

# Benchmarking Mendelian Randomization methods for causal inference using genome-wide association study summary statistics

Xianghong Hu<sup>1,2</sup>, Mingxuan Cai<sup>4</sup>, Jiashun Xiao<sup>5</sup>, Xiaomeng Wan<sup>1,2</sup>, Zhiwei Wang<sup>1,2</sup>, Hongyu Zhao<sup>\*6</sup>, and Can Yang<sup>\*1,2,3</sup>

<sup>1</sup>Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China.

<sup>2</sup>Guangzhou HKUST Fok Ying Tung Research Institute, Guangzhou 511458, China.

<sup>3</sup>Big Data Bio-Intelligence Lab, The Hong Kong University of Science and Technology, Hong Kong SAR, China

<sup>4</sup>Department of Biostatistics, City University of Hong Kong, Hong Kong, China.

<sup>5</sup>Shenzhen Research Institute of Big Data, Shenzhen 518172, China.

<sup>6</sup>Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA.

## Abstract

1 Mendelian Randomization (MR), which utilizes genetic variants as instrumental variables (IVs),  
2 has gained popularity as a method for causal inference between phenotypes using genetic data.  
3 While efforts have been made to relax IV assumptions and develop new methods for causal  
4 inference in the presence of invalid IVs due to confounding, the reliability of MR methods  
5 in real-world applications remains uncertain. To bridge this gap, we conducted a benchmark  
6 study evaluating 15 MR methods using real-world genetic datasets. Our study focused on  
7 three crucial aspects: type I error control in the presence of various confounding scenarios  
8 (e.g., population stratification, pleiotropy, and assortative mating), the accuracy of causal  
9 effect estimates, replicability and power. By comprehensively evaluating the performance of  
10 compared methods over one thousand pairs of exposure-outcome traits, our study not only

---

\*To whom correspondence may be addressed: Hongyu Zhao ([hongyu.zhao@yale.edu](mailto:hongyu.zhao@yale.edu)) and Can Yang ([macyang@ust.hk](mailto:macyang@ust.hk)).

11 provides valuable insights into the performance and limitations of the compared methods but  
12 also offers practical guidance for researchers to choose appropriate MR methods for causal  
13 inference.

## 14 Introduction

15 Understanding the causal relationships between exposures and outcomes is crucial in biomedical  
16 and social science research, as it enables discoveries in etiology, aids in drug development, and  
17 informs policy-making. While randomized controlled trials (RCTs) are considered the gold  
18 standard for assessing causality, they can be time-consuming, costly, and sometimes ethically  
19 challenging [1]. Causal inference based on observational data presents its own challenges,  
20 such as unmeasured confounding or reverse causality. Mendelian randomization (MR) offers a  
21 promising approach to performing causal inference using observed genetic data [2, 3]. According  
22 to Mendel’s law of inheritance, genotypes are randomly inherited from parents to offspring,  
23 thereby ideally being independent of environmental confounding factors. This characteristic  
24 motivates researchers to explore genetic data in order to study the causal effects of one phenotype  
25 (exposure) on another phenotype (outcome). In recent years, MR has gained popularity due  
26 to the availability of summary statistics from thousands of genome-wide association studies  
27 (GWAS) covering a wide range of phenotypes. Leveraging the rich genetic data resources  
28 available, researchers worldwide can investigate the potential causal relationships between  
29 exposures and outcomes of interest, encompassing diverse applications such as identifying disease  
30 risk causation [4], providing evidence for epidemiological associations [5], and prioritizing targets  
31 in drug development [6, 7].

32 To perform causal inference using MR approaches, genetic variants (typically Single  
33 Nucleotide Polymorphisms, i.e., SNPs) serve as instrument variables (IVs). A valid IV should  
34 satisfy the following three IV assumptions [8, 9]: (1) it is associated with the exposure of  
35 interest; (2) it is not associated with the confounders of the exposure and outcome traits; and  
36 (3) it affects the outcome only through the exposure of interest. However, these assumptions  
37 underlying MR are often too strong to be satisfied in real applications. In recent years, much  
38 effort has been devoted to relaxing these assumptions and new MR methods have been designed  
39 to enable causal inference in the presence of invalid IVs. To name a few, MR-PRESSO [10],  
40 cML-MA [11], and MR-Lasso [12] use outlier detection to identify invalid IVs and remove them  
41 from the MR analysis. MR-Robust[12], weighted-median [13], and weighted-mode [14] use  
42 outlier-robust techniques to mitigate the effects of invalid IVs. Additionally, methods like Egger  
43 [15], RAPS [16], and BWMR [17] employ probabilistic models to correct for different types of  
44 pleiotropy, while CAUSE [18], MRAPSS [19], MRMix [20], MR-ConMix [21], and MR-CUE  
45 [22] employ mixture component models to characterize valid and invalid signals, enabling causal  
46 inference based on the component of valid signals.

47 Although considerable progress has been made in the development of MR methods, their  
48 robustness to the violation of underlying assumptions in real-world applications remains  
49 largely unclear. Due to the complexity of human genetics, several factors can significantly  
50 impact the performance of existing MR methods. Firstly, complex traits often exhibit high  
51 polygenicity, meaning that individual SNPs have small effect sizes. To satisfy the IV assumption  
52 (1), researchers select SNPs as IVs from the exposure GWAS using a  $p$ -value threshold (IV

53 threshold). However, this selection process may inadvertently include weakly associated SNPs,  
54 which can introduce bias into MR estimates. Moreover, using the same exposure dataset for  
55 IV selection and MR estimation in two-sample MR settings can induce non-ignorable bias,  
56 known as selection bias [16]. Second, population stratification and family-level confounders  
57 (e.g., assortative mating and dynastic effects) are well-known issues in population-based GWAS,  
58 which can introduce associations between genetic instruments and unobserved confounders  
59 [23, 24, 25], leading to the violation of IV assumption (2). Despite the significance of population  
60 stratification and family-level confounders, many existing MR methods have not explicitly  
61 accounted for these. Third, pleiotropy is a ubiquitous phenomenon in human genetics, referring  
62 to a single genetic variant influencing multiple traits, thereby violating IV assumption (3) [26].  
63 Carefully accounting for pleiotropy is crucial for reliable causal inference using MR approaches.  
64 Given these complexities, it is crucial to conduct benchmarking studies to assess the reliability  
65 of existing MR methods when their model assumptions may be violated. Such studies would  
66 provide valuable insights into the performance and limitations of these methods in real-world  
67 scenarios.

68 In this study, we present a benchmarking analysis of MR methods for causal inference  
69 with real-world genetic datasets. Our focus is on MR methods that utilize GWAS summary  
70 statistics as input, as they do not require access to individual-level GWAS data and are widely  
71 applicable. Specifically, we consider 15 MR methods, including the standard IVW (fixed) [27]  
72 and IVW (random) [28] and 13 other advanced MR methods: Egger, RAPS, Weighted-median,  
73 Weighted-mode, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, CAUSE,  
74 MRAPSS and MR-ConMix. To assess the performance of these MR methods, we utilized  
75 real-world datasets and focused on three key aspects: type I error control, the accuracy of  
76 causal effect estimates, and replicability. Particularly, in evaluating type I error control, we  
77 used GWAS summary-level datasets for over one thousand exposure-outcome trait pairs of no  
78 causal effect, serving as negative controls. These trait pairs were carefully selected to represent  
79 scenarios involving confounding factors, such as population stratification and pleiotropy. We  
80 conducted a comparison between population-based MR and family-based MR to evaluate the  
81 influence of family-level confounders. Through our comprehensive experiments using real-world  
82 datasets, we found that the performance of MR methods is heavily influenced by confounding  
83 factors that arise from various sources in practical scenarios. We also investigated the influence  
84 of summary-level data pre-processing steps, such as the inclusion of SNPs with different minor  
85 allele frequencies and the choice of reference genome panels. Our study offers practical guidelines  
86 for researchers in choosing appropriate MR methods and improving the reliability of causal  
87 inference in MR analyses.

## 88 Results

### 89 The experimental design for benchmarking MR methods

90 We conducted a benchmarking of 15 summary-level data-based MR methods, which were  
91 categorized into four groups: IVW-class, outlier detection and removal methods, model-based  
92 methods, and outlier robust methods (Fig. 1-A, Table S1, section 1 of the supplementary note).  
93 The procedure for running the MR methods is outlined in Fig. 1-B and described in detail in

94 the Method section. To ensure a comprehensive evaluation, we utilized real-world datasets and  
95 focused on three crucial aspects: type I error control, the accuracy of causal effect estimates,  
96 replicability and power (see Fig. 1-C).

97 To assess type I error control, we applied the MR methods to GWAS summary statistics  
98 for three sets of exposure-outcome trait pairs with no causal effect. The three sets of trait  
99 pairs represented three different confounding scenarios, including (a) population stratification,  
100 (b) pleiotropy, and (c) family-level confounders. Specifically, in scenario (a), we used 1,130  
101 trait pairs between 226 exposures from UK Biobank and five negative control outcomes to  
102 investigate the influence of population stratification on MR methods (Supplementary data 1).  
103 The negative control study was designed carefully based on two criteria: First, the outcomes  
104 should not be causally affected by the exposures. Second, both the outcomes and exposures  
105 should be affected by population stratification. In this scenario, we chose four hair color-related  
106 traits and tanning ability as negative control outcomes. These traits are mainly determined at  
107 birth and are likely influenced by population stratification [24]. In scenario (b), we analyzed  
108 trait pairs between 11 exposures and seven negative control outcomes. The selected exposures  
109 included five adult behavior-related traits and six aging-related traits, while the negative  
110 control outcomes were seven childhood-related traits (Supplementary data 2). This choice  
111 was based on the convention that traits developed after adulthood are unlikely to affect traits  
112 developed before adulthood causally. These negative control outcomes exhibited non-zero  
113 genetic correlations with most of the exposures (SFig S4), indicating that pleiotropy is a major  
114 confounder here. In scenario (c), we analyzed 82 trait pairs using both population-based GWASs  
115 and family-based GWASs to examine the influence of family-level confounders (Supplementary  
116 data 3). Population-based GWAS estimates, which are derived from unrelated individuals, are  
117 known to be susceptible to bias due to the influence of family-level confounders. Conversely,  
118 family-based GWAS designs offer the advantage of accounting for the effects of family-level  
119 confounders when estimating GWAS effects [29, 30]. By comparing the results of MR analyses  
120 obtained from the population-based GWAS and family-based GWAS designs, we can assess  
121 the effectiveness of MR methods in controlling for type I errors in the presence of family-level  
122 confounding, such as assortative mating and dynastic effects. For this scenario, we required  
123 the trait pairs to be genetically uncorrelated. Based on the principle “no correlation implies  
124 no causal relationship”, we treated these trait pairs as negative controls. By applying MR  
125 methods to the three datasets representing different confounding scenarios, we investigated  
126 their ability to control type I errors in the presence of different confounding factors.

127 In evaluating the accuracy of causal effect estimates, we examined six pairs of traits where  
128 each pair comprised the same trait as both the “exposure” and the “outcome” (Supplementary  
129 data 4). The UK Biobank dataset was divided equally to obtain exposure and outcome GWAS  
130 data. Importantly, the true causal effects in this analysis were known to be exactly one  
131 [16, 31]. This design allowed us to assess the accuracy of MR methods in estimating causal  
132 effects. To evaluate replicability and power, we focused on a positive control example involving  
133 low-density lipoprotein cholesterol (LDL-C) and coronary artery disease (CAD). We applied  
134 all the MR methods to six GWAS datasets for LDL-C obtained from five distinct studies  
135 (Supplementary data 5). By analyzing multiple GWAS datasets for the same trait, we assessed  
136 the replicability of the causal effect estimates across different study designs and sample sizes.  
137 Detailed information about the datasets used in this study can be found in the Method section,

138 and specific details regarding the sources of GWAS data are summarized in Stable 1-5.

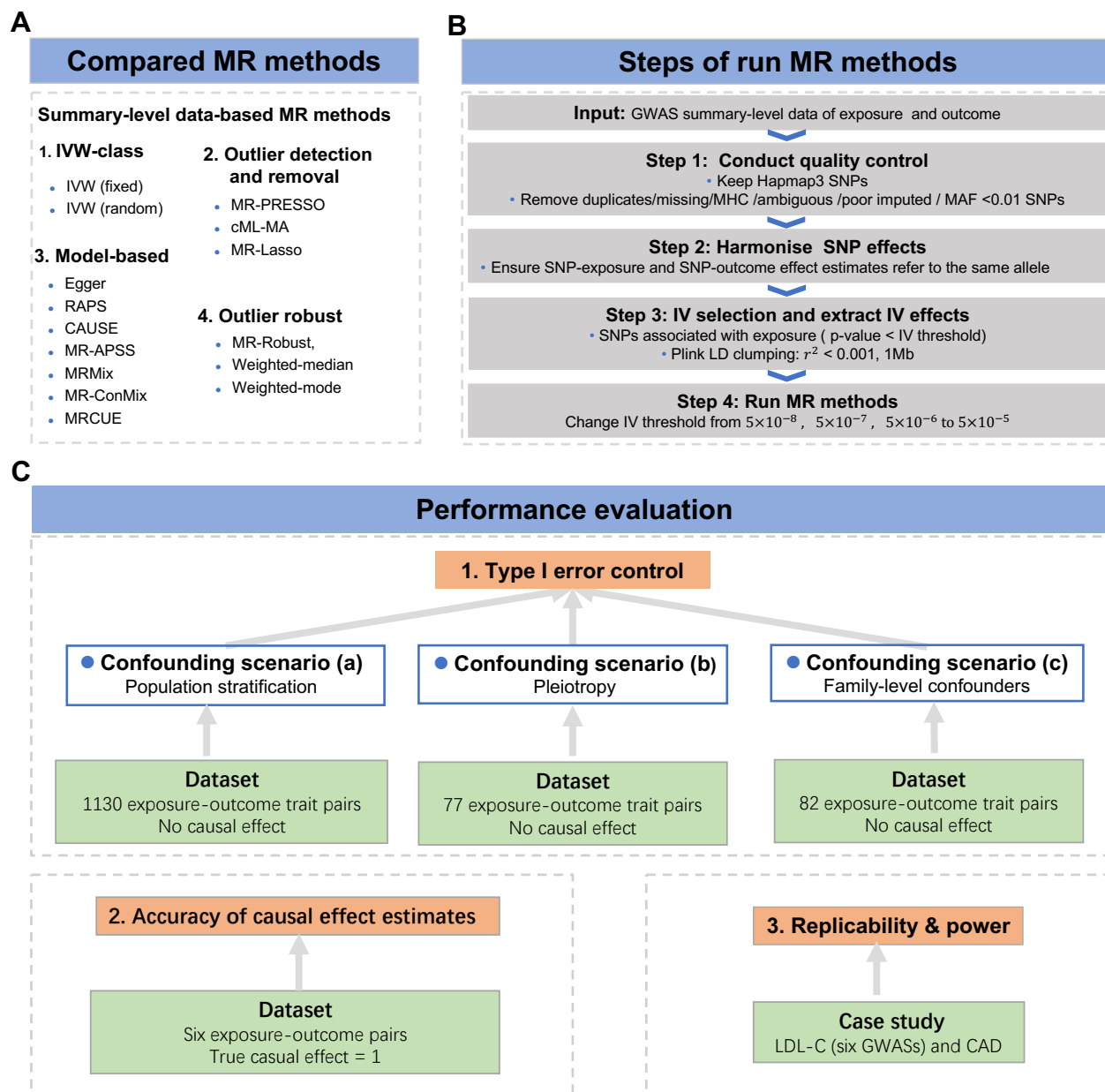


Figure 1: **Experimental design for benchmarking MR methods.** **A** We compared the performance of 15 GWAS summary-level data-based MR Methods. **B** We designed a four-step procedure for running MR methods. **C** We used real-world datasets to evaluate the performance of MR methods on three aspects: Type I error control in three confounding scenarios, including (a) population stratification, (b) pleiotropy, and (c) family-level confounders, the accuracy of causal effect estimates, replicability and power.

139 Throughout the evaluation process, our first step was to assess the performance of MR  
 140 methods when IVs were selected based on the default  $p$ -value thresholds in the exposure GWAS.  
 141 Specifically, among the compared methods, MR-APSS and MR-CUE utilized a default IV  
 142 threshold of  $5 \times 10^{-5}$ , CAUSE employed a default IV threshold of  $1 \times 10^{-3}$ , and the remaining  
 143 methods required strong IVs with a default IV threshold of  $5 \times 10^{-8}$ . All of the compared MR

144 methods, except for MR-CUE, require Plink LD clump ( $r^2 = 0.001, 1\text{Mb}$ ) to obtain independent  
145 SNPs as IVs. Furthermore, we introduced variations in the  $p$ -value thresholds used for IV  
146 selection to examine the methods' performance across a range of IV thresholds, including a  
147 stringent threshold of  $5 \times 10^{-8}$ , as well as more relaxed thresholds of  $5 \times 10^{-7}$ ,  $5 \times 10^{-6}$ , and  
148  $5 \times 10^{-5}$ . As the IV thresholds become looser, the number of IVs, including both valid IVs and  
149 invalid IVs, may increase. Moreover, the number of IVs with weaker effects may also increase.  
150 This analysis allowed us to assess the robustness of these methods in handling invalid IVs due  
151 to confounding factors and determine whether they are sensitive to the choice of IVs used in  
152 MR analysis.

## 153 **MR-APSS, Egger, Weighted-mode, and CAUSE achieve better performance** 154 **in type I error control**

155 We conducted a comprehensive evaluation to assess the effectiveness of 15 MR methods in  
156 controlling type I errors across various confounding scenarios. The evaluation utilized three  
157 real-world datasets and focused on three specific scenarios: population stratification, pleiotropy,  
158 and family-level confounders. To evaluate the performance of these methods, we generated QQ  
159 plots to visualize the  $p$ -values produced by each method for the three datasets, as shown in Figs.  
160 2-4. These QQ plots provide a visual tool to identify deviations from the expected diagonal  
161 line, which helps determine if the methods are generating systematically inflated or deflated  
162  $p$ -values. In this analysis, we initially assessed the MR methods using their default setting for  
163 IV selection. Specifically, we first examined the performance of MR-APSS and MR-CUE at  
164 the IV threshold of  $5 \times 10^{-5}$ , the performance of CAUSE at the IV threshold of  $1 \times 10^{-3}$ , and  
165 the performance of other methods at the IV threshold of  $5 \times 10^{-8}$ .

166 In scenario (a), characterized by the presence of strong population stratification, MR-APSS  
167 and Weighted-mode consistently generated well-calibrated  $p$ -values, using their default IV  
168 thresholds. However, Egger's  $p$ -values were slightly inflated. The  $p$ -values of CAUSE initially  
169 showed deflation but later exhibited inflation. Further analysis of causal effect estimates  
170 reveals that CAUSE's confidence intervals are more reliable compared to its  $p$ -values. On  
171 the other hand, the remaining 11 methods, including IVW (fixed), IVW (random), RAPS,  
172 Weighted-median, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, and  
173 MR-ConMix, exhibited highly inflated  $p$ -values at the default IV threshold of  $5 \times 10^{-8}$ . Notably,  
174 IVW (fixed) demonstrated the most severe inflation, which is expected as it is a basic MR  
175 method that does not account for IV invalidity, leading to bias and inflation in the estimates.  
176 While other methods incorporated different assumptions to address invalid IVs, they still failed  
177 to effectively control type I error inflation. These findings highlight the limitations of existing  
178 methods in handling scenarios involving strong population stratification, where their model  
179 assumptions do not align well with real-world situations.

180 In scenario (b), where pleiotropy is present, several methods exhibited effective control  
181 of type I errors. Notably, CAUSE, Egger, MR-APSS, and Weighted-mode demonstrated the  
182 absence of inflated  $p$ -values, indicating their capability to address pleiotropy. However, it was  
183 observed that CAUSE's  $p$ -values were deflated. On the other hand, several other methods,  
184 including IVW (random), MR-Lasso, MR-PRESSO, RAPS, Weighted-median, IVW (fixed),  
185 cML-MA, MR-ConMix, and MR-Robust, exhibited inflated  $p$ -values at the default IV threshold.

186 Notably, IVW (fixed), cML-MA, and MR-ConMix showed more pronounced inflation compared  
187 to the other methods. Despite these methods' primary focus on addressing pleiotropy, their  
188 performance in controlling type I errors was not entirely satisfactory. This observation indicates  
189 the ongoing challenge in effectively handling pleiotropy in MR analysis and the need for further  
190 methodological advancements.

191 In scenario (c), we conducted a comparison between the results of MR methods using  
192 both population-based and family-based GWAS data for 82 negative control trait pairs. The  
193 objective was to assess the effectiveness of MR methods in controlling for type I errors in  
194 the presence of family-level confounders. The QQ plots for MR methods at their default IV  
195 thresholds, using both population-based GWAS and within-family-based GWAS summary-level  
196 data, are depicted in Fig. 4. When using population-based GWAS data, Egger, Weighted-mode,  
197 and MRMix did not yield inflated  $p$ -values. CAUSE produced deflated  $p$ -values, and MR-APSS  
198 exhibited very slight inflation in the  $p$ -values. On the other hand, other methods such as IVW  
199 (fixed), MR-Lasso, cML-MA, and MR-CUE produced inflated  $p$ -values, indicating challenges in  
200 adequately addressing family-level confounding using these methods. However, when utilizing  
201 family-based GWAS data, all MR methods produced well-calibrated  $p$ -values, demonstrating  
202 effective control of type I error inflation. Our results provide further evidence for the usefulness  
203 of family-based MR in mitigating the influence of family-level confounders in MR analysis.

#### 204 **IV selection largely affects the performance of MR methods**

205 We conducted a comprehensive investigation into the performance of various MR methods by  
206 analyzing their behavior across a range of IV thresholds, including a stringent threshold of  
207  $5 \times 10^{-8}$ , as well as more relaxed thresholds of  $5 \times 10^{-7}$ ,  $5 \times 10^{-6}$ , and  $5 \times 10^{-5}$ . The QQ plots  
208 in Figs. 2 and 3 depict the results obtained in confounding scenarios (a) and (b), respectively.  
209 From these plots, we can observe that MR-APSS, Egger, and weighted-mode consistently  
210 generated well-calibrated  $p$ -values across varying IV thresholds. However, it is worth noting  
211 that the  $p$ -values obtained from CAUSE were consistently deflated. On the other hand, the  
212 remaining 11 methods, including IVW (fixed), IVW (random), RAPS, Weighted-median, MR-  
213 PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, and MR-ConMix, exhibited  
214 substantially inflated  $p$ -values. Furthermore, the degree of  $p$ -value inflation tended to increase  
215 as the IV threshold became looser. MRMix was an exception with slightly inflated  $p$ -values  
216 at a less stringent IV threshold of  $5 \times 10^{-5}$  but more inflated  $p$ -values at the IV threshold of  
217  $5 \times 10^{-8}$ , as observed in Fig. 3. This observation suggests that MRMix can be sensitive to the  
218 number of IVs used. It tends to produce more false positives when there is a limited number of  
219 IVs. Our results indicate that causal inference results obtained from most of the methods are  
220 sensitive to the IV threshold.

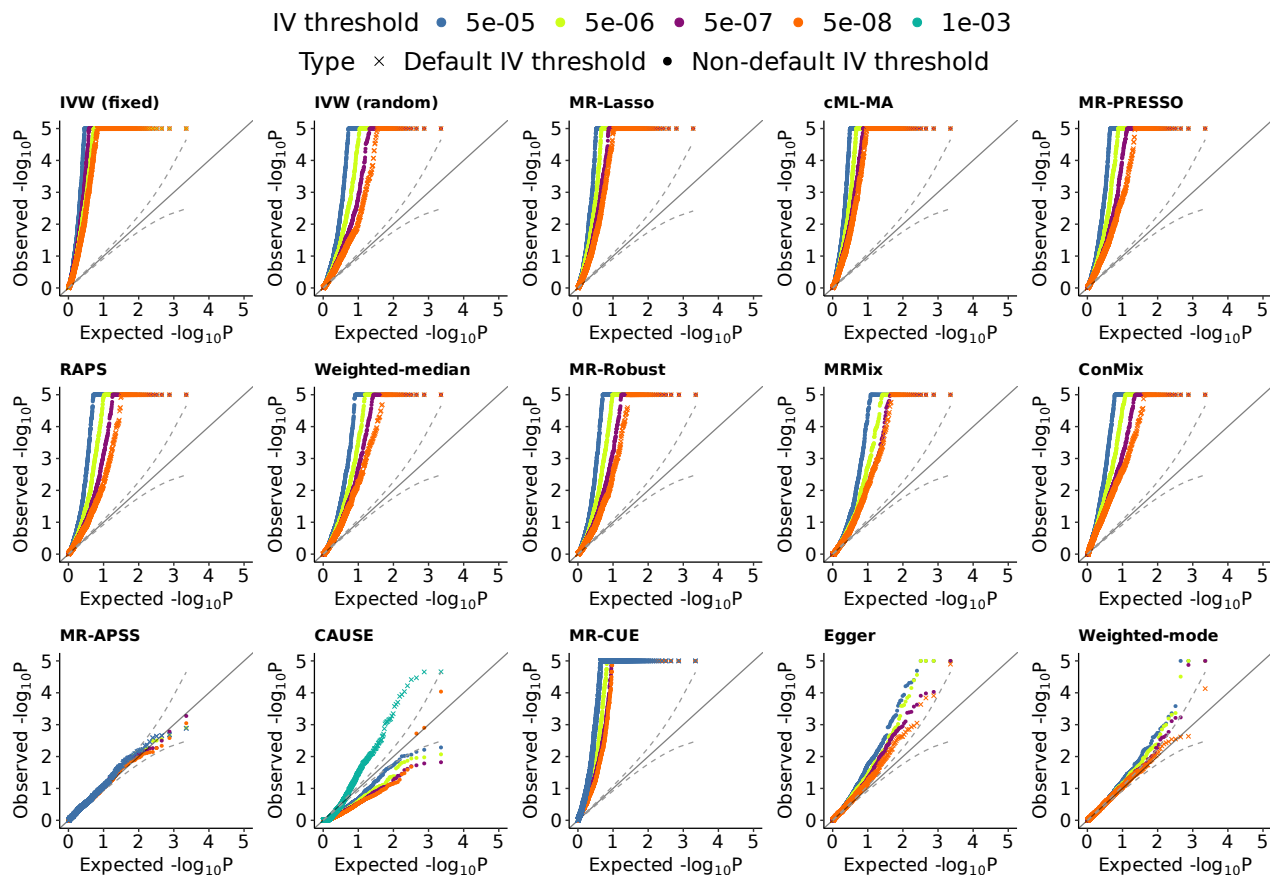


Figure 2: **Evaluation of type I error control in confounding scenario (a) of population stratification.** Type I error is evaluated by quantile-quantile plots of  $-\log_{10}(p)$  values from the 15 compared methods when testing the causal effect for 1130 negative control trait pairs at different IV thresholds. The 15 compared methods include IVW (fixed), IVW (random), Egger, RAPS, Weighted-median, Weighted-mode, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, CAUSE, MRAPSS and MR-ConMix. Each distinct color on the plot represents the results at a specific IV threshold and the results at the default IV thresholds of the compared MR methods are marked by a cross symbol. The default IV thresholds for MR-APSS and MR-CUE were set at  $5 \times 10^{-5}$ , while the default IV threshold for CAUSE was set at  $1 \times 10^{-3}$ . The remaining methods utilized a default IV threshold of  $5 \times 10^{-8}$ .



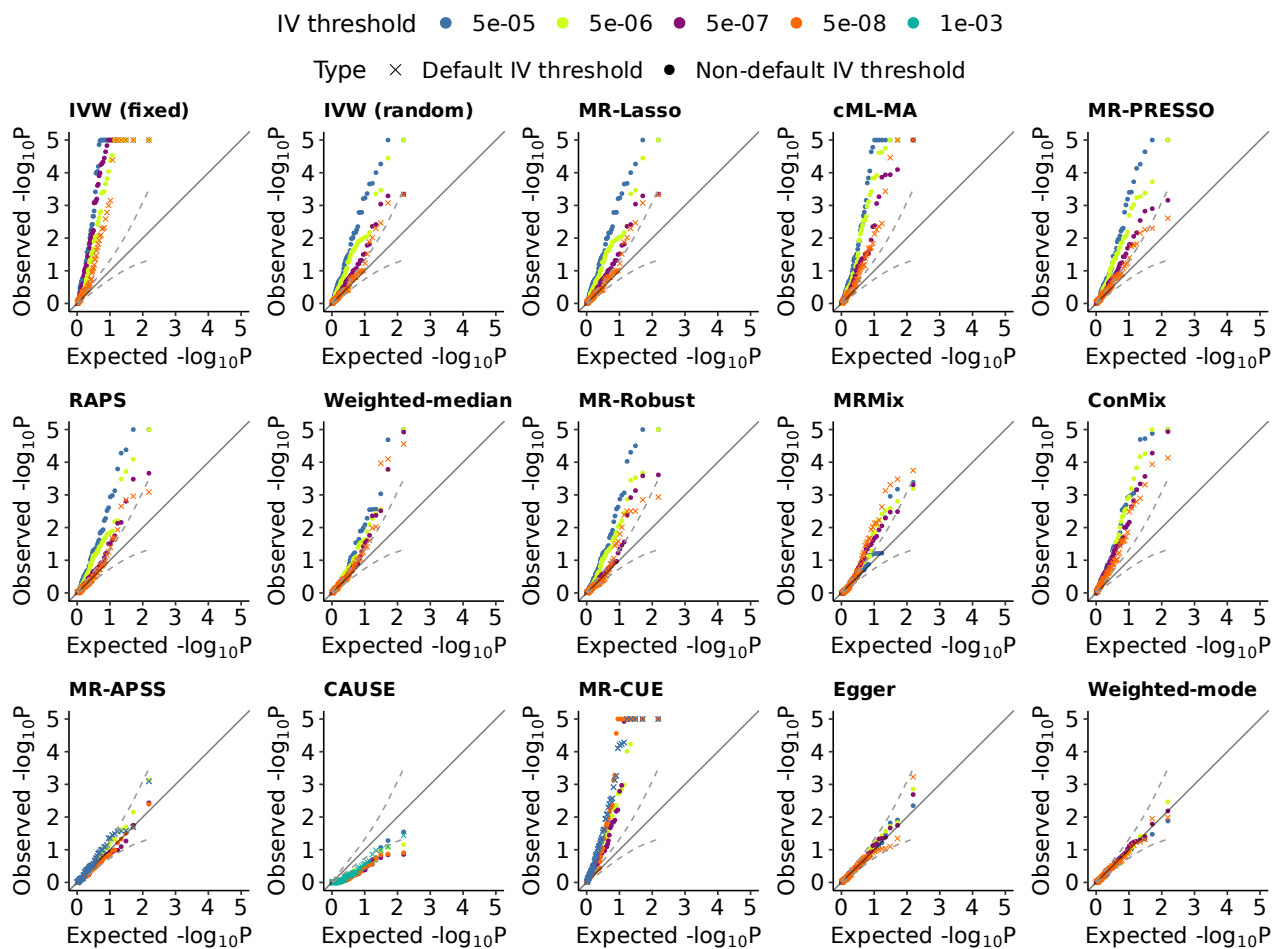


Figure 3: **Evaluation of type I error control in confounding scenario (b) of pleiotropy.** Type I error is evaluated by quantile-quantile plots of  $-\log_{10}(p)$  values from the 15 compared methods when testing the causal effect for 77 negative control trait pairs at different IV thresholds. The 15 compared methods include IVW (fixed), IVW (random), Egger, RAPS, Weighted-median, Weighted-mode, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, CAUSE, MRAPSS and MR-ConMix. Each distinct color on the plot represents the results at a specific IV threshold and the results at the default IV thresholds of the compared MR methods are marked by a cross symbol. The default IV thresholds for MR-APSS and MR-CUE were set at  $5 \times 10^{-5}$ , while the default IV threshold for CAUSE was set at  $1 \times 10^{-3}$ . The remaining methods utilized a default IV threshold of  $5 \times 10^{-8}$ .

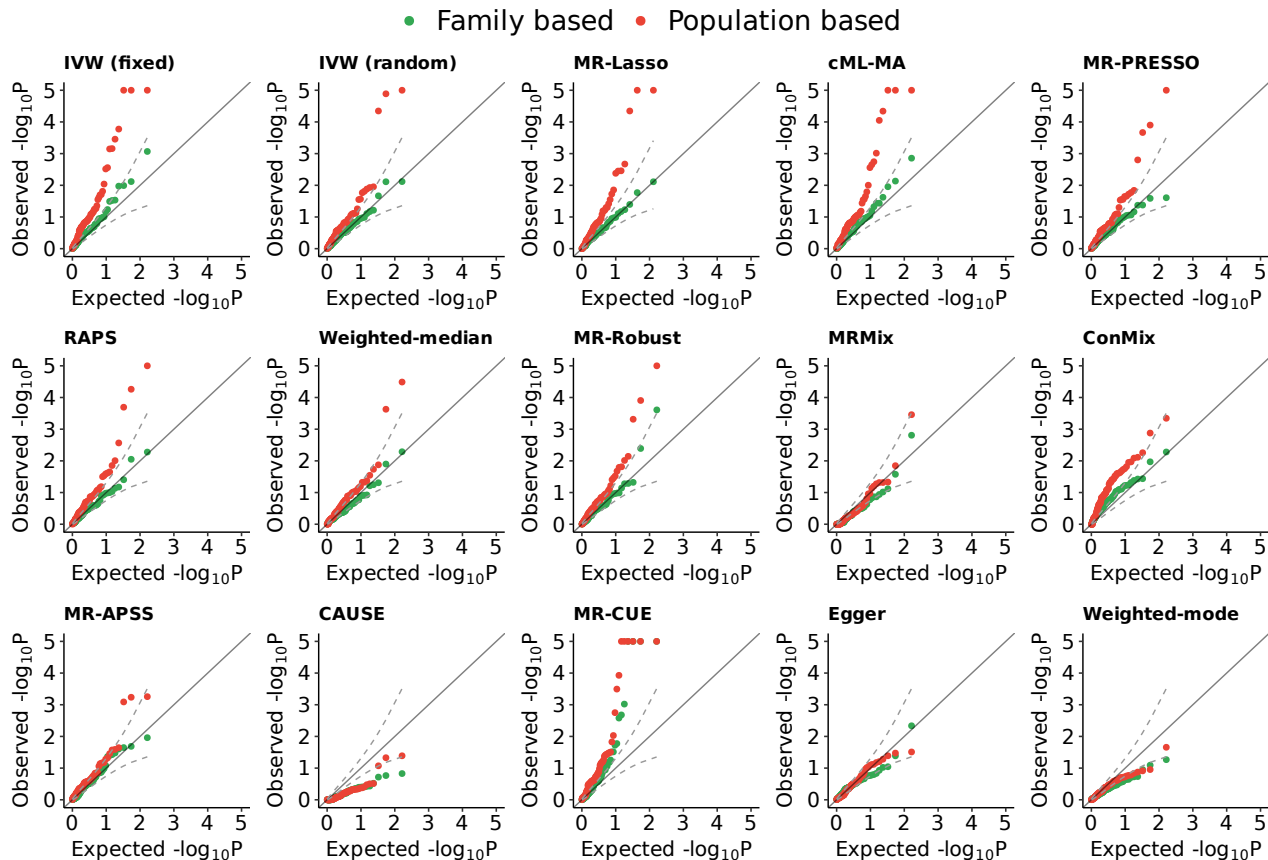


Figure 4: **Evaluation of Type I error control in the confounding scenario (c) of family-level confounders.** Quantile-quantile (Q-Q) plots illustrating the  $-\log_{10}(p)$  values for testing causal effects on 82 trait pairs using 15 different methods at their default IV thresholds. The comparison includes results from both population-based GWASs (depicted as red triangles) and sibling-based GWASs (depicted as green dots). The evaluated methods consist of IVW-fixed, IVW-random, Egger, RAPS, Weighted-median, Weighted-mode, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, CAUSE, MRAPSS, and MR-ConMix. MR-APSS and MR-CUE employ an IV threshold of  $5 \times 10^{-5}$ , CAUSE uses a threshold of  $1 \times 10^{-3}$ , while the remaining methods use a threshold of  $5 \times 10^{-8}$ .

## 221 Accuracy of causal effect estimates

222 To assess the accuracy of MR methods in estimating causal effects, we examined six pairs of  
 223 traits, where each pair involved the same trait being considered as both the “exposure” and  
 224 the “outcome” [16, 31]. In this specific scenario, the true causal effects for these trait pairs  
 225 were precisely known to be equal to one. This knowledge enabled us to compare the accuracy  
 226 of causal effect estimates given by different MR methods. We included three continuous traits:  
 227 Height, Waist Circumference (WC), and Educational Attainment (EA), as well as three binary  
 228 traits: Hypertension, High cholesterol, and Asthma. For each trait, we divided the UK Biobank  
 229 samples into two halves, representing the exposure GWAS and the outcome GWAS. We utilized  
 230 the Bolt-LMM software [32] to obtain GWAS summary statistics from these subsets. The  
 231 exposure GWAS summary statistics were used for both IV selection and causal effect estimation.

232 We varied the IV thresholds, starting with a stringent threshold of  $5 \times 10^{-8}$ , and progressively  
233 relaxed the thresholds to  $5 \times 10^{-7}$ ,  $5 \times 10^{-6}$ , and  $5 \times 10^{-5}$ . As such, we can investigate the  
234 robustness of MR methods to weak IV bias and selection bias. However, it is important to  
235 note that in this analysis, we cannot assess the robustness of MR methods to pleiotropy or  
236 other forms of confounding when testing the effect of the trait on itself using data from the  
237 same population. The causal effect estimates and their confidence intervals for 15 MR methods  
238 at different IV thresholds are presented in Fig. 5.

239 Our study found that MR-APSS outperformed other MR methods and produced more  
240 accurate causal effect estimates that were closer to the true value. Importantly, all of the  
241 confidence intervals produced by MR-APSS at different IV thresholds covered the true value.  
242 This indicates that MR-APSS is a promising method for accurately estimating causal effects in  
243 MR analyses, robust to weak IV bias and selection bias. Furthermore, MR-APSS produces  
244 narrower confidence intervals as the IV selection threshold was relaxed and weaker IVs were  
245 included in the MR analysis. These findings highlight the potential advantages of including  
246 more weak IVs in MR analysis to increase statistical power. Weighted-mode, at its default  
247 IV threshold of  $5 \times 10^{-8}$ , delivered estimates comparable to those of MR-APSS in terms of  
248 accuracy and coverage of true causal effects within the confidence intervals. Egger, while  
249 producing larger estimation errors, provided unbiased estimates when a stringent IV threshold  
250 was applied ( $5 \times 10^{-8}$ ). However, it tended to overestimate causal effects when a looser IV  
251 threshold was used. CAUSE, on the other hand, produced confidence intervals covering the  
252 true causal effect only at a stringent threshold of  $5 \times 10^{-8}$ .

253 The majority of existing MR methods, including IVW (fixed), IVW (random), MR-Lasso,  
254 cML-MA, MR-PRESSO, RAPS, Weighted-median, MR-Robust, MRMix, and MR-ConMix,  
255 displayed limitations in estimation accuracy in the presence of weak IV bias and selection  
256 bias. As the IV threshold became looser and weaker instruments were included, these methods  
257 produced estimates that were biased toward the null effect. Moreover, their confidence intervals  
258 failed to cover the true causal effects in most cases. This indicates that these methods are not  
259 capable of dealing with weak IV bias and selection bias, which compromises the accuracy of  
260 the causal effect estimates. It is crucial to acknowledge that the current strategy of using a  
261 stringent IV threshold for IV selection, such as  $5 \times 10^{-8}$ , is not a foolproof solution to address  
262 weak IV bias and selection bias. This approach has its limitations, including reduced power  
263 due to a limited number of IVs and susceptibility to weak IV bias and selection bias even with  
264 a stringent threshold. Our findings highlight the need for more robust MR methods that can  
265 effectively handle weak instruments and mitigate selection bias to accurately estimate causal  
266 effects.

267 In addition to biased causal effect estimation, methods such as MRMix and MR-Lasso have  
268 their specific limitations. MRMix exhibited some instability when varying IV thresholds. For  
269 instance, when examining the effect of height on itself, MRMix estimated the causal effect as  
270 0 with a standard error of 0.015 using 302 IVs at a threshold of  $5 \times 10^{-8}$ . Similarly, at an  
271 IV threshold of  $5 \times 10^{-6}$  with 666 IVs, MRMix again estimated the causal effect as 0 with  
272 a standard error of 0.017. However, the estimation result by MRMix at an IV threshold of  
273  $5 \times 10^{-7}$  was much more reliable. In this case, the causal effect of height on itself was estimated  
274 as 0.93 with a standard error of 0.055, utilizing 517 IVs. MR-Lasso failed to report causal  
275 estimates in some cases. For example, when testing the causal effect of WC on itself at the

276 IV threshold of  $5 \times 10^{-7}$  and that of EA on itself at the IV threshold of  $5 \times 10^{-6}$ , MR-Lasso  
 277 detected all IVs as invalid outliers and did not report any causal estimates.

Type	Threshold	Height (#SNP)	WC (#SNP)	EA (#SNP)	Hypertension (#SNP)	High cholesterol (#SNP)	Asthma (#SNP)
After LD clumping	5e-08	392	54	5	37	19	13
	5e-07	517	87	12	52	24	20
	5e-06	666	176	34	79	29	32
	5e-05	923	350	125	176	71	80
After thresholding	5e-08	6567	619	132	318	206	182
	5e-07	8631	902	164	459	301	256
	5e-06	11848	1588	342	723	416	365
	5e-05	16916	3130	770	1487	625	705

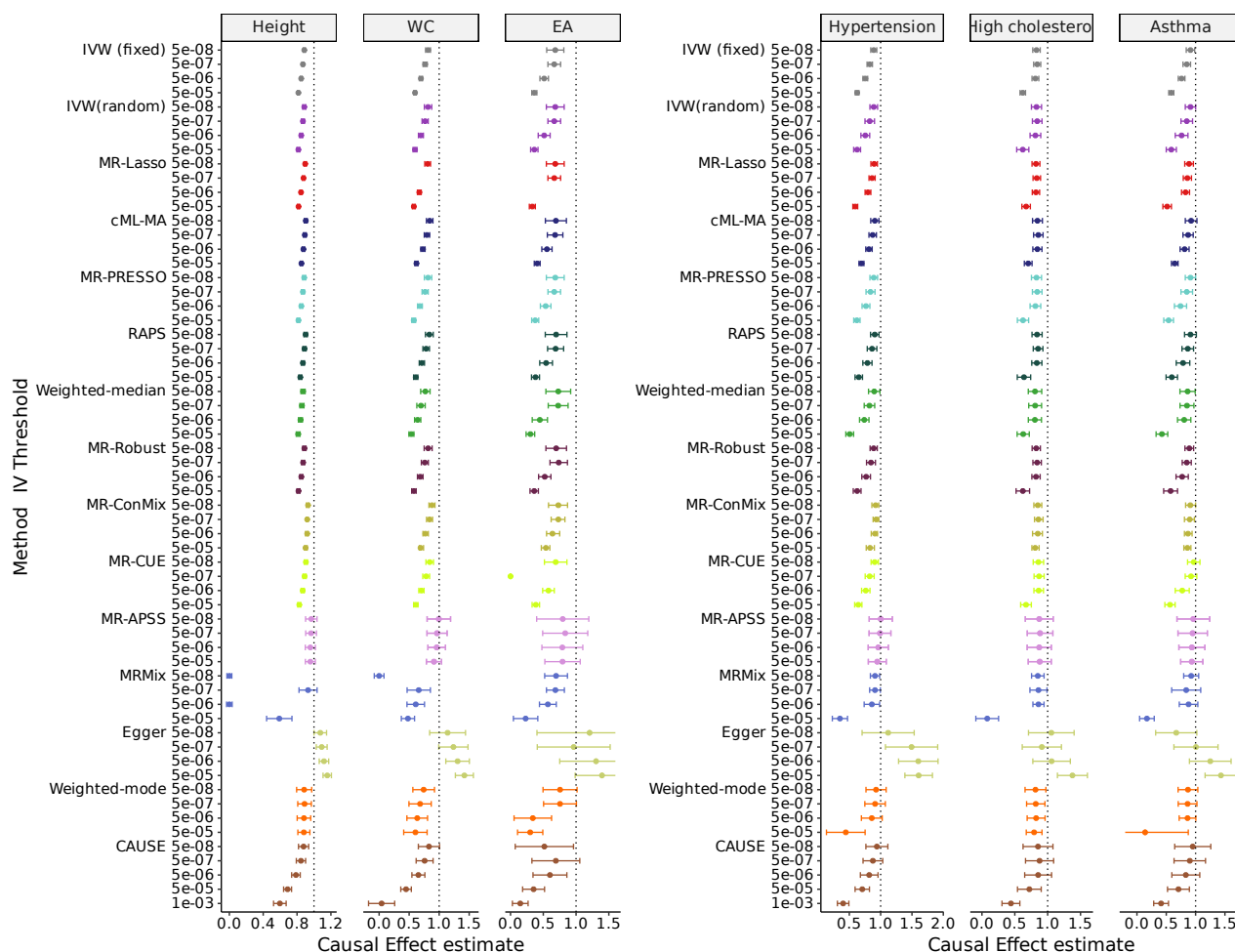


Figure 5: **Evaluation of the accuracy of causal effect estimation of 15 MR methods for six trait pairs.** Each pair comprised the same trait as both the “exposure” and the “outcome” Analyzed traits include three continuous traits, i.e., Height, Waist Circumference (WC), and Educational Attainment (EA), and three binary traits, i.e., Hypertension, High cholesterol, and Asthma. The top panel shows the number of IVs selected using different IV thresholds with/without LD clumping. The bottom panel shows the point estimates and 95% confidence intervals of different methods at different IV thresholds for the three continuous traits (bottom left) and the three binary traits (bottom right). The vertical dashed gray line represents the true causal effect size 1. Each of the 15 MR methods is represented by a distinct color.

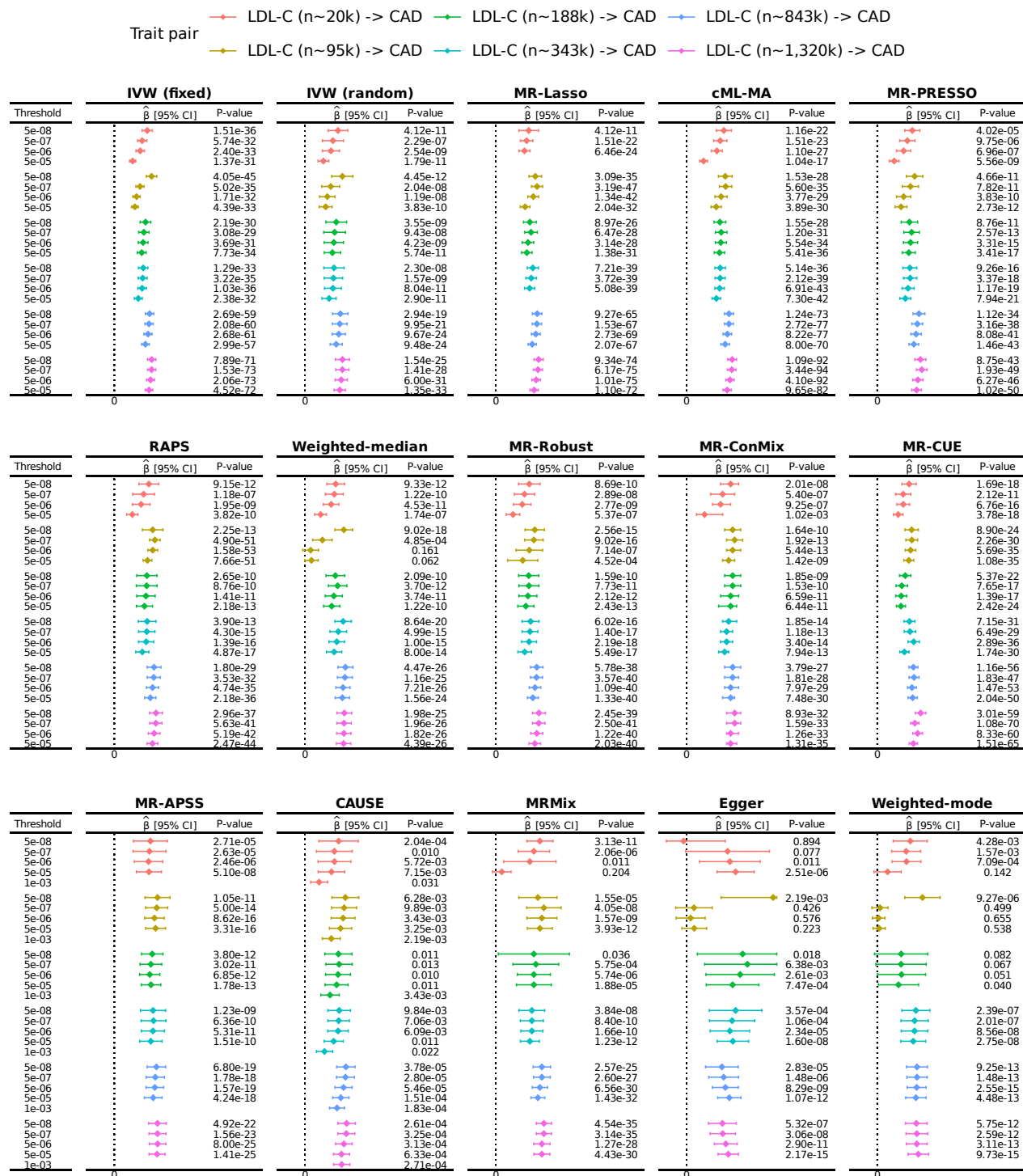


Figure 6: MR analysis results of LDL-C and CAD from 15 MR methods at different IV thresholds. Six GWAS datasets for LDL-C were used, each represented by a distinct color. The vertical dashed gray line represents the zero causal effect size. Compared methods include IVW-fixed, IVW-random, Egger, RAPS, Weighted-median, Weighted-mode, MR-PRESSO, MRMix, cML-MA, MR-Robust, MR-Lasso, MR-CUE, CAUSE, MRAPSS and MR-ConMix represented by different colors. The IV threshold is varied from  $5 \times 10^{-5}$ ,  $5 \times 10^{-6}$ ,  $5 \times 10^{-7}$  and  $5 \times 10^{-8}$ . The result of CAUSE at its default IV threshold  $1 \times 10^{-3}$  is also presented.

## 278 Replicability and power

279 To assess the replicability and power of the MR methods, we applied all the compared methods  
280 to infer the causal effect between LDL-C and CAD using six LDL-C GWAS datasets collected  
281 from five separate studies. By comparing the causal effects estimated from different GWAS  
282 datasets of the same trait, we were able to assess the reliability and generalizability of their  
283 causal effect estimates across different study designs and sample sizes. We also considered the  
284 IV thresholds varied from the stringent  $5 \times 10^{-8}$  to the relaxed  $5 \times 10^{-7}$ ,  $5 \times 10^{-6}$ , and  $5 \times 10^{-5}$   
285 to examine the sensitivity of the methods to different levels of instrument strength. The causal  
286 effect estimates, their 95% confidence intervals, and  $p$ -values produced by each method for  
287 different datasets and different IV thresholds are shown in Fig. 6.

288 Among all the compared methods, CAUSE and MR-APSS achieved outstanding performance  
289 in terms of replicability. Both CAUSE and MR-APSS are capable of producing confidence  
290 intervals that reject the null causal effect. The causal effect estimates and confidence intervals  
291 produced by both methods were highly consistent across different studies and different IV  
292 thresholds. The high consistency and generalizability of the results produced by these methods  
293 are particularly noteworthy, as they suggest that the causal effect estimates obtained using  
294 CAUSE and MR-APSS are likely to be more accurate and reliable than those obtained using  
295 other methods. However, we note that the  $p$ -values produced by CAUSE do not agree well with  
296 its confidence intervals. Consistent with previous results, the  $p$ -values produced by CAUSE are  
297 likely to be deflated. Therefore, caution should be exercised when interpreting the  $p$ -values  
298 produced by CAUSE.

299 Most of the MR methods we compared detected a significant causal relationship for  
300 all 24 tests between LDL-C and CAD using six datasets of LDL-C at four different IV  
301 thresholds. However, Egger, Weighted-mode, Weighted-median, and MRMix were unable to  
302 detect significant causal relationships in some cases. Specifically, among the 24 tests, Egger,  
303 Weighted-mode, Weighted-median, and MRMix failed to detect significant causal effects for  
304 five, four, two, and one test between LDL-C and CAD at the nominal level of 0.05, respectively.  
305 Although Egger and Weighted-mode showed good performance in terms of type I error control,  
306 our analysis revealed that Egger tended to produce estimates with large estimation errors, and  
307 Weighted-mode may have low power. Moreover, Weighted-median and MRMix, which are likely  
308 to produce false positives as shown in our previous analysis, can also lead to causal effects being  
309 wrongly shrunk to zero in some cases. However, we found that all methods, including Egger,  
310 Weighted-mode, Weighted-median, and MRMix, were able to detect significant causal effects  
311 in the cases of LDL-C( $n \sim 843k$ ) and LDL-C( $n \sim 1,320k$ ), indicating improved performance  
312 with large sample sizes. This suggests that larger sample sizes may lead to more accurate and  
313 reliable causal inference in MR analyses.

314 Consistent with previous findings, our analysis showed that most MR methods produced  
315 causal effect estimates that were sensitive to the choice of IV thresholds. Specifically, we found  
316 that causal effect estimates produced by most MR methods were closer to zero at a looser  
317 IV threshold of  $5 \times 10^{-5}$  compared to a more stringent IV threshold of  $5 \times 10^{-8}$ . However,  
318 we observed better consistency in the causal estimates produced by the MR methods across  
319 IV thresholds for the cases of LDL-C( $n \sim 843k$ ) and LDL-C( $n \sim 1,320k$ ). Importantly, as  
320 GWAS sample sizes increase to the scale of millions, the influence of weak IV bias and selection

321 bias may be greatly alleviated. Our analysis highlights the potential benefits of larger sample  
322 sizes for improving the accuracy and reliability of MR analyses. Therefore, researchers should  
323 consider using larger sample sizes in MR studies to improve the robustness of their causal  
324 inference.

## 325 Discussion

326 We present a benchmarking study of 15 two-sample summary-level data-based MR methods  
327 for causal inference. Our evaluation focuses on three crucial aspects: type I error control  
328 in the presence of various confounding scenarios (e.g., population stratification, pleiotropy,  
329 and assortative mating), the accuracy of causal effect estimates, replicability and power.  
330 Additionally, we explored the robustness of MR methods by evaluating their performance across  
331 a range of IV thresholds, assessing their ability to handle invalid IVs, and their sensitivity to  
332 IV selection. What sets our study apart is that our benchmark study is based on real-world  
333 datasets. Rather than relying on simulated or synthetic data, we carefully curated five diverse  
334 genetic datasets containing over one thousand trait pairs. The utilization of real-world datasets  
335 provides a more realistic and comprehensive evaluation of the performance of MR methods in  
336 practical scenarios.

337 Through the innovative designs of experiments that include a wide range of scenarios  
338 using real-world datasets, our study revealed that the performance of MR methods depends  
339 on underlying confounding factors that are very prevalent in real-world scenarios. Among  
340 the methods analyzed, Egger, weighted mode, and MR-APSS consistently demonstrated  
341 effective control of type I error across all three datasets representing different confounding  
342 scenarios, including population stratification, pleiotropy, and family-level confounders. However,  
343 CAUSE failed to control type I error using its default IV threshold of  $1 \times 10^{-3}$  in the  
344 presence of strong population stratification, although it exhibited deflation in other confounding  
345 scenarios. The remaining 11 MR methods displayed varying performances across the datasets  
346 representing different confounding scenarios. In the dataset representing strong population  
347 stratification (confounding scenario a), all 11 methods exhibited significant inflation of type  
348 I error. Conversely, in the datasets representing confounding scenarios of pleiotropy and  
349 family-level confounders (scenarios b and c), some methods, such as IVW (random), MR-Lasso,  
350 MR-PRESSO, RAPS, Weighted-median, IVW (fixed), cML-MA, MR-ConMix, and MR-Robust,  
351 demonstrated less severe inflation. Notably, MRMix displayed effective control of type I error  
352 in the dataset representing family-level confounders (confounding scenario c) but exhibited  
353 inflation in the dataset representing population stratification (confounding scenario a) and  
354 pleiotropy (confounding scenario b). These findings underscore the necessity of considering  
355 the characteristics of the datasets when selecting an appropriate MR method for analysis.  
356 Researchers should carefully assess the specific confounding factors present in their data and  
357 choose a method that has demonstrated robustness in handling those confounders.

358 Our study emphasized the significant impact of IV selection on the performance of MR  
359 methods. We found that using a looser threshold for IV selection resulted in inflated type  
360 I errors and increased bias in causal effect estimates for most methods. This highlights the  
361 limitations of certain MR methods in handling invalid or weak IVs and emphasizes the need to  
362 mitigate the potential bias associated with IV selection.

363 Based on our findings, we put forward the following recommendations as guidelines for best  
364 practices, aiming to assist researchers in choosing the most suitable summary-level MR methods  
365 for studying causal relationships between specific exposure-outcome trait pairs. By adhering  
366 to these guidelines, researchers can enhance the reliability and validity of their MR analyses.  
367 Firstly, we recommend conducting an analysis using negative controls. By incorporating  
368 negative controls, such as using hair colors as negative control outcomes, researchers can  
369 detect the presence of confounding bias and evaluate the robustness of different methods to  
370 confounding. This helps in selecting methods that can effectively handle confounding and  
371 provide more reliable results. Secondly, we advocate for adopting multiple standards for IV  
372 selection. Instead of relying solely on a single  $p$ -value threshold, researchers should consider  
373 various criteria and adjust the threshold accordingly to select IVs. By employing multiple  
374 standards, researchers can assess the sensitivity of MR methods to IV selection and invalid IVs.  
375 This allows for a more thorough evaluation of the methods' performance and helps prioritize  
376 methods that are robust to IV selection. Lastly, whenever feasible, we encourage researchers  
377 to gather data from multiple independent sources for the exposure and outcome of interest.  
378 This could involve incorporating data from different study populations, cohorts, or databases.  
379 By considering data from diverse sources, researchers can prioritize methods that demonstrate  
380 high replicability across multiple sources. This increases the reliability of the findings and  
381 strengthens the credibility of the robustness of the selected MR method.

382 While our benchmark study provides valuable insights into the performance of summary-  
383 level MR methods, it does have certain limitations. Firstly, the selection and measurement of  
384 confounding factors in real-world datasets can be a challenging task. Although we made careful  
385 efforts to include datasets that represented specific confounding scenarios, it is important to  
386 recognize that different types of confounders may coexist in these datasets. Secondly, due  
387 to the difficulty in collecting true positive cases from real data, we assessed the estimation  
388 accuracy of causal effect by treating the same trait as both exposure and outcome and examined  
389 replicability with a case study by employing multiple GWASs of the same exposure trait. While  
390 these strategies indirectly reflect the performance of MR methods in terms of power, a more  
391 comprehensive power analysis using multiple positive cases would provide valuable insights into  
392 the methods' ability to detect causal effects under different conditions. Thirdly, our evaluation  
393 of the estimation accuracy of MR methods utilized trait pairs where the exposure and outcome  
394 were the same trait. This design choice is currently the only possible way to ensure the true  
395 causal effects between trait pairs are known. However, we have to admit that the downside of  
396 this design is that this example does not test the methods' robustness to confounding factors  
397 like pleiotropy because the exposure and outcome are the same traits. Lastly, our benchmark  
398 study focused solely on summary-level MR methods, but it is important to recognize the  
399 availability of individual-level MR methods such as GENIUS[33] and GENIUS-MAWII [34] and  
400 MR-MiSTERI [35]. Although these methods are beyond the scope of our study, researchers  
401 should consider exploring them when they align with the study design and data availability, as  
402 they may provide additional insights and benefits in specific research contexts.



## 403 **Methods**

### 404 **Datasets for evaluation of type I error control in different confounding** 405 **scenarios**

#### 406 **Confounding scenario (a): Population stratification**

407 We aim to assess the effectiveness of MR methods in controlling type I errors in the presence of  
408 population stratification. To achieve this, we chose four hair color-related traits (Hair color:  
409 black, Hair color: blonde, Hair color: light brown, Hair color: dark brown) and skin tanning  
410 ability (Tanning) as our negative control outcomes. The GWAS summary statistics for the  
411 negative control outcomes were obtained from the GWAS ATLAS resource [30], which contains  
412 GWAS data from 600 traits in the UK Biobank. These traits were selected based on their  
413 characteristics: they are primarily determined at birth and thus are unlikely to be influenced  
414 by traits occurring after birth, and they are susceptible to confounding due to population  
415 stratification as indicated by LDSC intercept values of 1.678 (se = 0.017) for Hair color: black,  
416 1.206 (se = 0.016) for Hair color: blonde, 1.335 (se = 0.013) for Hair color: light brown, 1.510  
417 (se = 0.018) for Hair color: dark brown, and 1.916 (se = 0.020) for Tanning.

418 Next, we focused on selecting suitable exposure traits from the remaining 555 traits available  
419 in the GWAS ATLAS. We applied specific criteria to identify traits that were unrelated to  
420 hair or skin, had LDSC heritability estimates greater than 0.01, and possessed a minimum of  
421 four IVs. Through this process, we identified 226 traits that met these criteria, which we then  
422 utilized as exposure traits in our MR analysis.

423 We then applied MR methods to the 1130 exposure-outcome trait pairs formed by the  
424 selected exposure and negative control outcome traits (Supplementary data 1) and evaluated  
425 the effectiveness of MR methods in controlling for type I errors in the presence of population  
426 stratification.

#### 427 **Confounding scenario (b): Pleiotropy**

428 We aim to assess the type I error control of MR methods in the presence of confounding factors,  
429 such as pleiotropy, which can induce genetic correlation between trait pairs that are not causally  
430 linked. To accomplish this, we analyzed trait pairs consisting of 11 exposures and seven negative  
431 control outcomes. The selected exposures included five adult behavior-related traits, namely  
432 Coffee consumption [36], Instant coffee consumption [36], Ground coffee consumption [36],  
433 automobile speeding propensity [37] and risk [37], as well as six aging-related traits, including  
434 Self-rated health (<http://www.nealelab.is/uk-biobank/>, Phenotype Code: 2178), Longevity  
435 [38], Parental lifespan [39], Health span [40], Perceived age [41], and Frailty Index [42]. On the  
436 other hand, the negative control outcomes comprised seven childhood-related traits, such as  
437 Childhood aggression [43], Childhood BMI [44], Childhood intelligence [45], Fetal birth weight  
438 [46], Maternal birth weight [46], Pubertal growth (a single height measurement at age 10 in  
439 girls and 12 in boys) [47], and Comparative body size at age 10 [30]. We chose negative control  
440 outcomes based on the convention that traits developed after adulthood are unlikely to affect  
441 traits developed before adulthood causally. Consequently, causal effects between the selected  
442 exposures and negative control outcomes were considered implausible.

443 To conduct our analysis, we collected GWAS summary statistics for the exposure and  
444 outcome traits from multiple GWAS sources (detailed information can be found in Supplementary  
445 data 2). Subsequently, we examined the LDSC intercept estimates of the outcomes and  
446 calculated the genetic correlation estimates between the trait pairs (see SFig. S4). The LDSC  
447 intercepts for the outcomes were found to be approximately one, suggesting that population  
448 stratification was not a prominent confounding factor among these trait pairs. Out of the 77  
449 trait pairs analyzed, 49 pairs exhibited significant genetic correlations at the nominal level of  
450 0.05. These analyses allowed us to evaluate the performance of the MR methods in the presence  
451 of pleiotropy or other types of confounding that could induce genetic correlation between trait  
452 pairs. By considering these factors, we gained valuable insights into how well the MR methods  
453 controlled type I errors in the presence of confounders, thereby enhancing our understanding of  
454 their performance in such scenarios.

### 455 **Confounding scenario (c): Family-level confounders**

456 To assess the type I error control of MR methods in the presence of family-level confounders like  
457 assortative mating or other indirect genetic effects, we conducted an analysis using summary  
458 data obtained from a recent within-sibship GWAS study [30]. This dataset provided summary  
459 statistics for 25 traits, encompassing both within-sibship and population-based GWAS estimates.  
460 Details on these GWASs are summarized in Supplementary data 3. To ensure that the trait  
461 pairs analyzed in our study were suitable for evaluating type I errors, we required them to be  
462 genetically uncorrelated. This criterion was established to ensure that pairs with zero genetic  
463 correlation are unlikely to be causally linked, indicating the absence of a causal effect. To  
464 achieve this, we utilized LDSC [48] to estimate the genetic correlation between trait pairs  
465 among the 25 phenotypes using both population-based GWAS and within-sibship GWAS. Our  
466 selection process involved identifying 82 trait pairs (Supplementary data 3) with insignificant  
467 genetic correlation at the nominal level of 0.05 in both types of GWAS analyses. Subsequently,  
468 we applied MR methods to these selected trait pairs using both population-based GWAS and  
469 within-sibship GWAS. By comparing the results obtained from each method based on the two  
470 types of GWAS designs, we were able to examine the ability of MR methods to control for the  
471 effects of family-level confounders.

### 472 **Datasets for evaluation of the accuracy of causal effect estimates**

473 To evaluate the accuracy of the causal effect estimates of each method, we consider a special  
474 setting where the exposure and outcome are the same traits. Under a linear model setting,  
475 the genetic effects of IVs on the exposure and the outcome are the same but the effect size  
476 estimates are different. Therefore, there is no pleiotropy or other forms of confounding, and we  
477 could expect the true causal effect known to be exactly one [16, 31, 49]. Specifically, we used  
478 six traits in this setting including three continuous traits, i.e. Height, Waist Circumference  
479 (WC), and Educational attainment (EA), and three binary traits, i.e. Hypertension, High  
480 cholesterol, and Asthma. To obtain the exposure GWAS and outcome GWAS, we split the UK  
481 Biobank samples into two halves. One half was used as the exposure GWAS and the other  
482 half was used as the outcome GWAS. The sample sizes of the GWASs ranged from 121,194 to  
483 168,300 (see details in supplementary data 4). The GWAS summary statistics are obtained

484 using the BOLT-LMM software [50]. In our analysis, GWAS estimates for the binary traits are  
485 also obtained using linear models through BOLT-LMM and are then used as input for MR  
486 analysis. We could thus expect the true causal effect between the same binary exposure and  
487 binary outcome also equal one.

## 488 Datasets for evaluation of replicability and power

489 We used the example of low-density lipoprotein cholesterol (LDL-C) and coronary artery disease  
490 (CAD) for evaluation of the replicability and power of MR methods. The use of the LDL-C and  
491 CAD example in our case study provides several benefits. First, it serves as a positive control for  
492 comparing the performance of MR methods. High-level LDL-C is a well-established important  
493 risk factor for CAD. Several randomized control trials have consistently shown that lowering  
494 LDL-C levels with statins is effective in the prevention of CAD [51, 52, 53, 54, 55, 56, 57]. This  
495 allows us to evaluate the accuracy and replicability of different MR methods in a setting where  
496 we have high confidence in the existence of the positive causal effect. Second, the availability  
497 of multiple GWAS summary datasets for LDL-C provides a rich source of data for evaluating  
498 the performance of MR methods. We can thus assess the replicability of different MR methods  
499 using datasets with varying sample sizes and study designs. In our analysis, we gathered six  
500 European ancestries GWAS summary datasets for LDL-C i.e., LDL-C ( $n \sim 20k$ ) [58], LDL-C  
501 ( $n \sim 95k$ ) [59], LDL-C ( $n \sim 188k$ ) [60], LDL-C ( $n \sim 343k$ ) by the Neale Lab, LDL-C ( $n \sim 843k$ )  
502 (without UK biobank samples) and LDL-C ( $n \sim 1,320k$ ) (with UK biobank samples) [61]. The  
503 GWAS sample size increased from 19,840 in 2009 to 1.35 million in 2022. We used the same  
504 outcome GWAS for CAD which was obtained from the CARDIoGRAMplusC4D Consortium  
505 [62]. More details for the GWAS sources can be found in Supplementary data 5.

## 506 Steps of running MR methods

### 507 Step 1: quality control of GWAS summary statistics

508 The aim of the quality control step is to identify a candidate set of SNPs with high quality for  
509 IV selection and MR analysis. In our analysis, we adopted several common QC measures for  
510 GWAS summary statistics, including

- 511 • Checking missingness. For each SNP, the required data information for performing MR  
512 analysis includes SNP identifier (we use rs number), effect allele, none effect allele, effect  
513 size, standard error, sample size ( $N$ ), and  $p$ -value. SNPs missing any of the required  
514 information should be removed.
- 515 • Checking duplicates. Duplicated SNPs are SNPs with the same SNP identifier. We  
516 removed SNPs with duplicates to avoid any potential errors.
- 517 • Keeping unambiguous SNPs. We only involved unambiguous SNPs in our analysis, i.e.,  
518 SNPs with the allele types A/G, A/C, T/G, or T/C.
- 519 • Removing poorly imputed SNPs. The imputed information score (Info) is a measure of  
520 the quality of the imputed SNPs. SNPs with  $\text{Info} < 0.9$  are likely to be poorly imputed

521 and were excluded from analysis. This QC step is applicable as long as the imputed  
522 information is available in GWAS summary statistics.

523 • Removing low minor allele frequency (MAF) SNPs. Low MAF SNPs are those with MAF  
524 below a certain threshold (e.g., 0.01 or 0.05). SNPs with low MAF were excluded from  
525 analysis as they are more prone to error. The QC threshold for MAF was chosen as 0.01  
526 in our analyses. We will show later that the MR analysis results from different methods  
527 are not sensitive to the QC threshold for MAF. This QC step is applicable as long as  
528 MAF is available in GWAS summary statistics.

529 • Keeping SNPs in the set of HapMap 3 list. Because MAF or imputed information may  
530 be missing from the GWAS summary statistics, like LDSC, we restricted the analysis to  
531 a set of common and well-imputed SNPs in the HapMap 3 reference panel.

532 • Removing SNPs in the complex Major Histocompatibility Region (Chromosome 6, 26Mb  
533 – 34Mb).

534 • Removing SNPs with extremely large  $\chi^2$ . We removed SNPs with  $\chi^2 > \max\{80, N/1000\}$   
535 to reduce the undue influence of outliers on MR analysis results

536 After the QC step, GWAS datasets were formatted by retaining only the necessary data  
537 information for a set of SNPs that meet pre-determined quality control criteria. The retained  
538 data information typically includes the rs number, effect allele, non-effect allele, effect size,  
539 standard error, and  $p$ -value. It is important to note that we assume the phenotype and  
540 genotypes in GWASs are scaled to have a mean of zero and a variance of one. This scaling  
541 allows for the effect size and standard error to be calculated from  $z$ -scores and sample size,  
542 which can be more easily obtained from GWAS summary statistics.

543 The QC step is also important for methods like MR-APSS and CAUSE, which use SNPs  
544 across the genome to estimate nuisance parameters for their model.

## 545 **Step 2: harmonizing SNP effects of the exposure and outcome**

546 Performing MR analysis for an exposure-outcome trait pair requires harmonizing the effect  
547 estimates of each SNP to refer to the same allele. This is crucial for accurate MR analysis, as it  
548 ensures that the effect estimates for each SNP are comparable and can be combined to estimate  
549 the causal effect of the exposure on the outcome. To achieve this, we first checked the strands  
550 of the exposure and outcome alleles and flipped the outcome allele to the same strand as the  
551 exposure allele if they differ. We then checked the effect alleles of the exposure and outcome,  
552 and if they differed, we flipped the direction of the SNP-outcome effect to ensure that all effect  
553 estimates were aligned to the same allele. For example, if an SNP had an effect/non-effect  
554 allele of A/G in the exposure GWAS and C/T in the outcome GWAS, we first flipped the  
555 outcome allele to G/A. As the outcome GWAS presents the effect for the non-effect allele in  
556 the exposure GWAS, we then flipped the direction of the outcome effect to its opposite. Note  
557 that only unambiguous SNPs with allele types A/G, A/C, T/G, or T/C were considered, and  
558 any ambiguous SNPs were discarded from the analysis.

### 559 Step 3: IV selection and extract IV effects

560 After obtaining the harmonized summary dataset for each exposure-outcome trait pair, we  
561 began selecting instrumental variables (IVs) by identifying SNPs that were reliably associated  
562 with the exposure trait using a  $p$ -value threshold. To examine the robustness of MR methods  
563 to weak IV bias, we varied the  $p$ -value threshold for IV selection from  $5 \times 10^{-8}$  to  $5 \times 10^{-5}$ . For  
564 those methods that require independence, we further used the Plink LD clumping procedure  
565 with a threshold of  $r^2 = 0.001$  and a window size of 1 Mb to obtain a set of nearly independent  
566 SNPs from the initial set of SNPs that passed the  $p$ -value threshold. It is important to note  
567 that we required each trait pair to be analyzed with a minimum of five IVs. The final dataset  
568 containing the summary data for the selected IV set was used as input for performing MR  
569 analysis, with the goal of estimating the causal effect of the exposure on the outcome.

570 While it is common practice to select independent IVs from the exposure dataset and  
571 obtain summary data from the outcome GWAS, we perform IV selection after harmonizing the  
572 exposure and outcome datasets. This approach may reduce IV loss due to LD clumping, as  
573 selected IVs may be absent from the outcome GWAS.

### 574 Step 4: run MR methods

575 The implementation details of the 15 compared methods are described as follows:

- 576 • IVW-fixed, IVW-random, Egger, Weighted-median, and Weighted-mode were performed  
577 using the legacy version of the TwoSampleMR R package with default options (<https://github.com/MRCIEU/TwoSampleMR>).
- 579 • RAPS was performed using the `mr.raps` package without diagnostics by setting `diagnostics=F`.
- 580 • MRMix was performed using the MRMix package with default option (<https://github.com/gqi/MRMix>).
- 582 • MR-PRESSO was performed using the MRPRESSO package (<https://github.com/rondolab/MR-PRESSO>) using `OUTLIERtest = TRUE`, `DISTORTIONtest = TRUE`,  
583 `SignifThreshold = 0.05`, `seed = 1234`, and `NbDistribution = 1000` options.
- 584 • MR-Robust was performed with `lmrob` in `robustbase` R package.
- 586 • MR-Lasso and MR-ConMix are performed using the `mr_lasso` and `mr_conmix` functions,  
587 respectively, with their default options in MendelianRandomization R package.
- 588 • cML-MA was performed using R function of `mr_cML` with default options in the MRcML  
589 R package. It is important to note that we did not compare the results of the Data  
590 perturbation (DP) versions of cML-MA in our analysis. This decision was based on the  
591 consideration that the default cML-MA version (without DP) is more time-efficient.
- 592 • MR-APSS was performed using the MR-APSS (<https://github.com/YangLabHKUST/MR-APSS>) R package.
- 593 • CAUSE was performed using the CAUSE (<https://github.com/jean997/cause>) R  
594 package.
- 595

- 596 • MR-CUE was performed using the MR.CUE ([https://github.com/QingCheng0218/MR.](https://github.com/QingCheng0218/MR.CUE)  
597 CUE) R package.

## 598 **The choice of minor allele frequency threshold in quality control step**

599 One of the QC measures for GWAS summary statistics was to exclude low MAF SNPs with  
600 the concern that they are more prone to error. Typically, studies use MAF thresholds of 0.01  
601 or 0.05. To assess the impact of the MAF threshold and to determine an appropriate MAF  
602 threshold for MR, we conducted an analysis to explore the effect of the MAF threshold by  
603 varying the MAF threshold. Specifically, we considered the analysis of the six UK Biobank  
604 trait pairs used to evaluate causal effect estimation. We used MAF thresholds of 0.01 and 0.05  
605 in the QC step for the summary statistics, and we then applied MR methods to the formatted  
606 summary datasets with different MAF QC thresholds. Results from different MR methods  
607 using different MAF thresholds are given in Sfig. S2. Our analysis shows that MR methods are  
608 generally not sensitive to the choice of MAF thresholds. However, to obtain more candidate  
609 IVs, we chose a threshold of 0.01 for MAF QC in our MR analysis.

## 610 **The choice of reference panel for LD clumping**

611 All of the compared MR methods, except for MR-CUE, require independent or weakly correlated  
612 IVs. For those methods, an LD reference panel was used to perform LD clumping in the IV  
613 selection step. In contrast, MR-CUE allows for correlated IVs, and an LD reference panel is  
614 used to model the correlation between SNPs. To examine whether MR methods are sensitive to  
615 the choice of LD reference panel, we conducted a sensitivity analysis by comparing the results  
616 obtained using different reference panels. Specifically, we used an in-sample UK Biobank LD  
617 reference panel and the 1000 Genomes reference panel of European ancestry for the six UK  
618 Biobank trait pairs used to evaluate causal effect estimation. Both reference panels are of the  
619 same ancestry as the study population. We present the results from different MR methods  
620 using different LD reference panels in Supplementary Figure Sfig. S3. Our analysis shows that  
621 MR methods are generally not sensitive to the choice of LD reference panel as long as the  
622 panels are from the same ancestry.

## 623 **Data availability**

624 The UK Biobank data are from UK Biobank resources under application number 30186. All  
625 GWAS summary statistics used in this study are downloadable at [https://github.com/](https://github.com/YangLabHKUST/MRbenchmarking)  
626 [YangLabHKUST/MRbenchmarking](https://github.com/YangLabHKUST/MRbenchmarking). Supplementary Data 1-3, and 5 provide the references of  
627 these datasets.

## 628 **Code availability**

629 The source codes to reproduce all the analyses can be accessed at the following location:  
630 <https://github.com/YangLabHKUST/MRbenchmarking>.

## 631 Acknowledgements

632 We acknowledge the following grants: Hong Kong Research Grant Council grants nos. 16301419,  
633 16308120, 16307221 and 16307322, Hong Kong University of Science and Technology Startup  
634 Grants R9405 and Z0428 from the Big Data Institute, Guangdong-Hong Kong-Macao Joint  
635 Laboratory grant no. 2020B1212030001 and the RGC Collaborative Research Fund grant no.  
636 C6021-19EF to C.Y., City University of Hong Kong Startup Grant 7200746 and Strategic  
637 Research Grant 21300423 to M.C.

## 638 References

- 639 [1] Lars Bondemark and Sabine Ruf. Randomized controlled trial: the gold standard or an  
640 unobtainable fallacy? *European Journal of Orthodontics*, 37(5):457–461, 2015.
- 641 [2] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable  
642 approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, 2007.
- 643 [3] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for  
644 causal inference in epidemiological studies. *Human Molecular Genetics*, 23, 2014.
- 645 [4] Kaitlin H Wade, James Yarmolinsky, Edward Giovannucci, Sarah J Lewis, Iona Y Millwood,  
646 Marcus R Munafò, Fleur Meddens, Kimberley Burrows, Joshua A Bell, Neil M Davies,  
647 et al. Applying mendelian randomization to appraise causality in relationships between  
648 nutrition and cancer. *Cancer Causes & Control*, 33(5):631–652, 2022.
- 649 [5] Jean-Baptiste Pingault, Paul F O’reilly, Tabea Schoeler, George B Ploubidis, Frühling  
650 Rijdsdijk, and Frank Dudbridge. Using genetic data to strengthen causal inference in  
651 observational research. *Nature Reviews Genetics*, 19(9):566–580, 2018.
- 652 [6] Stephen Burgess, Amy M Mason, Andrew J Grant, Eric AW Slob, Apostolos Gkatzionis,  
653 Verena Zuber, Ashish Patel, Haodong Tian, Cunhao Liu, William G Haynes, et al. Using  
654 genetic association data to guide drug discovery and development: Review of methods  
655 and applications. *The American Journal of Human Genetics*, 110(2):195–214, 2023.
- 656 [7] Amand F Schmidt, Chris Finan, Maria Gordillo-Marañón, Folkert W Asselbergs, Daniel F  
657 Freitag, Riyaz S Patel, Benoît Tyl, Sandesh Chopade, Rupert Faraway, Magdalena  
658 Zwierzyna, et al. Genetic drug target validation using mendelian randomisation. *Nature*  
659 *communications*, 11(1):3255, 2020.
- 660 [8] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George  
661 Davey Smith. Mendelian randomization: using genes as instruments for making causal  
662 inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.
- 663 [9] Vanessa Didelez, Sha Meng, and Nuala A. Sheehan. Assumptions of IV Methods for  
664 Observational Epidemiology. *Statistical Science*, 25(1):22 – 40, 2010.

- 665 [10] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread  
666 horizontal pleiotropy in causal relationships inferred from Mendelian randomization between  
667 complex traits and diseases. *Nature genetics*, 50(5):693, 2018.
- 668 [11] Haoran Xue, Xiaotong Shen, and Wei Pan. Constrained maximum likelihood-based  
669 Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects.  
670 *The American Journal of Human Genetics*, 108(7):1251–1269, 2021.
- 671 [12] Jessica MB Rees, Angela M Wood, Frank Dudbridge, and Stephen Burgess. Robust  
672 methods in mendelian randomization via penalization of heterogeneous causal estimates.  
673 *PloS one*, 14(9):e0222362, 2019.
- 674 [13] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent  
675 estimation in mendelian randomization with some invalid instruments using a weighted  
676 median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.
- 677 [14] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference  
678 in summary data mendelian randomization via the zero modal pleiotropy assumption.  
679 *International journal of epidemiology*, 46(6):1985–1998, 2017.
- 680 [15] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization  
681 with invalid instruments: effect estimation and bias detection through Egger regression.  
682 *International journal of epidemiology*, 44(2):512–525, 2015.
- 683 [16] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small.  
684 Statistical inference in two-sample summary-data mendelian randomization using robust  
685 adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769, 2020.
- 686 [17] Jia Zhao, Jingsi Ming, Xianghong Hu, Jin Liu, and Can Yang. Bayesian Weighted  
687 Mendelian Randomization for Causal Inference based on Summary Statistics. *arXiv*  
688 *preprint arXiv:1811.10223*, 2018.
- 689 [18] Jean Morrison, Nicholas Knoblauch, Joseph H. Marcus, Matthew Stephens, and Xin He.  
690 Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects  
691 using genome-wide summary statistics. *Nature Genetics*, 2020.
- 692 [19] Xianghong Hu, Jia Zhao, Zhixiang Lin, Yang Wang, Heng Peng, Hongyu Zhao, Xiang Wan,  
693 and Can Yang. Mendelian randomization for causal inference accounting for pleiotropy  
694 and sample structure using genome-wide summary statistics. *Proceedings of the National*  
695 *Academy of Sciences*, 119(28):e2106858119, 2022.
- 696 [20] Guanghao Qi and Nilanjan Chatterjee. Mendelian randomization analysis using mixture  
697 models for robust and efficient estimation of causal effects. *Nature Communications*,  
698 10(1):1941, 2019.
- 699 [21] Stephen Burgess, Christopher N Foley, Elias Allara, James R Staley, and Joanna MM  
700 Howson. A robust and efficient method for mendelian randomization with hundreds of  
701 genetic variants. *Nature communications*, 11(1):376, 2020.



- 702 [22] Qing Cheng, Xiao Zhang, Lin S Chen, and Jin Liu. Mendelian randomization accounting  
703 for complex correlated horizontal pleiotropy while elucidating shared genetic etiology.  
704 *Nature Communications*, 13(1):6490, 2022.
- 705 [23] Ben Brumpton, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison,  
706 Gunnhild Åberge Vie, Yoonsu Cho, Laura D Howe, Amanda Hughes, Dorret I Boomsma,  
707 et al. Avoiding dynastic, assortative mating, and population stratification biases  
708 in mendelian randomization through within-family analyses. *Nature communications*,  
709 11(1):3519, 2020.
- 710 [24] Eleanor Sanderson, Tom G Richardson, Gibran Hemani, and George Davey Smith. The use  
711 of negative control outcomes in Mendelian Randomisation to detect potential population  
712 stratification or selection bias. *International Journal of Epidemiology*, 50(4):1350–1361,  
713 2021.
- 714 [25] Fernando Pires Hartwig, Neil Martin Davies, and George Davey Smith. Bias in mendelian  
715 randomization due to assortative mating. *Genetic epidemiology*, 42(7):608–620, 2018.
- 716 [26] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller.  
717 Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 2013.
- 718 [27] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization  
719 analysis with multiple genetic variants using summarized data. *Genetic epidemiology*,  
720 37(7):658–665, 2013.
- 721 [28] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala Sheehan,  
722 and John Thompson. A framework for the investigation of pleiotropy in two-sample  
723 summary data mendelian randomization. *Statistics in medicine*, 36(11):1783–1802, 2017.
- 724 [29] Alexander I Young, Stefania Benonisdottir, Molly Przeworski, and Augustine Kong.  
725 Deconstructing the sources of genotype-phenotype associations in humans. *Science*,  
726 365(6460):1396–1400, 2019.
- 727 [30] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed,  
728 Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A Lind, Teemu Palviainen, et al.  
729 Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic  
730 effects. *Nature genetics*, 54(5):581–592, 2022.
- 731 [31] Zhongshang Yuan, Lu Liu, Ping Guo, Ran Yan, Fuzhong Xue, and Xiang Zhou.  
732 Likelihood-based mendelian randomization analysis with automated instrument selection  
733 and horizontal pleiotropic modeling. *Science Advances*, 8(9):eabl5744, 2022.
- 734 [32] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model  
735 association for biobank-scale datasets. *Nature genetics*, 50(7):906–908, 2018.
- 736 [33] Eric Tchetgen Tchetgen, BaoLuo Sun, and Stefan Walter. The GENIUS approach to  
737 robust Mendelian randomization inference. *Statistical Science*, 36(3):443–464, 2021.

- 738 [34] Ting Ye, Zhonghua Liu, Baoluo Sun, and Eric Tchetgen Tchetgen. GENIUS-MAWII: For  
739 Robust Mendelian Randomization with Many Weak Invalid Instruments. *arXiv preprint*  
740 *arXiv:2107.06238*, 2021.
- 741 [35] Zhonghua Liu, Ting Ye, Baoluo Sun, Mary Schooling, and Eric Tchetgen Tchetgen. On  
742 mendelian randomization mixed-scale treatment effect robust identification (mr misteri)  
743 and estimation for causal inference. *arXiv preprint arXiv:2009.14484*, 2020.
- 744 [36] Nicola Pirastu, Ciara McDonnell, Eryk J Grzeszkowiak, Ninon Mounier, Fumiaki Imamura,  
745 Jordi Merino, Felix R Day, Jie Zheng, Nele Taba, Maria Pina Concas, et al. Using genetic  
746 variation to disentangle the complex relationship between food intake and health outcomes.  
747 *PLoS Genetics*, 18(6):e1010162, 2022.
- 748 [37] Richard Karlsson Linnér, Pietro Biroli, Edward Kong, S Fleur W Meddens, Robbee  
749 Wedow, Mark Alan Fontana, Maël Lebreton, Stephen P Tino, Abdel Abdellaoui, Anke R  
750 Hammerschlag, et al. Genome-wide association analyses of risk tolerance and risky  
751 behaviors in over 1 million individuals identify hundreds of loci and shared genetic  
752 influences. *Nature genetics*, 51(2):245–257, 2019.
- 753 [38] Joris Deelen, Daniel S Evans, Dan E Arking, Niccolò Tesi, Marianne Nygaard, Xiaomin  
754 Liu, Mary K Wojczynski, Mary L Biggs, Ashley van Der Spek, Gil Atzmon, et al. A  
755 meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nature*  
756 *communications*, 10(1):3669, 2019.
- 757 [39] Paul RHJ Timmers, Ninon Mounier, Kristi Lall, Krista Fischer, Zheng Ning, Xiao Feng,  
758 Andrew D Bretherick, David W Clark, Xia Shen, et al. Genomics of 1 million parent  
759 lifespans implicates novel pathways and common diseases and distinguishes survival chances.  
760 *elife*, 8:e39856, 2019.
- 761 [40] Aleksandr Zenin, Yakov Tsepilov, Sodbo Sharapov, Evgeny Getmantsev, LI Menshikov,  
762 Peter O Fedichev, and Yurii Aulchenko. Identification of 12 genetic loci associated with  
763 human healthspan. *Communications biology*, 2(1):41, 2019.
- 764 [41] Victoria Roberts, Barry Main, Nicholas J Timpson, and Simon Haworth. Genome-wide  
765 association study identifies genetic associations with perceived age. *Journal of Investigative*  
766 *Dermatology*, 140(12):2380–2385, 2020.
- 767 [42] Janice L Atkins, Juulia Jylhävä, Nancy L Pedersen, Patrik K Magnusson, Yi Lu, Yunzhang  
768 Wang, Sara Hägg, David Melzer, Dylan M Williams, and Luke C Pilling. A genome-wide  
769 association study of the frailty index highlights brain pathways in ageing. *Aging Cell*,  
770 20(9):e13459, 2021.
- 771 [43] Irene Pappa, Beate St Pourcain, Kelly Benke, Alana Cavadino, Christian Hakulinen,  
772 Michel G Nivard, Ilja M Nolte, Carla MT Tiesler, Marian J Bakermans-Kranenburg,  
773 Gareth E Davies, et al. A genome-wide approach to children’s aggressive behavior:  
774 The eagle consortium. *American Journal of Medical Genetics Part B: Neuropsychiatric*  
775 *Genetics*, 171(5):562–572, 2016.

- 776 [44] Suzanne Voegeleang, Jonathan P Bradfield, Tarunveer S Ahluwalia, John A Curtin,  
777 Timo A Lakka, Niels Grarup, Markus Scholz, Peter J Van der Most, Claire Monnereau,  
778 Evie Stergiakouli, et al. Novel loci for childhood body mass index and shared heritability  
779 with adult cardiometabolic traits. *PLoS genetics*, 16(10):e1008718, 2020.
- 780 [45] Beben Benyamin, BSt Pourcain, Oliver S Davis, Gail Davies, Narelle K Hansell, M-JA  
781 Brion, RM Kirkpatrick, Rolieke AM Cents, Sanja Franić, MB Miller, et al. Childhood  
782 intelligence is heritable, highly polygenic and associated with *fnbp1l*. *Molecular psychiatry*,  
783 19(2):253–258, 2014.
- 784 [46] Nicole M Warrington, Robin N Beaumont, Momoko Horikoshi, Felix R Day, Øyvind  
785 Helgeland, Charles Laurin, Jonas Bacelis, Shouneng Peng, Ke Hao, Bjarke Feenstra, et al.  
786 Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic  
787 risk factors. *Nature genetics*, 51(5):804–814, 2019.
- 788 [47] Diana L Cousminer, Diane J Berry, Nicholas J Timpson, Wei Ang, Elisabeth Thiering,  
789 Enda M Byrne, H Rob Taal, Ville Huikari, Jonathan P Bradfield, Marjan Kerkhof, et al.  
790 Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal  
791 height growth, pubertal timing and childhood adiposity. *Human molecular genetics*,  
792 22(13):2735–2747, 2013.
- 793 [48] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang,  
794 Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson,  
795 Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes  
796 confounding from polygenicity in genome-wide association studies. *Nature genetics*,  
797 47(3):291–295, 2015.
- 798 [49] Jingshu Wang, Qingyuan Zhao, Jack Bowden, Gibran Hemani, George Davey Smith,  
799 Dylan S Small, and Nancy R Zhang. Causal inference for heritable phenotypic risk factors  
800 using heterogeneous genetic instruments. *PLoS genetics*, 17(6):e1009575, 2021.
- 801 [50] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsón, Hilary K  
802 Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie  
803 Berger, et al. Efficient bayesian mixed-model analysis increases association power in large  
804 cohorts. *Nature genetics*, 47(3):284–290, 2015.
- 805 [51] Scandinavian Simvastatin Survival Study Group et al. Randomised trial of cholesterol  
806 lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival  
807 study (4s). *The Lancet*, 344(8934):1383–1389, 1994.
- 808 [52] C Packard, J Shepherd, S Cobbe, I Ford, CG Isles, JH McKillop, PW Macfarlane,  
809 AR Lorimer, and J Norrie. Influence of pravastatin and plasma lipids on clinical events in  
810 the west of scotland coronary prevention study (woscops). *Circulation*, 97(15):1440–1445,  
811 1998.
- 812 [53] Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group.  
813 Prevention of cardiovascular events and death with pravastatin in patients with coronary

- 814 heart disease and a broad range of initial cholesterol levels. *New England Journal of*  
815 *Medicine*, 339(19):1349–1357, 1998.
- 816 [54] National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation,  
817 and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third  
818 report of the national cholesterol education program (ncep) expert panel on detection,  
819 evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii)  
820 final report. *Circulation*, 106(25):3143–3421, December 2002.
- 821 [55] Heart Protection Study Collaborative Group et al. Mrc/bhf heart protection study  
822 of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised  
823 placebocontrolled trial. *The Lancet*, 360(9326):7–22, 2002.
- 824 [56] Epidemiological Studies Unit. Efficacy and safety of cholesterol-lowering treatment:  
825 prospective meta-analysis of data from 90 056 participants in 14 randomised trials of  
826 statins. *Lancet*, 366(9493):1267–1278, 2005.
- 827 [57] Haruo Nakamura, Kikuo Arakawa, Hiroshige Itakura, Akira Kitabatake, Yoshio Goto,  
828 Takayoshi Toyota, Noriaki Nakaya, Shoji Nishimoto, Masaharu Muranaka, Akira  
829 Yamamoto, et al. Primary prevention of cardiovascular disease with pravastatin in japan  
830 (mega study): a prospective randomised controlled trial. *The Lancet*, 368(9542):1155–1163,  
831 2006.
- 832 [58] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru,  
833 Eric E Schadt, Lee Kaplan, Derrick Bennett, Yun Li, Toshiko Tanaka, et al. Common  
834 variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*, 41(1):56–65,  
835 2009.
- 836 [59] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M  
837 Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman,  
838 Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood  
839 lipids. *Nature*, 466(7307):707–713, 2010.
- 840 [60] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson,  
841 Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, et al.  
842 Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274,  
843 2013.
- 844 [61] Sarah E Graham, Shoa L Clarke, Kuan-Han H Wu, Stavroula Kanoni, Greg JM Zajac,  
845 Shweta Ramdas, Ida Surakka, Ioanna Ntalla, Sailaja Vedantam, Thomas W Winkler,  
846 et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*,  
847 600(7890):675–679, 2021.
- 848 [62] the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 genomes–based genome-  
849 wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121–  
850 1130, 2015.