GWAS studies. Let γ_{ij} $(i = 1, 2, \ldots, K; j = 1, 2, \ldots, p)$ and Γ_j be the true marginal associations of the j-th IV with the i-th exposure X_i and outcome Y, respectively. The estimated effects and standard

errors of the associations between the j-th SNP and the i-th exposure and outcome from GWAS are denoted as $\left(\hat{\gamma}_{ij}, \hat{S}_{\gamma_{ij}}\right)$ and $\left(\hat{\Gamma}_j, \hat{S}_{\Gamma_j}\right)$, respectively.

Moreover, to account for the impact of rare variants, we integrate them by calculating PRS. For this purpose, we consider an extra independent individual-level genotype matrix $\boldsymbol{G}$ to calculate the **PRS** using the following formula:

$$\boldsymbol{PRS} = \sum_{k=1}^{n_r} \hat{\gamma}_k \boldsymbol{G}_k, \tag{3}$$

where $\hat{\gamma}_k$ is the coefficient from the GWAS data representing the associations between the k-th rare variants and the primary exposure; and $\boldsymbol{G}_k$ is a genotype vector representing the k-th rare variant from the matrix $\boldsymbol{G}$. Typically, each element in the genotype vector takes on a value of 0, 1, or 2. Note that the rare variants mentioned here specifically relate to be the primary exposure X_1, meaning that these rare SNPs are only required to be associated with the primary exposure, and are not necessarily associated with other secondary exposures. We finally treat the **PRS** as a quasi exposure and perform univariable regression of **PRS** onto each variant to calculate its effect and standard error, defined as $\left(\hat{\gamma}_{Pj}, \hat{S}_{\gamma_{Pj}}\right)$.

Decomposition of UHP effect

For univariable MR analysis, genetic variants affect outcome via other pathways other than exposure is defined as total UHP (tUHP). In the RARE model, we assume the tUHP can be decomposed into three parts: (1) variants affect outcome via known pathways (IVs → secondary exposures → Y, denoted as kUHP); (2) variants affect outcome via PRS (IVs → PRS → Y, denoted as pUHP); and (3) variants affect outcome via potential unknown pathways, denoted as uUHP. Therefore, according to Fig. 1B, Γ_j can be expressed as Eq. (1). Specifically, the second and the third part on the right of Eq. (1) correspond to the kUHP and pUHP, respectively. Without loss of generality, from hereon in, we focus on the case $K = 2$, although our RARE can be extended to higher dimensions (Supplementary Fig. 8). Thus, Eq. (2) can be further simplified as follows:

$$\Gamma_j = \beta_1\gamma_{1j} + \beta_2\gamma_{2j} + \beta_P\gamma_{Pj} + \kappa_{yj} + \theta_j = \boldsymbol{\beta}'\boldsymbol{\gamma_j} + \kappa_{yj} + \theta_j, \tag{4}$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_P)$ and $\boldsymbol{\gamma_j} = (\gamma_{1j}, \gamma_{2j}, \gamma_{Pj})'$.

RARE model for independent SNPs

With the assumptions that the GWAS data for the primary exposure, the secondary exposure and the outcome are from different studies and that all eligible SNPs are independent, once given the true effects, we can model the estimated effect vector $\left(\hat{\gamma}_{1j}, \hat{\gamma}_{2j}, \hat{\gamma}_{Pj}, \hat{\Gamma}_j\right)'$ using the following multivariable normal distribution:

$$\left(\hat{\gamma}_{1j}, \hat{\gamma}_{2j}, \hat{\gamma}_{Pj}, \hat{\Gamma}_j\right)' \sim N\left((\gamma_{1j}, \gamma_{2j}, \gamma_{Pj}, \Gamma_j)', \boldsymbol{\Sigma}_j\right), \tag{5}$$

where $\boldsymbol{\Sigma}_j = \mathrm{diag}\left(\hat{S}^2_{\gamma_{1j}}, \hat{S}^2_{\gamma_{2j}}, \hat{S}^2_{\gamma_{Pj}}, \hat{S}^2_{\Gamma_j}\right)$.

In RARE, IV-to-exposure effect of the main exposure is decomposed into DIE (δ_j) and IIE (κ_{xj}), where DIE and IIE are *i.i.d* with $\delta_j \sim N\left(0, \sigma^2_\delta\right)$ and $\kappa_{x_j} \sim N\left(0, \sigma^2_{\kappa_x}\right)$, respectively. Thus, the overall IV-to-exposure effect of the primary exposure γ_{1j} follows a normal distribution with mean $\delta_j + \kappa_{x_j}$. Moreover, the CHP (κ_{yj}) and uCHP (θ_j) similarly follow the prior distributions $\kappa_{yj} \sim N\left(0, \sigma^2_{\kappa_y}\right)$ and $\theta_j \sim N\left(0, \sigma^2_\theta\right)$, respectively. We further assume that the true IV-to-exposure effect of the secondary exposure (γ_{2j}) and PRS (γ_{Pj}) are *i.i.d* with $\gamma_{2j} \sim N\left(0, \sigma^2_{\gamma_2}\right)$ and $\gamma_{Pj} \sim N\left(0, \sigma^2_{\gamma_P}\right)$, respectively. Then, the IV-to-outcome effect Γ_j, given other parameters, can be written as follows:

$$\Gamma_j \mid \boldsymbol{\gamma}_j, \boldsymbol{\beta}, \kappa_{yj}, \sigma^2_\theta \sim N\left(\boldsymbol{\beta}'\boldsymbol{\gamma}_j + \kappa_{yj}, \sigma^2_\theta\right). \tag{6}$$

By combing Eq. (5) and Eq. (6), we build the Bayesian hierarchical model with conjugate priors for hyper parameters, $\sigma^2_{\kappa_y} \sim IG\left(a_{\kappa_y}, b_{\kappa_y}\right)$, $\sigma^2_\theta \sim IG\left(a_\theta, b_\theta\right)$, $\sigma^2_{\gamma_2} \sim IG\left(a_{\gamma_2}, b_{\gamma_2}\right)$, $\sigma^2_{\gamma_P} \sim IG\left(a_{\gamma_P}, b_{\gamma_P}\right)$.

Generation of simulated data

We construct individual-level structural model to obtain the input data of the proposed RARE. We first generate the genotype matrix $\boldsymbol{G}_s \in R^{n_s \times p}$ to generate the reference panel of the associations between the variants and the primary exposure. Similarly, we generate the genotype matrix $\boldsymbol{G}_i \in R^{n_i \times p}$ to calculate **PRS**, and then calculate the summary-level statistics of the associations between the variants and the primary exposure, the secondary exposure, the **PRS**, and the outcome, respectively. We should state that n_s and n_i are corresponding sample sizes and $p = n_c + n_r$ is the total number of eligible SNPs. For all simulations in the main manuscript, we set $n_s = n_i = 50,000$, $n_c = 1,000$, $n_r = 20$ and the number of confounders is set as $q = 20$.

To generate the genotype matrix $\boldsymbol{G}_s$, we first create a data matrix from a multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_r)$, where $\boldsymbol{\Sigma}_r(i,j) = |r|^{|i-j|}$ is a p-by-p dimensional matrix depicting the correlation between the i-th and j-th IV. We then simulate the genotype matrix by categorizing data matrix into dosage values according to MAF, that is uniformly distributed in [0.05, 0.5] for common SNPs and in [0.005, 0.01] for rare SNPs. Similar approaches can be used to obtain $\boldsymbol{G}_i$ as well. We then consider the following structural model to generate individual-level data for both $\boldsymbol{G}_s$ and $\boldsymbol{G}_i$:

$$\begin{aligned}
\boldsymbol{X}_1 &= \boldsymbol{G}\boldsymbol{\delta} + \boldsymbol{G}\boldsymbol{\kappa}_x + \boldsymbol{U}\boldsymbol{\psi}_x + \boldsymbol{\varepsilon}_{\boldsymbol{X}_1} \\
\boldsymbol{X}_2 &= \boldsymbol{G}\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}_{\boldsymbol{X}_2} \\
\boldsymbol{Y} &= \beta_1\boldsymbol{X}_1 + \beta_2 X_2 + \boldsymbol{G}\boldsymbol{\kappa}_y + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{U}\boldsymbol{\psi}_y + \boldsymbol{\varepsilon}_{\boldsymbol{Y}},
\end{aligned} \tag{7}$$

where, $\boldsymbol{U}$ denotes the confounder matrix for q confounders, and each column is sampled from a standard normal distribution; $\boldsymbol{\psi}_x \in R^{q\times 1}$ and $\boldsymbol{\psi}_y \in R^{q\times 1}$ are the corresponding vector of coefficients, and are sampled from a bivariate normal distribution $N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\psi\right)$, where $\boldsymbol{\Sigma}_\psi$ is a two-by-two diagonal matrix with elements 1. For IIE and CHP, we similarly assume $\boldsymbol{\kappa}_x$ and $\boldsymbol{\kappa}_y$ follow a bivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_\kappa)$, where $\boldsymbol{\Sigma}_\kappa$ is a two-by-two matrix with diagonal elements of 1 and off-diagonal elements of ρ_κ denoting the correlation between them. Additionally, we generate $\boldsymbol{\delta}$, $\boldsymbol{\gamma}_2$ and $\boldsymbol{\theta}$ from $N\left(0, \sigma^2_\delta\right)$, $N\left(0, \sigma^2_{\gamma_2}\right)$, and $N\left(0, \sigma^2_\theta\right)$, respectively. $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{Y}$ are the primary exposure, the secondary exposure and the outcome, respectively; $\boldsymbol{\varepsilon}_{\boldsymbol{X}_1}$, $\boldsymbol{\varepsilon}_{\boldsymbol{X}_2}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{Y}}$ are the random errors; β_1 and β_2 are the causal effects of the primary exposure and the secondary exposure on the outcome, respectively.

We first let $\boldsymbol{G} = \boldsymbol{G}_s$, and perform the single variant analysis on $\boldsymbol{G}_s$ and $\boldsymbol{X}_1$ to obtain the reference summary statistics between the rare variants and $\boldsymbol{X}_1$; then we calculate **PRS** based on $\boldsymbol{G}_i$. To illustrate the effect of the rare SNPs on the outcome, we modified the expression of $\boldsymbol{Y}$

$$\boldsymbol{Y} = \beta_1\boldsymbol{X}_1 + \beta_2\boldsymbol{X}_2 + \beta_P\boldsymbol{PRS} + \boldsymbol{G}\boldsymbol{\kappa}_y + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{U}\boldsymbol{\psi}_y + \boldsymbol{\varepsilon}_{\boldsymbol{Y}}. \tag{8}$$

We then let $\boldsymbol{G} = \boldsymbol{G}_i$ and conduct the single variant analysis on $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, **PRS** and $\boldsymbol{Y}$ with respect to $\boldsymbol{G}_i$ to calculate the summary statistics $\left(\hat{\gamma}_{1j}, \hat{S}_{\gamma_{1j}}\right)$, $\left(\hat{\gamma}_{2j}, \hat{S}_{\gamma_{2j}}\right)$, $\left(\hat{\gamma}_{Pj}, \hat{S}_{\gamma_{Pj}}\right)$ and $\left(\hat{\Gamma}_j, \hat{S}_{\Gamma_j}\right)$, $\forall j = 1, 2, \ldots, p$. In the simulation studies, we control the magnitudes for $\boldsymbol{\delta}$, $\boldsymbol{\kappa}_x$, $\boldsymbol{\gamma}_2$, $\boldsymbol{\kappa}_y$ and $\boldsymbol{\theta}$ using $h^2_{\boldsymbol{\delta}} = \frac{\text{var}(\beta_1 \boldsymbol{G\delta})}{\text{var}(\boldsymbol{Y})}$, $h^2_{\boldsymbol{\kappa}_x} = \frac{\text{var}(\beta_1 \boldsymbol{G\kappa}_x)}{\text{var}(\boldsymbol{Y})}$, $h^2_{\boldsymbol{\kappa}_y} = \frac{\text{var}(\boldsymbol{G\kappa}_y)}{\text{var}(\boldsymbol{Y})}$, $h^2_{\boldsymbol{\gamma}_2} = \frac{\text{var}(\beta_2 \boldsymbol{G\gamma}_2)}{\text{var}(\boldsymbol{Y})}$, and $h^2_{\boldsymbol{\theta}} = \frac{\text{var}(\boldsymbol{G\theta})}{\text{var}(\boldsymbol{Y})}$, repectively. We define $h^2_{\boldsymbol{\gamma}_1} = h^2_{\boldsymbol{\delta}} + h^2_{\boldsymbol{\kappa}_x}$, and let $h^2_{\boldsymbol{\delta}} = 4h^2_{\boldsymbol{\kappa}_x}$. We vary $h^2_{\boldsymbol{\gamma}_1}$ from 0 to 0.1 to evaluate the control of type I error rate ($h^2_{\boldsymbol{\gamma}_1} = 0$) and to examine the power of RARE ($h^2_{\boldsymbol{\gamma}_1} \neq 0$).

Key Points

- This paper proposes a novel multivariable Mendelian randomization method named **RARE**;
- **RARE** models the impact of rare variants on inferring causal effect of primary exposure on outcome;
- **RARE** accounts for both uncorrelated horizontal pleiotropy (UHP) and correlated horizontal pleiotropy (CHP);
- **RARE** decomposes UHP into three parts;
- This paper demonstrates superiority of **RARE** when combing rare variants into polygenic risk score.

Acknowledgements

We express our gratitude to all the researchers who have made their GWAS data publicly available, and we also extend our gratitude to the entire staff of the UK Biobank for their significant contributions to epidemiology.

Author contributions

Fangrong Yan, Tiantian Liu and Yu Cheng conceived the design of the study. Fangrong Yan provided funding support. Fangrong Yan, Tiantian Liu and Liyun Jiang supervised all the proposals and derivations of the statistical methods. Yu Cheng, Xinjia Ruan, Liyun Jiang, Xiaofan Lu, Yuqing Yang, Yuhang Wang, Shangjin Yan and Yuzhe Sun performed all the statistical and computational analyses. Yu Cheng and Xinjia Ruan developed the R package. Yucheng and Xinjia Ruan wrote the manuscript. Fangrong Yan and Tiantian Liu provided comments to refine the manuscript and approved the final manuscript.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declared no competing interests.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities of China Pharmaceutical University [2632023FY04, 2632023FY05], the National Natural Science Foundation of China [No. 82273735, No. 82304252, No. 82473737], the National Science Foundation for Young Scientists of China [No. 82404380], the 2024 National Key Research on Multi-Target Natural Drugs [SKLNMZZ2024JS32], and Key Laboratory of Scientific and Engineering Computing (Ministry of Education) and the Shanghai Frontiers Science Center of Modern Analysis.

Data availability

All GWAS summary statistics used in this study are publicly available in the GWAS Catalog (https://www.ebi.ac.uk/gwas/) and fastGWA (https://yanglab.westlake.edu.cn/data/ukb_fastgwa/imp_binary/), and all these data can be obtained by sending an email to the first author or the corresponding author. The UK Biobank is a paid database, and the individual-level genotype data can be obtained after submitting an application to the UK Biobank. All the analysis code are publicly available, and RARE is implemented in an open-source R package available through Github at https://github.com/Hide-in-lab/RARE/tree/main. The code used to reproduce the analysis can be found at https://github.com/Hide-in-lab/RARE/tree/main/simulation.

References

1. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22. https://doi.org/10.1093/ije/dyg070
2. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004;**33**:30–42. https://doi.org/10.1093/ije/dyh132
3. de Leeuw C, Savage J, Bucur IG. *et al.* Understanding the assumptions underlying Mendelian randomization. *Eur J Hum Genet* 2022;**30**:653–60. https://doi.org/10.1038/s41431-022-01038-5
4. Murphy N, Song M, Papadimitriou N. *et al.* Associations between glycemic traits and colorectal cancer: a Mendelian randomization analysis. *J Natl Cancer Inst* 2022;**114**:740–52. https://doi.org/10.1093/jnci/djac011
5. Bull CJ, Bell JA, Murphy N. *et al.* Adiposity, metabolites, and colorectal cancer risk: Mendelian randomization study. *BMC Med* 2020;**18**:396. https://doi.org/10.1186/s12916-020-01855-9
6. Wang X, Dai JY, Albanes D. *et al.* Mendelian randomization analysis of C-reactive protein on colorectal cancer risk. *Int J Epidemiol* 2019;**48**:767–80. https://doi.org/10.1093/ije/dyy244
7. Burgess S, Davey Smith G, Davies NM. *et al.* Guidelines for performing Mendelian randomization investigations: Update for summer 2023. *Wellcome Open Res* 2019;**4**:186. https://doi.org/10.12688/wellcomeopenres.15555.3
8. Do R, Stitziel NO, Won H-H. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2015;**518**:102–6. https://doi.org/10.1038/nature13917
9. Helgadottir A, Gretarsdottir S, Thorleifsson G. *et al.* Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat Genet* 2016;**48**:634–9. https://doi.org/10.1038/ng.3561
10. Surakka I, Horikoshi M, Mägi R. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* 2015;**47**:589–97. https://doi.org/10.1038/ng.3300
11. Hou K, Ding Y, Xu Z. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat Genet* 2023;**55**:549–58. https://doi.org/10.1038/s41588-023-01338-6
12. Sullivan PF, Yao S, Hjerling-Leffler J. Schizophrenia genomics: genetic complexity and functional insights. *Nat Rev Neurosci* 2024;**25**:611–24. https://doi.org/10.1038/s41583-024-00837-7
13. Lambert JC, Ramirez A, Grenier-Boley B. *et al.* Step by step: towards a better understanding of the genetic architecture of

Alzheimer's disease. *Mol Psychiatry* 2023;**28**:2716–27. https://doi.org/10.1038/s41380-023-02076-1
14. Morrison J, Knoblauch N, Marcus JH. *et al.* Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat Genet* 2020;**52**:740–7. https://doi.org/10.1038/s41588-020-0631-4
15. Cheng Q, Zhang X, Chen LS. *et al.* Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nat Commun* 2022;**13**:6490. https://doi.org/10.1038/s41467-022-34164-1
16. Sun BB, Maranville JC, Peters JE. *et al.* Genomic atlas of the human plasma proteome. *Nature* 2018;**558**:73–9. https://doi.org/10.1038/s41586-018-0175-2
17. Astle WJ, Elding H, Jiang T. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 2016;**167**:1415–1429 e1419. https://doi.org/10.1016/j.cell.2016.10.042
18. Shin SY, Fauman EB, Petersen A-K. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014;**46**:543–50. https://doi.org/10.1038/ng.2982
19. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* 2015;**181**:251–60. https://doi.org/10.1093/aje/kwu283
20. Waterworth DM, Ricketts SL, Song K. *et al.* Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 2010;**30**:2264–76. https://doi.org/10.1161/ATVBAHA.109.201020
21. Lin Z, Xue H, Pan W. Robust multivariable Mendelian randomization based on constrained maximum likelihood. *Am J Hum Genet* 2023;**110**:592–605. https://doi.org/10.1016/j.ajhg.2023.02.014
22. Rees JMB, Wood AM, Burgess S. Extending the MR-egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med* 2017;**36**:4705–18. https://doi.org/10.1002/sim.7492
23. Grant AJ, Burgess S. Pleiotropy robust methods for multivariable Mendelian randomization. *Stat Med* 2021;**40**:5813–30. https://doi.org/10.1002/sim.9156
24. Wang J, Zhao Q, Bowden J. *et al.* Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS Genet* 2021;**17**:e1009575. https://doi.org/10.1371/journal.pgen.1009575
25. Bond TA, Richmond RC, Karhunen V. *et al.* Exploring the causal effect of maternal pregnancy adiposity on offspring adiposity: Mendelian randomisation using polygenic risk scores. *BMC Med* 2022;**20**:34. https://doi.org/10.1186/s12916-021-02216-w
26. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* 2020;**12**:44. https://doi.org/10.1186/s13073-020-00742-5
27. Dudbridge F. Polygenic epidemiology. *Genet Epidemiol* 2016;**40**:268–72. https://doi.org/10.1002/gepi.21966
28. Shapland CY, Zhao Q, Bowden J. Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy. *Stat Med* 2022;**41**:1100–19. https://doi.org/10.1002/sim.9320
29. Lee S, Abecasis GR, Boehnke M. *et al.* Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014;**95**:5–23. https://doi.org/10.1016/j.ajhg.2014.06.009
30. Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. *Ann Transl Med* 2021;**9**:961. https://doi.org/10.21037/atm-21-1635
31. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: an R package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front Genet* 2020;**11**:157. https://doi.org/10.3389/fgene.2020.00157
32. Cao X, Tang Z, Zhang J. *et al.* Association between high-density lipoprotein cholesterol and type 2 diabetes mellitus among Chinese: the Beijing longitudinal study of aging. *Lipids Health Dis* 2021;**20**:71. https://doi.org/10.1186/s12944-021-01499-5
33. Haase CL, Tybjaerg-Hansen A, Nordestgaard BG. *et al.* HDL cholesterol and risk of type 2 diabetes: a Mendelian randomization study. *Diabetes* 2015;**64**:3328–33. https://doi.org/10.2337/db14-1603
34. Sunil B, Ashraf AP. Dyslipidemia in Pediatric type 2 diabetes mellitus. *Curr Diab Rep* 2020;**20**:53. https://doi.org/10.1007/s11892-020-01336-6
35. Wan KS, Hairi NN, Mustapha F. *et al.* Five-year LDL-cholesterol trend and its predictors among type 2 diabetes patients in an upper-middle-income country: a retrospective open cohort study. *Peer J* 2022;**10**:e13816. https://doi.org/10.7717/peerj.13816
36. Casula M, Colpani O, Xie S. *et al.* HDL in atherosclerotic cardiovascular disease: In search of a role. *Cells* 2021;**10**. https://doi.org/10.3390/cells10081869
37. Slomski A. High HDL cholesterol linked with death in coronary artery disease. *JAMA* 2022;**328**:10. https://doi.org/10.1001/jama.2022.10273
38. Mortensen MB, Dzaye O, Bøtker HE. *et al.* Low-density lipoprotein cholesterol is predominantly associated with atherosclerotic cardiovascular disease events in patients with evidence of coronary atherosclerosis: the western Denmark heart registry. *Circulation* 2023;**147**:1053–63. https://doi.org/10.1161/CIRCULATIONAHA.122.061010
39. Quan W, Han H, Liu L. *et al.* Influence of LDL-cholesterol lowering on coronary plaque progression of non-target lesions in patients undergoing percutaneous coronary intervention: findings from a retrospective study. *J Clin Med* 2023;**12**. https://doi.org/10.3390/jcm12030785
40. Carter AR, Sanderson E, Hammerton G. *et al.* Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur J Epidemiol* 2021;**36**:465–78. https://doi.org/10.1007/s10654-021-00757-1
41. Beddhu S, Chertow GM, Cheung AK. *et al.* Influence of baseline diastolic blood pressure on effects of intensive compared with standard blood pressure control. *Circulation* 2018;**137**:134–43. https://doi.org/10.1161/CIRCULATIONAHA.117.030848
42. Liang F, Wang Y. Coronary heart disease and atrial fibrillation: a vicious cycle. *Am J Physiol Heart Circ Physiol* 2021;**320**:H1–12. https://doi.org/10.1152/ajpheart.00702.2020
43. Graham SE, Clarke SL, Wu KHH. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* 2021;**600**:675–9. https://doi.org/10.1038/s41586-021-04064-3

Briefings in Bioinformatics, 2025, **26(3)**, bbaf214
https://doi.org/10.1093/bib/bbaf214
Problem Solving Protocol