Likelihood based Mendelian randomization analysis with automated instrument selection and horizontal pleiotropic modeling

Zhongshang Yuan[1,2], Lu Liu[1,2], Ping Guo[1,2], Ran Yan[1,2], Fuzhong Xue[1,2], Xiang Zhou[3,4*]

[1] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, 250012 Jinan, Shandong, China;

[2] Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, 250012 Jinan, Shandong, China;

[3] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA;

[4] Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

*Corresponding author. Email: xzhousph@umich.edu

Abstract

Mendelian randomization (MR) is a common tool for identifying causal risk factors underlying diseases. Here, we present a method, MRAID, for effective MR analysis. MRAID borrows ideas from fine mapping analysis to model an initial set of candidate SNPs that are in potentially high linkage disequilibrium with each other and automatically selects among them the suitable instruments for causal inference. MRAID also explicitly models both uncorrelated and correlated horizontal pleiotropic effects that are widespread for complex trait analysis. MRAID achieves both tasks through a joint likelihood framework and relies on a scalable sampling-based algorithm to compute calibrated *p*-values. Comprehensive and realistic simulations show MRAID can provide calibrated type I error control, reduce false positives, while being more powerful than existing approaches. We illustrate the benefits of MRAID for an MR screening analysis across 645 trait pairs in UK Biobank, identifying multiple lifestyle causal risk factors of cardiovascular disease-related traits.

Introduction

Investigating causal relationship among complex traits and identifying causal risk factors are an important first step towards understanding the biology of diseases. A common statistical tool for performing such causal inference in observational studies is Mendelian randomization (MR). MR is a form of instrumental variable analysis that uses SNPs to serve as instruments for inferring the causal effect of an exposure variable on an outcome variable (*1*). MR requires only summary statistics from genome-wide association studies (GWASs) and is often performed in a two-sample study setting where the exposure variable and the outcome variable are measured in two separate studies (*2*). With the abundant availability of GWAS summary statistics, numerous MR analyses are being carried out, identifying important causal risk factors for various common diseases. These MR studies are facilitated by many recently developed MR methods that include the inverse variance weighted (IVW) method, MR-Egger (*3*), median-based regression (*4*), BWMR (*5*), RAPS (*6*), MRMix (*7*), CAUSE (*8*), to name a few. Different MR methods differ in their modeling assumptions and inference algorithms, but the majority of them encounter two important modeling and algorithmic challenges that have so far limited the effectiveness of MR analysis.

First, almost all existing MR methods rely on a pre-selected set of independent SNPs to serve as instruments for MR analysis. The instruments are selected to be independent from each other to ensure the validity of the statistical inference framework used in many common MR methods such as IVW. The independent SNPs are often selected through linkage disequilibrium (LD) clumping, a procedure that first ranks SNPs based on their marginal association evidence with the exposure variable and then retains SNPs that are not in high LD with the SNPs on top of the ranking list. Using LD clumping to select SNPs may be suboptimal, however, as the selected SNPs may only

represent tagging SNPs that are in LD with the causal SNPs rather than the causal ones themselves. The tagging SNPs obtained through clumping can sometimes be quite far away from the causal ones (*9-12*), often explain a much smaller phenotypic variance than expected (*13-15*), and have limited prediction power on the outcome trait (*16*). Indeed, the parallel research field of GWAS fine-mapping highlights the benefits of performing SNP selection through formal modeling based approaches to refine the set of potentially causal associations (*9-11, 17-19*). Consequently, using a formal SNP selection procedure to identify instruments instead of directly using the tagging SNPs via clumping may help increase the power of MR analysis. In addition, perhaps more importantly, selecting independent SNPs for MR analysis may not be ideal either, as complex traits can be influenced by multiple causal SNPs residing in the same local region that are in potential LD with each other. Consequently, selecting independent SNPs may only capture a small proportion of the phenotypic variance in the exposure variable (*14, 15*), again leading to a loss of power in the subsequent MR analysis (*1, 2, 20, 21*). Indeed, in the parallel research field of transcriptome-wide association studies (TWAS), it has been well documented that incorporating correlated SNPs can substantially improve gene expression prediction accuracy (*22*), and consequently TWAS analysis power, than using independent SNPs only (*14, 23-25*). Therefore, incorporating correlated SNPs and developing effective approaches to select instruments among them are important to fully captivate the potential of MR.

Second, only a limited number of MR methods model horizontal pleiotropy and even fewer can effectively control for it during MR analysis (*26*). Horizontal pleiotropy occurs when the SNP instruments exhibit effects on the outcome through pathways other than the exposure. Horizontal pleiotropy has been widely observed in complex trait analysis (*14, 26*) and often comes in two distinct types. The first type of horizontal

pleiotropy arises through paths independent of the exposure, with the resulting horizontal pleotropic effects being independent of the SNP effects on the exposure. The second type of horizontal pleiotropy arises through unobserved exposure-outcome confounders and induces correlation between the horizontal pleotropic effects and the SNP effects on the exposure. For example, insulin resistance (IR) occurs when excess glucose in the blood reduces the ability of the cells to absorb and use blood sugar for energy. IR is an important cause of type II diabetes (T2D) (*27*), and also has profound effects on lipoproteins such as low density lipoprotein (LDL) (*28*). When investigating the causal effect of LDL on T2D, the selected candidate instrumental SNPs may include IR-associated SNPs. These IR-associated SNPs are likely to be associated with both LDL and T2D, leading to correlated pleiotropy in the MR analysis. The presence of either type of horizontal pleiotropy violates standard MR modeling assumptions and can lead to biased causal effect estimates and increased false discoveries. Early MR analyses control for horizontal pleiotropy by simply removing instrumental SNPs that are potentially associated with the outcome variable (*26, 29-31*). Removing SNPs associated with the outcome would result in a conservative set of selected instruments and lead to a loss of power in the subsequent MR analysis. Recent MR methods explicitly model horizontal pleiotropy by specifying modeling assumptions on the horizontal pleiotropic effects. For example, the Egger assumption assumes the same horizontal pleiotropic effect across SNP instruments (*3, 14*), while PMR-VC (*14*) and BWMR (*5*) assume the horizontal pleiotropic effects to follow a normal distribution; all these methods model the first type of horizontal pleiotropy. MRMix (*7*) and CAUSE (*8*), by contrast, employ a normal-mixture model to control for both types of horizontal pleiotropy. Unfortunately, modeling both types of horizontal pleiotropy has been technically challenging, as the resulting likelihood function of the MR model often

consists of an integration that cannot be solved analytically. Consequently, both MRMix and CAUSE rely on non-likelihood based approaches to perform MR inference. Specifically, MRMix searches on a grid of causal effect candidates to identify the one that maximizes the proportion of GWAS summary statistics residing in the expected sub-model without horizontal pleiotropy. CAUSE contrasts the out-of-sample prediction accuracy between two different models, one with the causal effect and the other without, by computing the expected log pointwise posterior density between the two, for causal inference. Non-likelihood based causal inference, however, can lead to a loss of power and/or uncalibrated test statistics that are essential for large-scale screening of causal risk factors underlying diseases. Indeed, as we will show here, MRMix is not robust to modeling misspecifications on the instrumental effect sizes and is prone to estimation bias, while CAUSE yields overly conservative *p*-values.

Here, we present a likelihood-based two-sample MR method for causal inference that overcomes the above two challenges. Specifically, our method models an initial set of candidate SNP instruments that are in high LD with each other and automatically selects among them the suitable instruments for MR analysis. In addition, our method accounts for both types of horizontal pleiotropy in a likelihood framework and relies on a scalable sampling-based algorithm for calibrated *p*-values computation. We refer to our method as the two-sample Mendelian Randomization with Automated Instrument Determination (MRAID). We demonstrate the effectiveness of MRAID through comprehensive and realistic simulations. We also apply MRAID for an MR screening analysis across 645 trait pairs in the UK Biobank (*32*), identifying lifestyle risk factors that may causally influence cardiovascular disease-related traits.

Results

Method overview and simple illustrative simulations

MRAID is described in the Materials and Methods, with its technical details provided in the Supplementary Text and a method schematic shown in Fig. 1. Briefly, MRAID is a two-sample MR method that aims to infer the causal effect of an exposure variable on an outcome variable using GWAS summary statistics. MRAID models jointly all genome-wide significant SNPs that are in potential LD with each other and performs automated instrument selection among them to identify suitable instruments for MR analysis. In addition, MRAID explicitly accounts for two types of horizontal pleiotropic effects through a maximum likelihood-based inference framework and is scalable to biobank datasets (Table 1).

We first performed simple simulations to develop intuition and illustrate the benefits of modeling multiple correlated SNPs (details in Materials and Methods). Here, we only compared MRAID with the MR method that uses only the top exposure-associated SNP (i.e. lead variant) for MR analysis. Consistent with the fine-mapping literature (*9-12*), the lead variant is the casual SNP in only 53.6% of the simulation replicates. As a result, MRAID produces calibrated type I error control, while the MR method using only the lead variant produces slightly deflated *p*-values (Fig. S1). In addition, MARID is more powerful than the MR method using only the lead variant regardless of the number of causal SNPs and the LD structure among different SNPs, the presence or absence of the causal SNPs (Fig. S2). Similar results are observed when comparing MRAID that uses the top two SNPs obtained from a stepwise regression with MRAID that uses only the lead SNP (Fig. S3). These results highlight the benefits of modeling correlated SNPs and performing SNP selection for MR analysis.

Simulations: Type I error control

We performed comprehensive and realistic simulations to evaluate the performance of MRAID and compare it with seven existing MR methods (details in Materials and Methods). We first examined type I error control of different methods in different scenarios. In the absence of both correlated and uncorrelated horizontal pleiotropic effects, most methods, including MRAID, IVW-R, Robust, RAPS, Weighted median and MRMix, all yield reasonably calibrated type I error control (Fig. 2A). Weighted mode and CAUSE, on the other hand, display overly conservative type I error control, which is consistent with the original studies (*7, 8, 33*). The null *p*-value distributions from different methods remain largely similar regardless of the number of SNPs that affect the exposure (Fig. S4A) and their total effects on the exposure (Fig. S4B). We further examined the robustness of different methods in settings where the SNP effects on the exposure do not follow a simple normal distribution but with some SNPs displaying larger effects than the others. In these settings, MRAID, IVW-R and RAPS remain calibrated, while both MRMix and Robust method show inflated type I errors, presumably due to their restricted normality assumptions on the SNP effect sizes (Fig. 2B). Note that we directly used correlated SNPs for MRAID but performed clumping to select independent SNPs for the other methods. Without clumping, all other MR methods produce overly inflated type I errors (Fig. S5).

We examined the effects of horizontal pleiotropy on type I error control for different methods. When horizontal pleiotropic effects are present but are uncorrelated with the instrumental effects, MRAID maintains type I error control (Fig. 2C). In contrast, both Weighted mode and CAUSE remain overly conservative, while MRMix, Robust, IVW-R, Weighted median and RAPS yield inflated *p*-values (Fig. 2C). Similar conclusion holds regardless of the effect size for the uncorrelated horizontal pleiotropy or the

proportion of SNPs that display uncorrelated pleiotropic effects (Fig. S6). The *p*-value inflation problem of MRMix and Weighted median relieves when the proportion of SNPs that display uncorrelated horizontal pleiotropic effects decreases. When correlated horizontal pleiotropic effects are also present in addition to the uncorrelated horizontal pleiotropic effects, MRAID maintains effective type I error control (Fig. 2D). In contrast, both Weighted mode and CAUSE remain overly conservative, while MRMix, Robust, IVW-R, Weighted median and RAPS produce inflated *p*-values. Similar conclusion holds regardless of the effect size of the correlated horizontal pleiotropy (Fig. S7A *vs* Fig. S7C), the proportion of SNPs that display uncorrelated horizontal pleiotropic effects (Fig. S7A *vs* Fig. S7B), the proportion of SNPs that display correlated horizontal pleiotropic effects (Fig. S7A *vs* Fig. S7D), or how the correlated horizontal pleiotropic effects are created (Fig. S8).

Simulations: Power comparison

We examined the power of different MR methods to detect non-zero causal effect. Because the same *p*-value from different methods may correspond to different type I errors, we computed power based on an false discovery rate (FDR) of 0.05 instead of a nominal *p*-value threshold to allow for fair comparison among methods. In the absence of both uncorrelated and correlated horizontal pleiotropic effects, MRAID, IVW-R and RAPS all have high power across different scenarios. Among these three methods, MRAID is slightly more powerful than the other two when the instrumental effects are small or when the causal effect is small (Fig. 3A, 3B), presumably due to the automated instrument selection procedure employed in MRAID. MRAID is slightly less powerful than the other two when the instrumental effects are large and the causal effect is large (Fig. S9), as the simple instrumental selection approaches used in the other methods

can be effective in these lesser challenging settings. The performance of these three methods is generally followed by Robust. While Weighted mode, MRMix, and, to a lesser extent, CAUSE, have low power.

We examined the influence of horizontal pleiotropy on the power of different methods. When horizontal pleiotropic effects are present but are uncorrelated with the instrumental effects, MRAID is more powerful than the other MR methods (Fig. 3C). The power gain brought by MRAID becomes more apparent with increasing horizontal pleiotropy, which is characterized by increased horizontal pleiotropic effect sizes and/or increased proportion of SNPs that display horizontal pleiotropic effects (Fig. S10). The performance of MRAID is often followed by RAPS, Robust, CAUSE, IVW-R and Weighted median, while MRMix and Weighted mode generally have low power (Fig. S10). Among these methods, the performance of IVW-R is particularly sensitive to the horizontal pleiotropic effect sizes or the proportion of SNPs that display horizontal pleiotropic effects. When correlated horizontal pleiotropic effects are also present in addition to the uncorrelated horizontal pleiotropic effects, the power of MRAID remains higher than the other methods. The higher power of MRAID maintains regardless of the correlated horizontal pleiotropic effect sizes, the proportion of instrumental SNPs that display correlated horizontal pleiotropic effects (Fig. 3D, Fig. S10D-F), or how the correlated horizontal pleiotropic effects are created (Fig. S11). The power gain brought by MRAID is particularly apparent with increased proportion of instrumental SNPs that display uncorrelated horizontal pleiotropic effects (Fig. 3D *vs* Fig. S10F, Fig. S10E *vs* Fig. S10D). Importantly, the power of MRAID is close to an oracle MR approach that uses the actual set of instrumental SNPs for MR inference, especially when the casual effect size is large, supporting the effectiveness of the automatic instrument selection procedure in MRAID (Fig. S12).

Next, we examined the ability of different MR methods in distinguishing the causal effect direction through reverse causality analysis. In particular, we tested the causal effect of the outcome on the exposure in the alternative simulations where the exposure had casual effect on the outcome but not vice versa. In the presence of horizontal pleiotropy, the SNP instruments obtained for the outcome in the reverse MR analysis would contain two sets of SNPs: a set of exposure SNP instruments that are indirectly associated with the outcome through the exposure and a set of SNPs that are directly associated with the outcome thought their horizontal pleiotropic effects on the outcome. Because the two sets of SNPs displayed heterogeneous effects on the exposure, we would fail to detect a non-zero causal effect of the outcome on the exposure. Therefore, the reverse causality analysis in the presence of horizontal pleiotropy effectively served as analysis on null simulations. Indeed, we found that MRAID provides effective type I error control and calibrated *p*-values in the reverse causality analysis across a range of simulation scenarios (Fig. S13 and S14). In contrast, the type I error control of the other methods is highly dependent on the extent of the horizontal pleiotropy. Specifically, when a small proportion of exposure instrumental SNPs display horizontal pleiotropy on the outcome, the majority of the candidate instrumental SNPs for the outcome in the reverse causality analysis would not display horizontal pleiotropic effects on the exposure. In this case, both CAUSE and Weighted mode remain overly conservative, while IVW-R, MRMix, RAPS and Robust yield slightly inflated *p*-values (Fig. S13A, S13C, S14A, and S14C). By contrast, when a large proportion of instrumental SNPs for the exposure display horizontal pleiotropic effects on the outcome, the majority of the candidate instrumental SNPs for the outcome in the reverse causality analysis would display horizontal pleiotropic effects on the exposure. In this case, MRMix, Robust, IVW-R, Weighted median and RAPS all start

to produce inflated *p*-values (Fig. S13B and S14B) as we have shown in the corresponding null scenarios. The *p*-value inflation of these methods becomes more prominent with smaller horizontal pleiotropic effect sizes, where it becomes increasingly hard to select the second set of SNPs to serve as outcome instruments (Fig. S13D and S14D). The comparison results hold when we force multiple causal SNPs to be in the same LD block (Fig. S15).

MRAID also produces reasonably unbiased causal effect estimates under the null (Fig. S16A) and under various alternatives (Fig. S16B-D), and produces reasonably well estimated proportional estimates (Fig. S17). In addition, MRAID is reasonably robust when the uncorrelated horizontal pleiotropic effects from the instrumental SNPs are either larger or smaller than that from the non-instrumental SNPs (Fig. S18).

Real data applications

We applied MRAID and the other MR methods to analyze 38 lifestyle risk factors and 11 CVD-related traits in the UK Biobank (details in Materials and Methods). Specifically, we divided the UK Biobank data into two separate, equal-sized subsets, representing an exposure GWAS and an outcome GWAS. We performed two sets of analysis. First, we focused on the eight CVD-related traits and examined the causal effect of each trait in the exposure GWAS on the same trait in the outcome GWAS, effectively examining the causal effect of the trait on itself. The true causal effect in such analysis is non-zero and equals exactly one, with the scatter plots displayed in Fig. S19. We found that all methods were able to detect a non-zero causal effect for the trait on itself across all eight CVD-related traits (Fig. 4). However, only MRAID and CAUSE were able to produce 95% confidence intervals that cover the true causal effects for all eight trait pairs, with CAUSE producing confidence intervals that are

2.39-5.69 times larger than MRAID (Fig. 4). For example, in the HDL-HDL analysis, MRAID (estimate = 0.98; 95% CI: 0.96-1.01), CAUSE (0.95; 0.82-1.09) and MRMix (0.96; 0.90-1.02) correctly inferred the causal effect, with MRAID providing the smallest confidence interval (Fig. 4H). In contrast, the confidence intervals from the other five methods did not cover the true causal effect of one. In the LDL-LDL analysis, MRAID (0.97; 0.94-1.01) and CAUSE (0.96; 0.84-1.08) correctly inferred the causal effect, with MRAID providing a smaller confidence interval (Fig. 4F). While the confidence intervals from the other six methods also did not cover the true causal effect of one. The results suggest that both MRAID and CAUSE can produce accurate causal effect estimates and calibrated confidence intervals for the trait on itself analysis, with MRAID being more powerful than CAUSE. We applied similar analysis for investigating the causal effect of each of the four blood lipid traits in the global lipid genomic consortium (GLGC) (*34*) on the same trait in UKBB and found similar results (Fig. S20).

Next, we investigated the causal relationship between 38 lifestyle risk factors and 11 CVD-related traits. The association of lifestyle risk factors on CVD-related traits has been extensively documented (*35, 36*). However, it remains controversial on whether the detected associations are causal as some of the association effects were estimated to have different signs in different studies (*37, 38*). We performed both forward causality analysis examining the causal effects of lifestyle factors on CVD-related traits and reverse causality analysis examining the causal effects of CVD-related traits on lifestyle factors. The distribution of *p*-values for the analyzed trait pairs from different methods are shown in Fig. 5A. Consistent with the simulations, we found that the *p*-values from MRAID (genomic inflation factor, GIF =0.90), and to a lesser extent MRMix (GIF = 0.78), are generally well behaved and slightly conservative across

analyzed trait pairs, more so than the other methods (Fig. 5A). Also consistent with the simulations, we found that the *p*-values from CAUSE are overly conservative (GIF = 0.12), while the *p*-values from RAPS (GIF = 1.96), Weighted mode (GIF = 1.70), IVW-R (GIF = 2.12), Weighted median (GIF=1.80), and Robust (GIF = 2.00) all show appreciable inflation (Fig. 5A). Indeed, only MRAID produces calibrated *p*-values in the permutation analysis where we permuted the outcome trait (Fig. 5B).

Based on a Bonferroni corrected *p*-value threshold (7.75×10^{-5}), MRAID detected eight causal associations (Table S1), all of which have strong biological support. For example, MRAID detected a negative causal effect of smoking on BMI. The negative association between smoking and obesity has been well documented in observational studies (*39, 40*) and MR studies (*41*). Specifically, nicotine intake during smoking decreases resting metabolic rate (*42, 43*) and inhibits lipoprotein lipase activity and other kinase pathways to reduce lipolysis (*40*), all of which lead to a reduction in the net energy storage in adipose tissues and subsequent weight loss(*44*). Nicotine also activates acetylcholine receptors in the hypothalamus and subsequently anorexigenic neurons (*45, 46*), which leads to suppressed appetite and food intake. As another example, MRAID detected an effect of age started smoking in the former smokers on HDL, suggesting a negative effect of smoking behavior on HDL. Smoking behavior in general is well known to be causally associated with HDL (*47*). In particular, smoking can modify the activity of critical enzymes for lipid transport, lower lecithin-cholesterol acyltransferase (LCAT) activity, and alter cholesterol ester transfer protein (CETP) and hepatic lipase activity, all of which can reduce HDL metabolism. In addition, smoking induces oxidative modifications that render HDL dysfunctional and deprive its atheroprotective properties (*48, 49*).

Importantly, MRAID did not mistakenly detect many false causal associations that were detected by the other methods. A well-known example of a potential false causal association is the effect of smoking on blood pressure. A negative association between smoking and blood pressure has been observed in observational studies (*36*). However, multiple subsequent MR studies on large datasets did not support a causal relationship between the two traits (*47, 50*). Indeed, the association between smoking and blood pressure in observational studies is likely confounded by factors that include, but not limited to age, BMI, social class, salt intake, drinking habits, as well as unmeasured confounders (*51*). Consistent with these previous MR studies, MRAID did not detect a significant causal effect from any of the eight smoking related traits on either SBP or DBP. In contrast, almost all other methods falsely detected causal effects of some of the smoking related traits on blood pressure. For example, the causal effect of the number of unsuccessful stop-smoking attempts on SBP is not detected by MRAID ($p = 0.44$), CAUSE ($p = 0.01$) nor Weighted mode ($p = 1.3 \times 10^{-4}$), but falsely identified by IVW-R ($p = 1.4 \times 10^{-6}$), Robust ($p = 4.1 \times 10^{-34}$), RAPS ($p = 5.5 \times 10^{-6}$), MRMix ($p = 2.8 \times 10^{-6}$), and Weighted median ($p = 2.4 \times 10^{-5}$). Similarly, the causal effect of age started smoking in former smokers on SBP is not detected by MRAID ($p = 0.06$) nor CAUSE ($p = 1.3 \times 10^{-3}$), but falsely detected by IVW-R ($p = 7.8 \times 10^{-5}$), Robust ($p = 1.5 \times 10^{-30}$), RAPS ($p = 8.9 \times 10^{-7}$), Weighted median ($p = 1.0 \times 10^{-6}$), Weighted mode ($p = 1.7 \times 10^{-6}$), and MRMix ($p = 4.1 \times 10^{-7}$). As another false example, BMI is unlikely to causally influence the time spent driving, at least not positively. Indeed, MRAID ($p = 0.01$), along with MRMix ($p = 0.12$), CAUSE ($p = 0.02$), Weighted median ($p = 0.04$), and Weighted mode ($p = 0.04$), did not detect any causal effect of BMI on time spent driving. However, both

IVW-R ($p = 2.5 \times 10^{-6}$) and RAPS ($p = 3.1 \times 10^{-5}$) detected a false positive effect of BMI on time spent driving.

Finally, we note that an important feature of MRAID is its ability to effectively decompose the SNP effects on the outcome into three distinct paths: one directly acts from SNPs to the outcome, one mediated through the exposure, and the other acts through a hidden confounding factor that influences both exposure and outcome. Consequently, MRAID can be used to estimate the proportion of SNPs in different categories, including the proportion of SNPs that are associated with the exposure among the genome-wide significant ones (π_β), the proportion of SNPs that exhibit correlated horizontal pleiotropy (π_c), the proportion of SNPs that exhibit uncorrelated horizontal pleiotropy among the selected instruments (π_1), and the proportion of SNPs that exhibit uncorrelated horizontal pleiotropy among the remaining candidate instruments (π_0). In the real data applications, we estimated the mean of π_β, π_c, π_1 and π_0 across the 645 analyzed trait pairs to be 14.6%, 6.4%, 16.4%, and 5%, respectively (Fig. S21). Note that these percentages were calculated based on the number of candidate instruments; thus, a value of 14.6% corresponds to an average of 107 variants per trait. Additional analysis illustrated that the *p*-values of MRAID remains consistent with each other regardless of the prior distribution of π_β (Fig. S22). In addition, we estimated their means in the eight significant trait pairs to be 6.2%, 5.7%, 11.4%, and 0.1%, respectively. The proportion of SNPs displaying correlated pleiotropy is also highly correlated with the proportion of SNPs displaying uncorrelated pleiotropy, with the latter generally being larger than the former (Fig. S23). These proportion estimates support the wide-spread horizontal pleiotropy previously identified in complex trait analysis (*26*) and provide detailed quantifications on the extent to which the two types of horizontal pleiotropy influence MR analysis.

Discussion

We have presented MRAID, a two-sample MR method that can automatically select suitable instruments from a candidate set of correlated SNPs and that can control for both correlated and uncorrelated horizontal pleiotropy in a likelihood-based inference framework. Overall, by automatically selecting instrumental SNPs and performing inference under a likelihood-based framework, MRAID yields calibrated *p*-values across a wide range of scenarios and improves power of MR analysis over existing approaches. We have illustrated the benefits of MRAID through simulations and applications to complex trait analysis.

We have primarily focused on modeling quantitative traits with MRAID in the present study. For binary exposures and outcomes, one could treat them as continuous variables and directly applied MRAID for MR analysis. Treating binary exposures and outcomes as continuous variables can be justified by recognizing the linear model as a first-order Taylor approximation to a generalized linear model such as the logistic regression (*52*). However, such approximation is accurate only when the SNP effects on the exposure and outcome are relatively small. While similar approaches have been applied in many previous MR studies (*53-55*), we caution that the interpretation of the causal effect estimate can be challenging when the linear models are used to fit binary exposures and outcomes, especially when a two-stage inference procedure is used for MR analysis (*56, 57*). For example, when a binary exposure is a dichotomization of a continuous risk factor, the causal effect estimation through modeling the binary exposure without the underlying continuous risk factor may require additional modeling assumptions, even when the main MR assumptions are satisfied. In addition, modeling binary exposure without the underlying continuous risk factor can lead to

violation of the exclusion restriction assumption, as the instruments can influence the outcome via the continuous risk factor even if the binary exposure does not change.

Therefore, extending MRAID to explicitly model data types beyond quantitative traits is important to ensure its wide applicability. Because MRAID builds upon a data generative model and performs inference on the SNP-exposure model and the SNP-outcome model jointly through a maximum likelihood-based framework, it can be naturally extended towards modeling binary outcome through a liability threshold model (*58*), and binary and other types of exposure or outcome data through a generalized linear model framework. To the best of our knowledge, the only likelihood-based MR method that accommodates both binary risk factors and outcome is IV-MVB (*59*). IV-MVB, however, requires individual-level data, applies to the one-sample analysis setting, and cannot easily handle multiple instruments in a computationally efficient fashion especially for those that are correlated. Therefore, exploring the benefits of MRAID extensions towards modeling generalized data types while keeping computation in check will be an important direction for future research.

Another important future direction is to extend MRAID to incorporate SNP functional annotations. Specifically, we could model the probability of a SNP exhibiting instrumental effects as a function of a given set of SNP annotations through a logistic function, similar to what was used in many SNP fine-mapping methods (*60-62*). In addition, we could also model the probability of a SNP exhibiting horizontal pleiotropic effects as a function of its functional annotations through a logistic function. Because the biological function and importance of a SNP can be predicted in certain degree by its functional annotations, incorporating SNP functional annotations can potentially improve the performance of MRAID. Certainly, incorporating functional annotations would inevitably increase the number of parameters, making it challenging

to carry out powerful MR inference given the small instrumental effects. Therefore, it would be important to incorporating informative functional annotations while mitigating the impact by the increased number of parameters to ensure optimal MR inference power.

MRAID is not without limitations. First, while MRAID performs automated selection on SNP instruments, such selection builds upon a sparsity inducing modeling assumption specified on the SNP effect sizes. The sparse modeling assumption contains multiple hyper-parameters that rely on a sampling-based algorithm for inference. Accurate and robust inference of the hyper-parameters will likely require at least a moderate number of candidate instruments. While the significance of the trait pairs evaluated by MRAID in our real data application does not appear to be dependent on the number of candidate instruments selected for the trait pair (Fig. S24), we caution that MRAID may incur low power when the instrumental effect size is small and the number of candidate instruments is low, which can happen in GWAS with small sample sizes and for exposure traits with a non-polygenic architecture. Second, MRAID primarily follows the approach of CAUSE to model correlated horizontal pleiotropy by introducing a single latent variable to serve as the confounder for both the exposure and the outcome. Because of its limitation in modeling only a single unobserved confounding factor, MRAID may not be fully effective in settings where multiple or other types of shared genetic components are present between the exposure and the outcome. Finally, the summary statistics version of MRAID requires as input two LD matrices, one from the exposure GWAS and another from the outcome GWAS. In the present study, we have estimated both LD matrices using individual level data. In the absence of individual level data, both LD matrices may be estimated from a reference panel with the same genetic ancestry (e.g. from the 1,000 Genomes Project). However,

care needs to be taken when the exposure and outcome GWASs are carried out on two populations with distinct genetic ancestries, or when the genetic ancestry of the reference panel does not match that of the GWAS (Fig. S25).

Materials and Methods

MRAID for individual level data

We provide an overview of our method here, with its inference and technical details provided in the Supplementary Text and an illustrative diagram displayed in Fig. 1. Our goal is to estimate and test the causal effect of an exposure variable on an outcome variable in the two-sample MR setting where the exposure and outcome variables are measured in two separate GWASs with no sample overlap. We refer to the two separate GWASs as the exposure GWAS and the outcome GWAS, respectively. To set up the notations, we denote $\mathbf{x}$ as an n_1-vector of the exposure variable measured on n_1 individuals in the exposure GWAS. We denote $\mathbf{y}$ as an n_2-vector of the outcome variable measured on n_2 individuals in the outcome GWAS. We scale both $\mathbf{x}$ and $\mathbf{y}$ to have zero mean and unit standard deviation. In the exposure GWAS, we perform an initial screening to select SNPs that are associated with the exposure variable with a marginal p-value below the genome-wide significance threshold of 5×10^{-8}. These SNPs are likely in LD with each other and are selected to serve as the initial set of candidate instruments. We denote $\mathbf{Z}_\mathrm{x}$ as the resulting n_1 by p genotype matrix for the p selected candidate instrumental SNPs in the exposure GWAS. We also denote $\mathbf{Z}_\mathrm{y}$ as an n_2 by p genotype matrix for the same p candidate instrumental SNPs in the outcome GWAS. We scale each column of the two genotype matrices to have mean zero and standard deviation of one. We model the relationship among the exposure, outcome and genotypes through the following three linear regressions:

$$\mathbf{x} = \mathbf{Z}_\mathrm{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathrm{x}, \qquad (1)$$

$$\tilde{\mathbf{x}} = \mathbf{Z}_\mathrm{y}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_\mathrm{x}, \qquad (2)$$

$$\mathbf{y} = \tilde{\mathbf{x}}\alpha + \mathbf{Z}_\mathrm{y}\boldsymbol{\eta}_0 + \mathbf{Z}_\mathrm{y}\boldsymbol{\eta}_1 + \boldsymbol{\varepsilon}_\mathrm{y}. \qquad (3)$$

Above, equation (1) describes the relationship between the genotypes $\mathbf{Z}_x$ and the exposure variable $\mathbf{x}$ in the exposure GWAS; equation (2) describes the relationship between the genotypes $\mathbf{Z}_y$ and the unobserved exposure $\tilde{\mathbf{x}}$ in the outcome GWAS; equation (3) describes the relationship among the genotypes $\mathbf{Z}_y$, the outcome $\mathbf{y}$, and the unobserved exposure $\tilde{\mathbf{x}}$ in the outcome GWAS; $\boldsymbol{\beta}$ is a p-vector of SNP effects on the exposure; both $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ are p-vectors of horizontal pleiotropy effects on the outcome; α is a scalar that represents the causal effect of the exposure on the outcome; $\boldsymbol{\varepsilon}_x$ is an n_1-vector of residual error with each element independently and identically distributed from a normal distribution $N(0, \sigma_x^2)$; $\tilde{\boldsymbol{\varepsilon}}_x$ is an n_2-vector of residual error with each element distributed from the same normal distribution $N(0, \sigma_x^2)$; and $\boldsymbol{\varepsilon}_y$ is an n_2-vector of residual error with each element distributed from a normal distribution $N(0, \sigma_y^2)$. We note that while the above three equations are specified based on two separate GWASs, they are connected to each other by the common parameter $\boldsymbol{\beta}$. We carefully consider the modeling assumptions on the SNP effects on the exposure variable $\boldsymbol{\beta}$ as well as the horizontal pleiotropic effects $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ as follows.

The p SNPs included in the above model represent an initial set of candidate instruments. While all the candidate instruments are marginally associated with the exposure, the majority of them are unlikely the causal SNPs for the exposure variable. Instead, most candidate instruments likely represent tagging SNPs that are associated with the exposure variable due to LD with the truly causal ones underlying the exposure. Therefore, it would be beneficial to perform additional selections on the candidate instruments to identify SNPs that are causal for the exposure and treat them as the instruments in order to maximize the power of MR analysis. To do so, we borrow ideas from fine-mapping approaches developed in the research field of GWAS and specify a sparsity inducing modeling assumption on the SNP effects on the exposure ($\boldsymbol{\beta}$) to

perform automated instrument selection. In particular, we assume that $\beta_j \sim \pi_\beta N(0, \sigma_\beta^2) + (1-\pi_\beta)\delta_0$, where δ_0 is the Dirac function that represents a point mass at zero. That is, with probability $1-\pi_\beta$, the j-th SNP has zero effect on the exposure; while with probability π_β, the j-th SNP has a non-zero effect on the exposure and its effect size follows a normal distribution with mean zero and variance σ_β^2, where the variance parameter σ_β^2 determines the magnitude of the effect sizes. The sparse assumption on $\boldsymbol{\beta}$ allows us to select SNPs with non-zero effects on the exposure to serve as the instruments in the MR model.

In addition, the p SNPs included in the above model can also exhibit horizontal pleiotropic effects and influence the outcome variable through pathways other than the exposure. To control for the potential horizontal pleiotropic effects and improve causal effect inference, we introduce two sets of parameters, $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, to model horizontal pleiotropic effects. The two sets of parameters are placed separately for the two SNP groups – the group of selected instrumental SNPs and the group of unselected non-instrumental SNPs – that are categorized by the sparse modeling assumption on $\boldsymbol{\beta}$. In particular, $\boldsymbol{\eta}_1$ represents the horizontal pleiotropic effects exhibited by the selected SNPs instruments with non-zero $\boldsymbol{\beta}$ while $\boldsymbol{\eta}_0$ represents the horizontal pleiotropic effects exhibited by the unselected non-instrumental SNPs with zero $\boldsymbol{\beta}$. Controlling for $\boldsymbol{\eta}_1$ can help mitigate the bias in causal effect estimation induced by horizontal pleiotropic effects from the instrumental SNPs. While controlling for $\boldsymbol{\eta}_0$ can reduce residual error variance in equation (3) and thus help improve the statistical efficiency of causal effect estimation.

To effectively control for the horizontal pleiotropic effects exhibited from both SNP groups, we specify separate modeling assumptions on $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$. Specifically,

for the selected SNP instruments, we assume that they can exhibit horizontal pleiotropic effects in two different ways: they can affect the outcome through a common confounder that is associated with both the exposure and outcome, and they can affect the outcome through paths independent of the exposure. For the first type of horizontal pleiotropy, we assume that each selected SNP instrument has a probability of π_c to induce pleiotropy through the confounder. Following(*8*), we assume that the confounder effect on the outcome is ρ times its effect on the exposure. Consequently, the effect of the selected SNP instrument acted through the confounder on the outcome becomes $\rho\beta_j$, if the SNP effect on the exposure is β_j. Thus, our assumption on η_{1j}^c, which represents the first type of horizontal pleiotropy as a part of $\boldsymbol{\eta}_1$ for the *j*-th SNP, is $\eta_{1j}^c|\beta_j \neq 0 \sim \pi_c I(\eta_{1j} = \rho\beta_j) + (1-\pi_c)\delta_0$, where $I(\cdot)$ is an indicator function that sets the horizontal pleiotropic effect to be $\rho\beta_j$. For the second type of horizontal pleiotropy, we assume that each selected SNP instrument has a probability of π_1 to exhibit a horizontal pleiotropic effect on the outcome directly, bypassing the exposure. We use η_{1j}^u to represent the second type of horizontal pleiotropy as a part of $\boldsymbol{\eta}_1$ for the *j*-th SNP. Our assumption on η_{1j}^u is thus $\eta_{1j}^u|\beta_j \neq 0 \sim \pi_1 N\left(0,\ \sigma_\eta^2\right) + (1-\pi_1)\delta_0$, where the variance σ_η^2 determines the strength of the horizontal pleiotropic effect. Note that the first type of horizontal pleiotropic effects are correlated with the instrumental effects on the exposure due to the confounder, while the second type of horizontal pleiotropic effects are uncorrelated with the instrumental effects on the exposure. The total horizontal pleiotropy is the summation of the two, with $\eta_{1j} = \eta_{1j}^c + \eta_{1j}^u$. Certainly, because $\boldsymbol{\eta}_1$ are the horizontal pleiotropic effects for the selected SNP instruments, we have $\eta_{1j} = 0$ if $\beta_j = 0$. For the unselected non-instrumental SNPs with a zero β_j, we assume that $\eta_{0j}|\beta_j = 0 \sim \pi_0 N\left(0,\ \sigma_\eta^2\right) + (1-\pi_0)\delta_0$. That is,

with probability π_0, the non-instrumental SNPs display horizontal pleiotropic effects characterized by the same variance parameter σ_η^2. We use the same variance parameter σ_η^2 for modeling the uncorrelated horizontal pleiotropic effects from both instrumental and non-instrumental SNPs because we often do not have enough number of SNPs to estimate two separate parameters accurately. Since $\boldsymbol{\eta}_0$ are the horizontal pleiotropic effects for the non-instruments, we also have $\eta_{0j} = 0$ if $\beta_j \neq 0$.

The above parameterization of the horizontal pleiotropic effects is based on the selection of SNP instruments. An equivalent and alternative parametrization of the horizontal pleiotropic effects is to partition them into a correlated horizontal pleiotropic component $\boldsymbol{\eta}_\mathrm{c}$ and an uncorrelated horizontal pleiotropic component $\boldsymbol{\eta}_\mathrm{u}$. Specifically, the correlated horizontal pleiotropy occurs only for the selected SNP instruments with $\eta_{cj}|\beta_j \neq 0 \sim \pi_c I(\eta_{1j} = \rho\beta_j) + (1-\pi_c)\delta_0$ and $\eta_{cj} = 0$ if $\beta_j = 0$. The uncorrelated horizontal pleiotropy, on the other hand, occurs for both instrumental and non-instrumental SNPs with $\eta_{uj}|\beta_j \neq 0 \sim \pi_1 N(0,\ \sigma_\eta^2) + (1-\pi_1)\delta_0$ and $\eta_{uj}|\beta_j = 0 \sim \pi_0 N(0,\ \sigma_\eta^2) + (1-\pi_0)\delta_0$. In other words, $\eta_{cj} = \eta_{1j}^c$ and $\eta_{uj} = \eta_{1j}^u + \eta_{0j}$.

Overall, the SNP effects on the outcome in our model are exhibited through three different paths: via the exposure on outcome causal effect α; via the correlated horizontal pleiotropic effects mediated by an unobserved confounder; and via the uncorrelated horizontal pleiotropic effects. SNPs in the model can exhibit none, one, or multiple types of these effects. Note that the SNP effects on the outcome through the causal effect and through the correlated horizontal pleiotropy are not distinguishable from each other unless we make further modeling assumptions. Here, following(*8*), we assume π_c to be small. Thus, among the selected SNP instruments with non-zero effects on the exposure, only a fraction of them exhibit correlated horizontal pleiotropic effects on the outcome.

Our key parameter of interest is the causal effect α. The causal interpretation of α in a standard MR model requires the selected SNP instruments to satisfy three conditions: (i) instruments are associated with the exposure (relevance condition); (ii) instruments are not associated with any other confounder that may be associated with both exposure and outcome (independence condition); (iii) instruments only influence the outcome through the path of exposure (exclusion restriction condition). Our modeling assumption on $\boldsymbol{\beta}$ allows us to select SNPs to satisfy the relevance condition. Our modeling assumptions on $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ allow us to explicitly model the violation of the independence and exclusion restriction conditions. Therefore, our model effectively replaces the general conditions (ii) and (iii) with specific modeling assumptions on $\boldsymbol{\beta}$, $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$. In addition, through explicit modeling of the correlation between the instrument-exposure effects and instrument-outcome effects through ρ, our model no longer requires the InSIDE assumption, which is sometimes referred to as the weak exclusion restriction condition(*3*). Consequently, the causal effect interpretation of α in our model only depends on the explicit assumptions made in the model.

We are interested in estimating the causal effect α and testing the null hypothesis $H_0: \alpha = 0$. Performing inference on α, however, is computationally challenging, as the likelihood defined based on the above modeling assumptions is in a complicated form and involves integrations that cannot be obtained analytically. Here, we develop an approximate inference algorithm under the maximum likelihood framework to perform numerical integration of the likelihood and obtain an approximate *p*-value for testing α. Our algorithm is based on the observation that the likelihood function of α can be expressed as a ratio between the posterior and the prior. Because the posterior of α is asymptotically normally distributed(*63, 64*), we can use Gibbs sampling to obtain

posterior samples of α and use the sample mean and sample standard deviation to summarize this posterior distribution. In addition, we can also specify a normal prior on α and obtain the prior mean and standard deviation. Because the likelihood of α is expressed as the posterior divided by the prior and is itself asymptotically normally distributed(*63, 64*), we can rely on the method of moments to obtain the approximate maximum likelihood estimate $\hat{\alpha}$ and its standard error $se(\hat{\alpha})$ based on the mean and standard deviation from both the posterior and the prior. Afterwards, we can construct an approximate Wald test statistic and obtain a *p*-value for hypothesis testing. Details of the algorithm is provided in the Supplementary Text. As a unique feature of our algorithm, we introduce a set of binary indicator variables to effectively explore the joint parameter space to alleviate the issue from the inter-dependence among the parameters $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_0$ and $\boldsymbol{\beta}$ (details in Supplementary Text). Note that, while our algorithm relies on Gibbs sampling, we do not perform a Bayesian analysis; rather, we treat the Gibbs sampling as a convenient and accurate numerical approximation tool to obtain the marginal likelihood of α, which is otherwise infeasible or inaccurate to obtain under various frequentist approaches.

We refer to our model and algorithm together as the two-sample Mendelian Randomization with Automated Instrument Determination (MRAID). The automated instrument determination part highlights the desirable feature of our model in automatically selecting instrumental variables from a set of candidate ones that are in potentially high LD with each other. Compared with existing two-sample MR approaches, MRAID relies on a likelihood inference framework, is capable of modeling correlated instruments, performs automated instrument selection, controls for both correlated and uncorrelated horizontal pleiotropy, and is computationally scalable.

MRAID is implemented in an R package, freely available at www.xzlab.org/software.html.

MRAID for summary statistics

While we have described MRAID using individual-level data, MRAID can be extended to make use of only summary statistics. Details for the summary statistics version of MRAID are provided in the Supplementary Text. Briefly, the summary statistics version of MRAID requires two types of input: the SNP marginal effect size estimates on the exposure and outcome; and the SNP correlation matrices in the exposure and outcome GWASs. Both input types are obtained based on standardized genotype data where the genotypes for each SNP have been standardized to have zero mean and unit standard deviation. Here, we denote the p-vector of the SNP marginal effect size estimates on the exposure as $\widehat{\boldsymbol{\beta}}_{\mathrm{x}}$ and the corresponding vector of marginal effect size estimates on the outcome as $\widehat{\boldsymbol{\beta}}_{\mathrm{y}}$. We denote the p by p SNP correlation matrix in the exposure GWAS as Σ_1 and the corresponding matrix in the outcome GWAS as Σ_2. Both Σ_1 and Σ_2 are positive semi-definite and can be estimated from the same LD reference panel (e.g. individuals with the same ancestry in the 1,000 Genomes Project). The MRAID model for summary statistics can be constructed based on the following two equations

$$\widehat{\boldsymbol{\beta}}_{\mathrm{x}} = \Sigma_1\boldsymbol{\beta} + \mathbf{e}_{\mathrm{x}}, \qquad (4)$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{y}} = \alpha\Sigma_2\boldsymbol{\beta} + \Sigma_2\boldsymbol{\eta}_0 + \Sigma_2\boldsymbol{\eta}_1 + \mathbf{e}_{\mathrm{y}}, \quad (5)$$

where $\mathbf{e}_{\mathbf{x}}$ is a p-vector of residual error that follows a multivariate normal distribution $N(0, \Sigma_1\sigma_x^2/(n_1-1))$; and $\mathbf{e}_{\mathrm{y}}$ is a p-vector of residual error that follows another a multivariate normal distribution $N\big(0, \Sigma_2\sigma_y^2/(n_2-1)\big)$. A similar approximate

inference algorithm under the maximum likelihood framework is developed for the summary version of MRAID.

Simulations

We performed realistic simulations to evaluate the performance of MRAID and compared it with seven existing MR methods. For simulations, we randomly selected 60,000 individuals from UK Biobank (*32*). We split these individuals randomly into two equal-sized sets: one set with 30,000 individuals to serve as the exposure GWAS and another set with the remaining 30,000 individuals to serve as the outcome GWAS. For these individuals, we obtained their genotypes from 649,695 SNPs on chromosome 1 that are overlapped with the GERA study we used before (*14*), standardized each SNP to have mean zero and unit standard deviation, and used the standardized genotypes to simulate the exposure and outcome. Specifically, in the exposure GWAS, we randomly selected K SNPs (K =100 or 1,000) to have non-zero effects on the exposure. We denoted the genotype matrix of the K SNPs as $\tilde{\boldsymbol{Z}}_x$. We simulated the K SNP effect sizes on the exposure ($\boldsymbol{\beta}$) from a normal distribution $N(0, PVE_{\tilde{Z}_x}/K)$, where the scalar $PVE_{\tilde{Z}_x}$ represents the proportion of variance in the exposure variable explained by these genetic effects. We summed the genetic effects across all K SNPs as $\tilde{\boldsymbol{Z}}_x\boldsymbol{\beta}$. In addition, we simulated the residual errors $\boldsymbol{\varepsilon}_{\mathrm{x}}$ from a normal distribution $N(0, 1 - PVE_{\tilde{Z}_x})$. We then summed the genetic effects and the residual errors to yield the simulated exposure variable $\mathbf{x}$. In the outcome GWAS, we obtained the genotypes for the same K SNPs as $\tilde{\boldsymbol{Z}}_y$ and used the same $\boldsymbol{\beta}$ from the exposure GWAS to compute

the genetic component underlying the outcome as $\widetilde{\mathbf{Z}}_y\boldsymbol{\beta}$. We set the causal effect α to be $\alpha = \sqrt{PVE_\alpha / PVE_{\widetilde{Z}_x}}$, so that the proportion of variance in the outcome variable explained by the causal effect term ($\widetilde{\mathbf{Z}}_y\boldsymbol{\beta}\alpha$) is PVE_α. We randomly obtained $\pi_c K$ SNPs (rounded to an integer) from the K SNPs to exhibit correlated pleiotropy. We simulated the correlated pleiotropic effect sizes to be $\rho\boldsymbol{\beta}$ and set ρ so that the proportion of variance in the outcome variable explained by correlated pleiotropy is PVE_c. In addition, we randomly obtained $\pi_1 K$ SNPs (again rounded to an integer) from the K SNPs and randomly obtained $100 - \pi_1 K$ SNPs from the remaining non-causal SNPs to exhibit uncorrelated pleiotropy, so that a total of 100 SNPs displayed uncorrelated pleiotropy. We simulated the uncorrelated horizontal pleiotropic effects for these 100 SNPs from a normal distribution and scaled them so that the proportion of phenotypic variance in the outcome explained by uncorrelated pleiotropy is PVE_u. We simulated the residual errors $\boldsymbol{\varepsilon}_\mathrm{y}$ from a normal distribution $N(0, 1 - PVE_\alpha - PVE_\mathrm{c} - PVE_u)$. We summed the causal effect term, correlated and uncorrelated horizontal pleiotropic effects, and the residual errors to yield the simulated outcome $\mathbf{y}$.

We treated the causal SNPs as unknown and followed standard MR procedure to select SNPs to serve as the instrumental variables. To do so, we used the linear regression model implemented in GEMMA (*65*) to perform association analysis in the exposure GWAS and selected SNPs with a p-value below 5×10^{-8} as the candidate instrumental variables for analysis. For the selected SNPs, we obtained their effect size estimates, standard errors, and Z scores to serve as the summary statistics input. We also denoted the standardized genotype matrices for the selected SNPs in the exposure

and outcome GWASs as $\boldsymbol{Z}_x$ and $\boldsymbol{Z}_y$, respectively. Based on the genotype matrices, we obtained the SNP correlation matrices as $\boldsymbol{\Sigma}_1 = \frac{\boldsymbol{Z}_x^T \boldsymbol{Z}_x}{n_1 - 1}$ and $\boldsymbol{\Sigma}_2 = \frac{\boldsymbol{Z}_y^T \boldsymbol{Z}_y}{n_2 - 1}$ to serve as input for MR model fitting.

In the simulations, we first examined a baseline simulation setting where we set $PVE_{\tilde{Z}_x} = 10\%$, $PVE_\alpha = 0$, $K = 100$, $\pi_c = 0$, $PVE_u = 0$, $PVE_c = 0$. On top of the baseline setting, we varied one parameter at a time to examine the influence of various parameters on method performance. For $PVE_{\tilde{Z}_x}$, we set it to be either 5% or 10%. For $\boldsymbol{\beta}$, in addition to simulating it from a normal distribution, we also simulated them from the Bayesian sparse linear mixed model (BSLMM) distribution(*52*). Specifically, we randomly selected either 1% or 10% of the *K* SNPs to have large effects and these large-effect SNPs explain 20% of $PVE_{\tilde{Z}_x}$. We set the remaining SNPs to have small effects to explain the remaining $PVE_{\tilde{Z}_x}$. For *K*, we set it to be either 100 or 1,000. For PVE_α, we set it to be zero in the null simulations and examined different values in the alternative simulations. In the alternative simulations, we set PVE_α to be 0.05%, 0.15% or 0.25% when *K*=100 and set it to be 0.5%, 1.5% and 2.5% when *K*=1,000 to ensure sufficient power. For the uncorrelated horizontal pleiotropic effects, we set PVE_u to be either 0, 2.5% or 5%. Under the null ($PVE_\alpha = 0$) in the absence of uncorrelated horizontal pleiotropy ($PVE_u = 0$), we set *K* to be 100 or 1,000. In the presence of uncorrelated horizontal pleiotropy, we set *K* to be 100 and set π_1 to be either 0, 10%, 20%, or 30%. We also simulated the correlated pleiotropy effects and set π_c to be either 5% or 10%, with ρ being $\sqrt{0.02}$ or $\sqrt{0.05}$ following the previous literature(*8*).

Note that MRAID used the same variance parameter for modeling the uncorrelated horizontal pleiotropic effects of both instrumental and non-instrumental SNPs, as the number of SNPs included in the model may not be sufficiently large for accurate inference of the two separate parameters. We conducted additional simulations to evaluate the robustness of MRAID against the violation of this assumption in the presence of horizontal pleiotropic effects ($PVE_u = 5\%$), with the proportion of instrumental SNPs having uncorrelated horizontal pleiotropy to be 20% and 30%, respectively. We set the variance parameter for generating the uncorrelated horizontal pleiotropic effects from the non-instrumental SNPs to be either 3 times or 1/3 of that from the instrumental SNPs.

For the above simulations, we examined the number of causal SNPs in each LD block. Specifically, we used LDetect (*66*) and followed its default settings to divide chromosome 1 into 133 independent LD blocks. We then examined the location of the randomly selected 100 causal SNPs in each simulation. We found that the mean proportion of LD blocks that contain at least one causal SNP across the 1,000 simulation replicates is 51%. In the LD blocks that contain at least one causal SNPs, we found the number of causal SNPs ranges from 1 to 7. In particular, an average of 64% of LD blocks with at least one causal SNP contain exactly one causal SNP and an average of 36% of LD blocks with at least one causal SNP contain more than one causal SNP. Therefore, it appears that multiple causal SNPs are presented in the same LD block in our simulations. For the SNPs that in the same LD block, the mean of the absolute value of pair-wise r^2 is 0.76.

For null simulations, we performed 1,000 simulation replicates in each scenario to examine type I error control. For power evaluation, we performed 100 alternative

simulations along with 900 null simulations, with which we computed power based on an FDR of 0.05. We then repeated such analysis five times and report the average power across these replicates. Note that we computed power based on FDR instead of a nominal *p*-value threshold to allow for fair comparison across methods, as the same *p*-value from different methods may correspond to different type I errors.

Besides the above comprehensive simulations, we also conducted a set of simple simulations to help build the intuition of the benefits of MRAID brought by modeling multiple correlated SNPs. Specifically, we first used LDetect with the default settings to divide chromosome 1 into 133 approximately independent LD blocks (*66*). We randomly selected one LD block and randomly selected 10 SNPs in the LD block for simulations. Among the 10 SNPs, we randomly selected one SNP to be causal and applied different methods to perform MR analysis either with the causal SNP (i.e. 10 SNPs total) or without the causal SNPs (i.e. 9 SNPs total). We also randomly selected two SNPs to be causal and performed MR analysis either with the causal SNPs (i.e. 10 SNPs total) or without the two causal SNPs (i.e. 8 SNPs total). In addition, we randomly selected two neighborhood LD blocks on chromosome 1 and randomly selected 10 SNPs from each block for another set of simulations. We then carried out similar simulations, with one SNP in each LD block randomly selected to be causal and with the MR analysis performed either with or without the two causal SNPs. We examined both the type I error control as well as power in the absence of both correlated and uncorrelated horizontal pleiotropic effects. We set $n_1 = 30000$, $n_2 = 30000$, $PVE_{\mathrm{zx}} = 0.25\%$, $PVE_{\alpha} = 0$ (null simulation) or $PVE_{\alpha} = 1\%$ (power simulation) to examine the performance of MRAID and the MR method that uses only the lead variant.

Real Data Applications

We applied MRAID and other MR methods to detect causal associations between 38 lifestyle risk factors and 11 CVD-related traits in the UK Biobank. The UK Biobank data consists of 487,298 individuals and 92,693,895 imputed SNPs (*32*). We followed the same sample QC procedure in Neale lab (https://github.com/Nealelab/UK_Biobank_GWAS/tree/master/imputed-v2-gwas) to retain a total of 337,129 individuals of European ancestry for analysis. We also filtered out SNPs with an HWE p-value $< 10^{-7}$, a genotype call rate < 95%, or an MAF < 0.001 to obtain a total of 13,876,958 SNPs for analysis. For the retained individuals, we obtained all lifestyle-related quantitative traits and CVD-related traits, removed those traits with a sample size less than 10,000, and focused on the remaining set of 38 lifestyle traits and 11 CVD-related traits for analysis. The 38 lifestyle traits include 8 physical activity traits, 12 alcohol intake traits, 10 diet traits (e.g. coffee and fruits) and 8 smoking related traits. The 11 CVD-related traits include four pulse wave traits, two blood pressure traits (SBP and DBP), four lipid traits (LDL, HDL, TC, and TG) and BMI. Details of these traits are listed in Table S2. Many of these lifestyle risk factors have been found to be associated with CVD-related traits in observational studies (*67-69*), though it remains unclear whether these associations represent causal relationship. For each trait in turn, we removed the effects of sex and top ten genotype principal components (PCs) to obtain the trait residuals, standardized the residuals to have a mean of zero and a standard deviation of one, and used these scaled residuals for MR analysis.

To mimic the two-sample MR design, we randomly split the 337,129 individuals into two non-overlap sets: an exposure GWAS set with 168,564 individuals and an outcome GWAS set with 168,565 individuals. The random data split strategy ensures sample homogeneity within each study and independence between studies, and was extensively used in the previous MR literature (*6, 70-72*). We examined the 38 lifestyle

traits in the exposure GWAS and examined the 11 CVD-related traits in the outcome GWAS. In both GWASs, we obtained summary statistics for each trait through linear regression implemented in GEMMA. When lifestyle traits in the exposure GWAS were treated as the exposure, we selected SNPs with a *p*-value below 5×10^{-8} to serve as the candidate instruments for each exposure trait. Because almost all MR methods require at least two instrumental SNPs and some methods can become unstable when the number of instrumental SNPs is too large, we removed exposure traits for which the number of candidate instruments is either below two or above 10,000. This way, we removed three traits with less than two candidate instruments and four traits with more than 10,000 candidate instruments. We paired the remaining 31 exposure lifestyle traits with 11 outcome CVD-related traits into 341 trait pairs. The mean number of significantly associated SNPs among the 31 traits is 286. When CVD-related traits in the outcome GWAS were treated as the exposure, we removed three traits with less than two candidate instruments, including Pulse wave reflection index with no candidate SNP, Pulse wave peak to peak time with one candidate SNP and Pulse wave Arterial Stiffness index with one candidate SNP. We still found that the remaining eight traits have a total of more than 10,000 candidate instruments. Therefore, for these remaining traits, we used a more stringent *p*-value threshold of 1×10^{-15} to select SNP instruments and analyzed the resulting 304 trait pairs. The mean number of associated SNPs among the eight CVD-related traits is 2,318. In total, we analyzed 645 trait pairs.

Compared Methods

We compared the performance of MRAID with seven existing methods that include the followings. (1) IVW-R, which is the random effects version of IVW. It obtains the

causal effect estimate through weighting and combining the effect estimates from individual instruments. It relies on random effects to account for pleiotropy and effect estimate heterogeneity across instruments(*73*). (2) Weighted mode, which is a mode version of IVW. It obtains the causal effect estimate as the mode, instead of the mean, of the effect estimates obtained from individual instruments(*74*). (3) Robust, which is a robust version of IVW. It uses the MM-estimation procedure consisting of an initial S-estimate followed by an M-estimate(*75*) that is further combined with Tukey's bi-weight loss function(*76*). (4) Weighted median, which can provide the consistent estimator even when up to 50% of the information comes from invalid instrumental variables. We fitted methods (1)-(4) using R package 'MendelianRandomization' with default settings. (5) RAPS, which is the MR Adjusted Profile Score method. It incorporates random effects and robust loss functions into the profile score to account for systematic and idiosyncratic pleiotropy (*6*). We fitted RAPS using R package 'mr.raps'; (6) MRMix, which relies on a mixture model to account for horizontal pleiotropic effects and their correlation with instrumental effect sizes (*7*). We fitted MRMix using R package 'MRMix'. (7) CAUSE, which identifies instrumental effect size patterns that are consistent with causal effects while accounting for correlated pleiotropy (*8*). We fitted CAUSE using R package 'cause'. We compared MRAID with the above seven methods because CAUSE is one of the most recently developed methods; IVW-R, Robust and RAPS all have been shown to have superior performance when the InSIDE assumption is satisfied; while MRMix and Weighted mode perform well even the InSIDE assumption is violated(*8, 33*). In both simulations and real data applications, we first obtained SNPs that achieve genome-wide significance level ($p < 5 \times 10^{-8}$) to serve as a candidate set of instrumental SNPs. We directly use this candidate set of instrumental SNPs for MRAID. Because all other MR methods require

independent instrumental SNPs, we performed LD clumping on the candidate set of instrumental SNPs to select independent ones for analysis. LD clumping is performed using PLINK, where we set the LD r^2 parameter to be 0.001. CAUSE also requires estimating some nuisance parameters in the model by using a random set of SNPs across the genome, and we did so by randomly selecting 100,000 SNPs following(*8*). Finally, we explored an oracle approach in the power simulations where we knew the actual set of instrumental SNPs that affect the exposure variable. In the oracle approach, we obtained the actual set of instrumental SNPs, selected among them the independent ones via pruning, and used the selected set of SNPs to serve as instruments using the IVW-R method. The compared methods and their corresponding software are listed in Table S3.

References

1. S. Burgess, D. S. Small, S. G. Thompson, A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res* **26**, 2333-2355 (2017).

2. S. Burgess, A. Butterworth, S. G. Thompson, Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology* **37**, 658-665 (2013).

3. J. Bowden, G. D. Smith, S. Burgess, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512-525 (2015).

4. J. Bowden, G. Davey Smith, P. C. Haycock, S. Burgess, Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic epidemiology* **40**, 304-314 (2016).

5. J. Zhao, J. Ming, X. Hu, G. Chen, J. Liu, C. Yang, Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics* **36**, 1501-1508 (2020).

6. Q. Zhao, J. Wang, G. Hemani, J. Bowden, D. S. J. A. o. S. Small, Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Stat.* **48**, 1742-1769 (2020).

7. G. Qi, N. Chatterjee, Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat. Commun.* **10**, 1941 (2019).

8. J. Morrison, N. Knoblauch, J. H. Marcus, M. Stephens, X. J. N. G. He, Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.*, 1-7 (2020).

9. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc B* **82**, 1273-1300 (2020).

10. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198**, 497-U484 (2014).

11. J. Yang, L. G. Fritsche, X. Zhou, G. Abecasis, C. International Age-Related Macular Degeneration Genomics, A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. *Am. J. Hum. Genet.* **101**, 404-416 (2017).

12. D. J. Schaid, W. N. Chen, N. B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491-504 (2018).

13. S. H. Gage, G. Davey Smith, J. J. Ware, J. Flint, M. R. Munafo, G = E: What GWAS Can Tell Us about the Environment. *PLoS Genet* **12**, e1005765 (2016).

14. Z. Yuan, H. Zhu, P. Zeng, S. Yang, S. Sun, C. Yang, J. Liu, X. Zhou, Testing and controlling for horizontal pleiotropy with probabilistic Mendelian

randomization in transcriptome-wide association studies. *Nat. Commun.* **11**, 3861 (2020).

15. H. Zhu, X. Zhou, Statistical methods for SNP heritability estimation and partition: A review. *Comput Struct Biotechnol J* **18**, 1557-1568 (2020).

16. Y. Ma, X. Zhou, Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends Genet.* **37**, 995-1011 (2021).

17. Y. Guan, M. J. T. A. o. A. S. Stephens, Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780-1815 (2011).

18. P. Carbonetto, M. Stephens, Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Anal.* **7**, 73-108, 136 (2012).

19. P. Carbonetto, X. Zhou, M. Stephens, varbvs: Fast Variable Selection for Large-scale Regression. *arXiv preprint arXiv:1608.02990*, (2017).

20. S. Burgess, F. Dudbridge, S. G. Thompson, Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.* **35**, 1880-1906 (2016).

21. S. Burgess, S. G. Thompson, Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat. Med.* **30**, 1312-1323 (2011).

22. P. Zeng, X. Zhou, S. Huang, Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics* **18**, 368 (2017).

23. A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. de Geus, D. I. Boomsma, F. A. Wright, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245-252 (2016).

24. P. Zeng, X. Zhou, Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* **8**, 456 (2017).

25. L. Liu, P. Zeng, F. Xue, Z. Yuan, X. Zhou, Multi-trait transcriptome-wide association studies with probabilistic Mendelian randomization. *Am. J. Hum. Genet.* **108**, 240-256 (2021).

26. M. Verbanck, C. Y. Chen, B. Neale, R. Do, Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693-698 (2018).

27. A. M. Johnson, J. M. Olefsky, The origins and drivers of insulin resistance. *Cell* **152**, 673-684 (2013).

28. W. T. Garvey, S. Kwon, D. Zheng, S. Shaughnessy, P. Wallace, A. Hutto, K. Pugh, A. J. Jenkins, R. L. Klein, Y. Liao, Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance. *Diabetes* **52**, 453-462 (2003).

29. Z. Zhu, Z. Zheng, F. Zhang, Y. Wu, M. Trzaskowski, R. Maier, M. R. Robinson, J. J. McGrath, P. M. Visscher, N. R. Wray, J. Yang, Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).

30. P. Zeng, T. Wang, J. Zheng, X. Zhou, Causal association of type 2 diabetes with amyotrophic lateral sclerosis: new evidence from Mendelian randomization using GWAS summary statistics. *BMC Med.* **17**, 225 (2019).

31. P. Zeng, X. Zhou, Causal effects of blood lipids on amyotrophic lateral sclerosis: a Mendelian randomization study. *Hum. Mol. Genet.* **28**, 688-697 (2019).

32. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

33. G. Qi, N. Chatterjee, A comprehensive evaluation of methods for Mendelian randomization using realistic simulations and an analysis of 38 biomarkers for risk of type 2 diabetes. *Int. J. Epidemiol.*, (2021).

34. C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M. L. Buchkovich, S. Mora, J. S. Beckmann, J. L. Bragg-Gresham, H. Y. Chang, A. Demirkan, H. M. Den Hertog, R. Do, L. A. Donnelly, G. B. Ehret, T. Esko, M. F. Feitosa, T. Ferreira, K. Fischer, P. Fontanillas, R. M. Fraser, D. F. Freitag, D. Gurdasani, K. Heikkila, E. Hypponen, A. Isaacs, A. U. Jackson, A. Johansson, T. Johnson, M. Kaakinen, J. Kettunen, M. E. Kleber, X. Li, J. Luan, L. P. Lyytikainen, P. K. E. Magnusson, M. Mangino, E. Mihailov, M. E. Montasser, M. Muller-Nurasyid, I. M. Nolte, J. R. O'Connell, C. D. Palmer, M. Perola, A. K. Petersen, S. Sanna, R. Saxena, S. K. Service, S. Shah, D. Shungin, C. Sidore, C. Song, R. J. Strawbridge, I. Surakka, T. Tanaka, T. M. Teslovich, G. Thorleifsson, E. G. Van den Herik, B.

F. Voight, K. A. Volcik, L. L. Waite, A. Wong, Y. Wu, W. Zhang, D. Absher, G. Asiki, I. Barroso, L. F. Been, J. L. Bolton, L. L. Bonnycastle, P. Brambilla, M. S. Burnett, G. Cesana, M. Dimitriou, A. S. F. Doney, A. Doring, P. Elliott, S. E. Epstein, G. Ingi Eyjolfsson, B. Gigante, M. O. Goodarzi, H. Grallert, M. L. Gravito, C. J. Groves, G. Hallmans, A. L. Hartikainen, C. Hayward, D. Hernandez, A. A. Hicks, H. Holm, Y. J. Hung, T. Illig, M. R. Jones, P. Kaleebu, J. J. P. Kastelein, K. T. Khaw, E. Kim, N. Klopp, P. Komulainen, M. Kumari, C. Langenberg, T. Lehtimaki, S. Y. Lin, J. Lindstrom, R. J. F. Loos, F. Mach, W. L. McArdle, C. Meisinger, B. D. Mitchell, G. Muller, R. Nagaraja, N. Narisu, T. V. M. Nieminen, R. N. Nsubuga, I. Olafsson, K. K. Ong, A. Palotie, T. Papamarkou, C. Pomilla, A. Pouta, D. J. Rader, M. P. Reilly, P. M. Ridker, F. Rivadeneira, I. Rudan, A. Ruokonen, N. Samani, H. Scharnagl, J. Seeley, K. Silander, A. Stancakova, K. Stirrups, A. J. Swift, L. Tiret, A. G. Uitterlinden, L. J. van Pelt, S. Vedantam, N. Wainwright, C. Wijmenga, S. H. Wild, G. Willemsen, T. Wilsgaard, J. F. Wilson, E. H. Young, J. H. Zhao, L. S. Adair, D. Arveiler, T. L. Assimes, S. Bandinelli, F. Bennett, M. Bochud, B. O. Boehm, D. I. Boomsma, I. B. Borecki, S. R. Bornstein, P. Bovet, M. Burnier, H. Campbell, A. Chakravarti, J. C. Chambers, Y. I. Chen, F. S. Collins, R. S. Cooper, J. Danesh, G. Dedoussis, U. de Faire, A. B. Feranil, J. Ferrieres, L. Ferrucci, N. B. Freimer, C. Gieger, L. C. Groop, V. Gudnason, U. Gyllensten, A. Hamsten, T. B. Harris, A. Hingorani, J. N. Hirschhorn, A. Hofman, G. K. Hovingh, C. A. Hsiung, S. E. Humphries, S. C. Hunt, K. Hveem, C. Iribarren, M. R. Jarvelin,

A. Jula, M. Kahonen, J. Kaprio, A. Kesaniemi, M. Kivimaki, J. S. Kooner, P. J. Koudstaal, R. M. Krauss, D. Kuh, J. Kuusisto, K. O. Kyvik, M. Laakso, T. A. Lakka, L. Lind, C. M. Lindgren, N. G. Martin, W. Marz, M. I. McCarthy, C. A. McKenzie, P. Meneton, A. Metspalu, L. Moilanen, A. D. Morris, P. B. Munroe, I. Njolstad, N. L. Pedersen, C. Power, P. P. Pramstaller, J. F. Price, B. M. Psaty, T. Quertermous, R. Rauramaa, D. Saleheen, V. Salomaa, D. K. Sanghera, J. Saramies, P. E. H. Schwarz, W. H. Sheu, A. R. Shuldiner, A. Siegbahn, T. D. Spector, K. Stefansson, D. P. Strachan, B. O. Tayo, E. Tremoli, J. Tuomilehto, M. Uusitupa, C. M. van Duijn, P. Vollenweider, L. Wallentin, N. J. Wareham, J. B. Whitfield, B. H. R. Wolffenbuttel, J. M. Ordovas, E. Boerwinkle, C. N. A. Palmer, U. Thorsteinsdottir, D. I. Chasman, J. I. Rotter, P. W. Franks, S. Ripatti, L. A. Cupples, M. S. Sandhu, S. S. Rich, M. Boehnke, P. Deloukas, S. Kathiresan, K. L. Mohlke, E. Ingelsson, G. R. Abecasis, C. Global Lipids Genetics, Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283 (2013).

35. D. Mozaffarian, P. W. Wilson, W. B. Kannel, Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. *Circulation* **117**, 3031-3038 (2008).

36. B. H. Brummett, M. A. Babyak, I. C. Siegler, M. Shanahan, K. M. Harris, G. H. Elder, R. B. Williams, Systolic blood pressure, socioeconomic status, and biobehavioral risk factors in a nationally representative US young adult sample. *Hypertension* **58**, 161-166 (2011).

37. A. R. Barker, L. Gracia-Marco, J. R. Ruiz, M. J. Castillo, R. Aparicio-Ugarriza, M. Gonzalez-Gross, A. Kafatos, O. Androutsos, A. Polito, D. Molnar, K. Widhalm, L. A. Moreno, Physical activity, sedentary time, TV viewing, physical fitness and cardiovascular disease risk in adolescents: The HELENA study. *Int. J. Cardiol.* **254**, 303-309 (2018).

38. N. T. Hadgraft, E. Winkler, R. E. Climie, M. S. Grace, L. Romero, N. Owen, D. Dunstan, G. Healy, P. C. Dempsey, Effects of sedentary behaviour interventions on biomarkers of cardiometabolic risk in adults: systematic review with meta-analyses. *Br. J. Sports Med.* **55**, 144-154 (2021).

39. S. Dare, D. F. Mackay, J. P. Pell, Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. *PLOS ONE* **10**, e0123579 (2015).

40. Z. Yuan, J. Ji, T. Zhang, Y. Liu, X. Zhang, W. Chen, F. Xue, A novel chi-square statistic for detecting group differences between pathways in systems epidemiology. *Stat. Med.* **35**, 5512-5524 (2016).

41. U. C. Winslow, L. Rode, B. G. Nordestgaard, High tobacco consumption lowers body weight: a Mendelian randomization study of the Copenhagen General Population Study. *Int. J. Epidemiol.* **44**, 540-550 (2015).

42. A. Hofstetter, Y. Schutz, E. Jequier, J. Wahren, Increased 24-hour energy expenditure in cigarette smokers. *N. Engl. J. Med.* **314**, 79-82 (1986).

43. R. J. Moffatt, S. G. Owens, Cessation from cigarette smoking: changes in body weight, body composition, resting metabolism, and energy consumption. *Metabolism* **40**, 465-470 (1991).

44. C. M. Ferrara, M. Kumar, B. Nicklas, S. McCrone, A. P. Goldberg, Weight gain and adipose tissue metabolism after smoking cessation in women. *Int. J. Obes.* **25**, 1322-1326 (2001).

45. Y. H. Jo, D. A. Talmage, L. W. Role, Nicotinic receptor-mediated effects on appetite and food intake. *J. Neurosci.* **53**, 618-632 (2002).

46. Y. S. Mineur, A. Abizaid, Y. Rao, R. Salas, R. J. DiLeone, D. Gundisch, S. Diano, M. De Biasi, T. L. Horvath, X. B. Gao, M. R. Picciotto, Nicotine decreases food intake through activation of POMC neurons. *Science* **332**, 1330-1332 (2011).

47. D. B. Rosoff, G. Davey Smith, N. Mehta, T. K. Clarke, F. W. Lohoff, Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study. *PLOS Med.* **17**, e1003410 (2020).

48. B. M. He, S. P. Zhao, Z. Y. Peng, Effects of cigarette smoking on HDL quantity and function: implications for atherosclerosis. *J. Cell. Biochem.* **114**, 2431-2436 (2013).

49. A. D. Gepner, M. E. Piper, H. M. Johnson, M. C. Fiore, T. B. Baker, J. H. Stein, Effects of smoking and smoking cessation on lipids and lipoproteins: outcomes from a randomized clinical trial. *Am. Heart J.* **161**, 145-151 (2011).

50. A. Linneberg, R. K. Jacobsen, T. Skaaby, A. E. Taylor, M. E. Fluharty, J. L. Jeppesen, J. H. Bjorngaard, B. O. Asvold, M. E. Gabrielsen, A. Campbell, R. E. Marioni, M. Kumari, P. Marques-Vidal, M. Kaakinen, A. Cavadino, I. Postmus, T. S. Ahluwalia, S. G. Wannamethee, J. Lahti, K. Raikkonen, A. Palotie, A. Wong, C. Dalgard, I. Ford, Y. Ben-Shlomo, L. Christiansen, K. O. Kyvik, D. Kuh, J. G. Eriksson, P. H. Whincup, H. Mbarek, E. J. de Geus, J. M. Vink, D. I. Boomsma, G. D. Smith, D. A. Lawlor, A. Kisialiou, A. McConnachie, S. Padmanabhan, J. W. Jukema, C. Power, E. Hypponen, M. Preisig, G. Waeber, P. Vollenweider, T. Korhonen, T. Laatikainen, V. Salomaa, J. Kaprio, M. Kivimaki, B. H. Smith, C. Hayward, T. I. Sorensen, B. H. Thuesen, N. Sattar, R. W. Morris, P. R. Romundstad, M. R. Munafo, M. R. Jarvelin, L. L. Husemoen, Effect of Smoking on Blood Pressure and Resting Heart Rate: A Mendelian Randomization Meta-Analysis in the CARTA Consortium. *Circulation: Cardiovascular Genetics* **8**, 832-841 (2015).

51. R. E. Luehrs, D. Zhang, G. L. Pierce, D. R. Jacobs, Jr., R. Kalhan, K. M. Whitaker, Cigarette Smoking and Longitudinal Associations With Blood Pressure: The CARDIA Study. *J. Am. Heart Assoc.* **10**, e019566 (2021).

52. X. Zhou, P. Carbonetto, M. Stephens, Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).

53. T. G. Richardson, E. Sanderson, B. Elsworth, K. Tilling, G. Davey Smith, Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. *BMJ* **369**, m1203 (2020).

54. J. Vaucher, B. J. Keating, A. M. Lasserre, W. Gan, D. M. Lyall, J. Ward, D. J. Smith, J. P. Pell, N. Sattar, G. Pare, M. V. Holmes, Cannabis use and risk of schizophrenia: a Mendelian randomization study. *Mol. Psychiatry* **23**, 1287-1292 (2018).

55. M. Gormley, T. Dudding, E. Sanderson, R. M. Martin, S. Thomas, J. Tyrrell, A. R. Ness, P. Brennan, M. Munafo, M. Pring, S. Boccia, A. F. Olshan, B. Diergaarde, R. J. Hung, G. Liu, G. Davey Smith, R. C. Richmond, A multivariable Mendelian randomization analysis investigating smoking and alcohol consumption in oral and oropharyngeal cancer. *Nat. Commun.* **11**, 6071 (2020).

56. S. Burgess, J. A. Labrecque, Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *Eur. J. Epidemiol.* **33**, 947-952 (2018).

57. M. Baiocchi, J. Cheng, D. S. Small, Instrumental variable methods for causal inference. *Stat. Med.* **33**, 2297-2340 (2014).

58. M. Lynch, B. Walsh, *Genetics and analysis of quantitative traits*. (Sinauer Associates,Sunderland,MA, 1998).

59. P. H. Allman, I. Aban, D. M. Long, S. L. Bridges, Jr., V. Srinivasasainagendra, T. MacKenzie, G. Cutter, H. K. Tiwari, A novel Mendelian randomization method with binary risk factor and outcome. *Genetic epidemiology*, (2021).

60. G. Kichaev, W. Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, B. Pasaniuc, Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).

61. G. Kichaev, B. Pasaniuc, Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet* **97**, 260-271 (2015).

62. G. Kichaev, M. Roytman, R. Johnson, E. Eskin, S. Lindstrom, P. Kraft, B. Pasaniuc, Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248-255 (2017).

63. A. J. Lea, J. Tung, X. Zhou, A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. *PLoS Genet.* **11**, e1005650 (2015).

64. S. Sun, M. Hood, L. Scott, Q. Peng, S. Mukherjee, J. Tung, X. Zhou, Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* **45**, e106 (2017).

65. X. Zhou, M. Stephens, Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821-824 (2012).

66. T. Berisa, J. K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283-285 (2016).

67. M. J. Stampfer, F. B. Hu, J. E. Manson, E. B. Rimm, W. C. Willett, Primary prevention of coronary heart disease in women through diet and lifestyle. *N. Engl. J. Med.* **343**, 16-22 (2000).

68. A. O. Odegaard, W. P. Koh, M. D. Gross, J. M. Yuan, M. A. Pereira, Combined lifestyle factors and cardiovascular disease mortality in Chinese men and women: the Singapore Chinese health study. *Circulation* **124**, 2847-2854 (2011).

69. R. R. Huxley, M. Woodward, Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *Lancet* **378**, 1297-1305 (2011).

70. E. Mountjoy, N. M. Davies, D. Plotnikov, G. D. Smith, S. Rodriguez, C. E. Williams, J. A. Guggenheim, D. Atan, Education and myopia: assessing the direction of causality by mendelian randomisation. *BMJ* **361**, k2022 (2018).

71. A. Henry, M. Katsoulis, S. Masi, G. Fatemifar, S. Denaxas, D. Acosta, V. Garfield, C. E. Dale, The relationship between sleep duration, cognition and dementia: a Mendelian randomization study. *Int. J. Epidemiol.* **48**, 849-860 (2019).

72. Q. Zhao, J. Wang, W. Spiller, J. Bowden, D. S. J. S. s. Small, Two-sample instrumental variable analyses using heterogeneous samples. *Stat. Sci.* **34**, 317-333 (2019).

73. J. Bowden, M. F. Del Greco, C. Minelli, G. Davey Smith, N. Sheehan, J. Thompson, A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783-1802 (2017).

74. F. P. Hartwig, G. Davey Smith, J. Bowden, Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985-1998 (2017).

75. M. Koller, W. A. J. C. S. Stahel, D. Analysis, Sharpening wald-type inference in robust regression for small samples. *Comput. Stat. Data Anal.* **55**, 2504-2515 (2011).

76. S. Burgess, J. Bowden, F. Dudbridge, S. G. J. a. p. a. Thompson, Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv preprint arXiv:1608.02990*, (2016).

Acknowledgements

Funding: ZY was supported by the National Natural Science Foundation of China (81872712), the Natural Science Foundation of Shandong Province (ZR2019ZD02) and the Young Scholars Program of Shandong University (2016WLJH23). XZ was supported by the National Institutes of Health grant R01HG009124. This study has been conducted using UK Biobank resource under Application Number 51470. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It also has funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK.

Author contributions: XZ conceived the idea. XZ and ZY developed the methods. ZY developed the software tool with assistance from LL. ZY performed simulations and real data analysis with assistance from LL, PG, RY and FX. XZ and ZY wrote the manuscript with input from all other authors. All authors reviewed and approved the final manuscript.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: No data are generated in the present study. The UK Biobank data is from UK Biobank resource under application number 51470. The MRAID is implemented in the R package MRAID, freely available on GitHub (https://github.com/yuanzhongshang/MRAID).

Table 1 Mean computational time (in minutes) of various two-sample MR methods.

#SNPs	MRAID	CAUSE	MRMix	IVW-R	Weighted mode	Weighted median	RAPS	Robust
1000	0.31(0.07)	2.63(1.90)	0.12(0.03)	0.0001(0.00001)	0.17(0.04)	0.011(0.003)	0.0004(0.0003)	0.0004(0.0001)
2000	1.57(0.26)	2.99(2.24)	0.13(0.03)	0.0001(0.00001)	0.18(0.03)	0.012(0.002)	0.0004(0.0006)	0.0004(0.0001)
3000	3.91(0.75)	3.66(2.36)	0.14(0.02)	0.0001(0.00002)	0.18(0.03)	0.012(0.003)	0.0004(0.0003)	0.0004(0.0001)
4000	6.82(1.35)	4.24(1.62)	0.15(0.04)	0.0001(0.00003)	0.18(0.04)	0.013(0.003)	0.0004(0.0001)	0.0004(0.0002)
5000	10.80(2.69)	4.92(2.29)	0.18(0.05)	0.0001(0.00004)	0.18(0.04)	0.013(0.002)	0.0004(0.0001)	0.0005(0.0003)

Computation is carried out on a single thread of a Xeon Gold 6138 CPU. The computation time is averaged across 20 replicates, with values inside parentheses denoting the standard deviation. #SNPs denotes the number of instrumental variables included in the model. The computational time for MRAID is based on 1,000 Gibbs sampling iterations with the first 200 as burn-in.

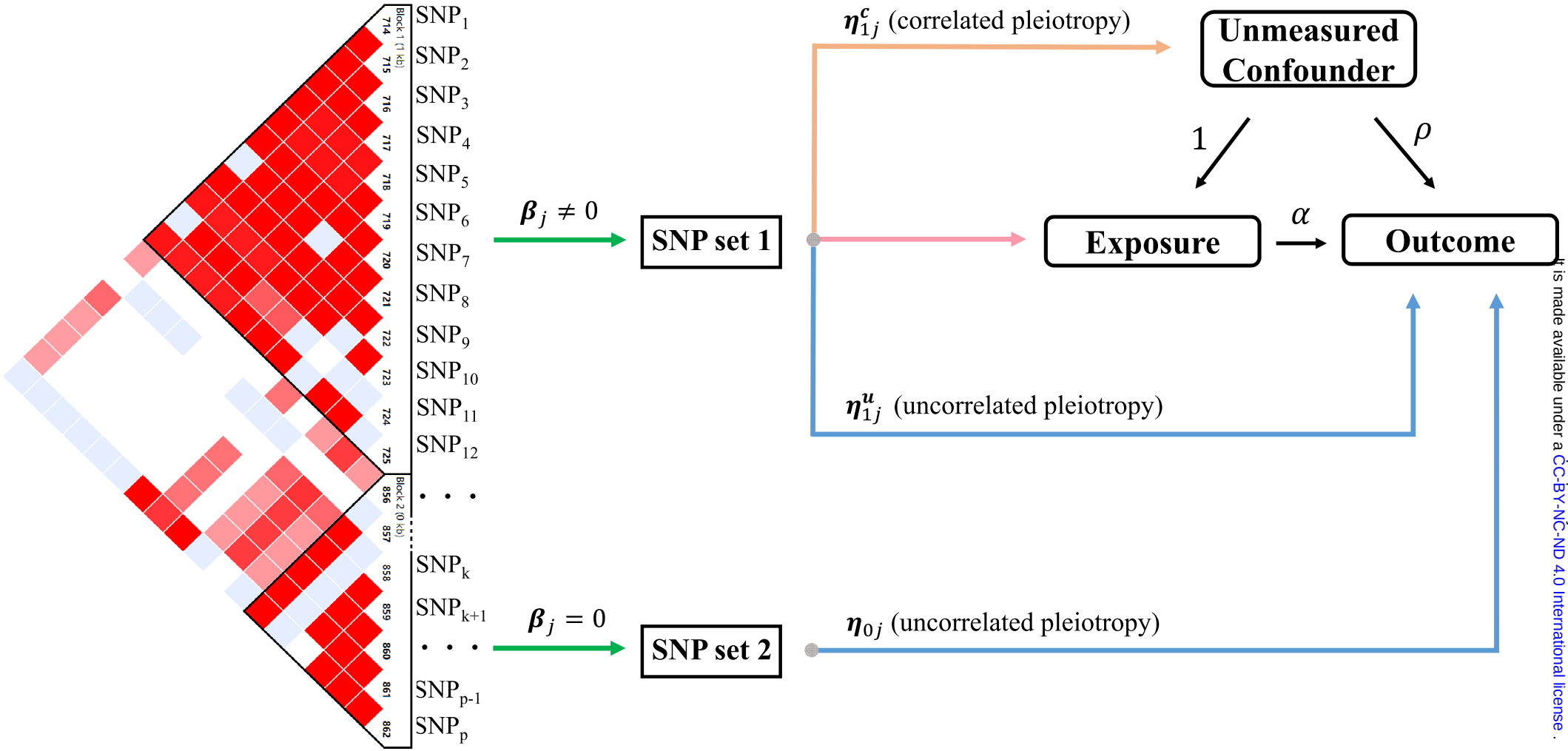

Fig. 1 Schematic of MRAID. MRAID is a Mendelian randomization method that infers the causal effect of an exposure on an outcome in the presence of unmeasured confounder by using SNPs as instrumental variables. MRAID first obtains an initial set of candidate SNP instruments that are marginally associated with the exposure (SNP1, …, SNP_p) and that are in potential LD with each other (LD plot on left). MRAID imposes a sparsity assumption on the instrumental effects of the candidate SNPs to divide instruments with non-zero effects (SNP set 1) and zero effects (SNP set 2) on the exposure. Among the selected instruments (SNP set 1), MRAID assumes that a proportion of them display horizontal pleiotropic effects that are uncorrelated with instrumental effects (blue path) and that another proportion of them display horizontal pleiotropic effects that are correlated with instrumental effects (orange path). Among the non-selected instrument candidates (SNP set 2), MRAID also assumes that a proportion of them display horizontal pleiotropic effects that are uncorrelated with instrumental effects (blue path). Overall, MRAID models jointly all genome-wide significant SNPs that are in potential LD with each other and performs automated instrument selection among them to identify suitable instruments. MRAID explicitly accounts for both correlated and uncorrelated horizontal pleiotropy and relies on a likelihood framework for effective and scalable inference.

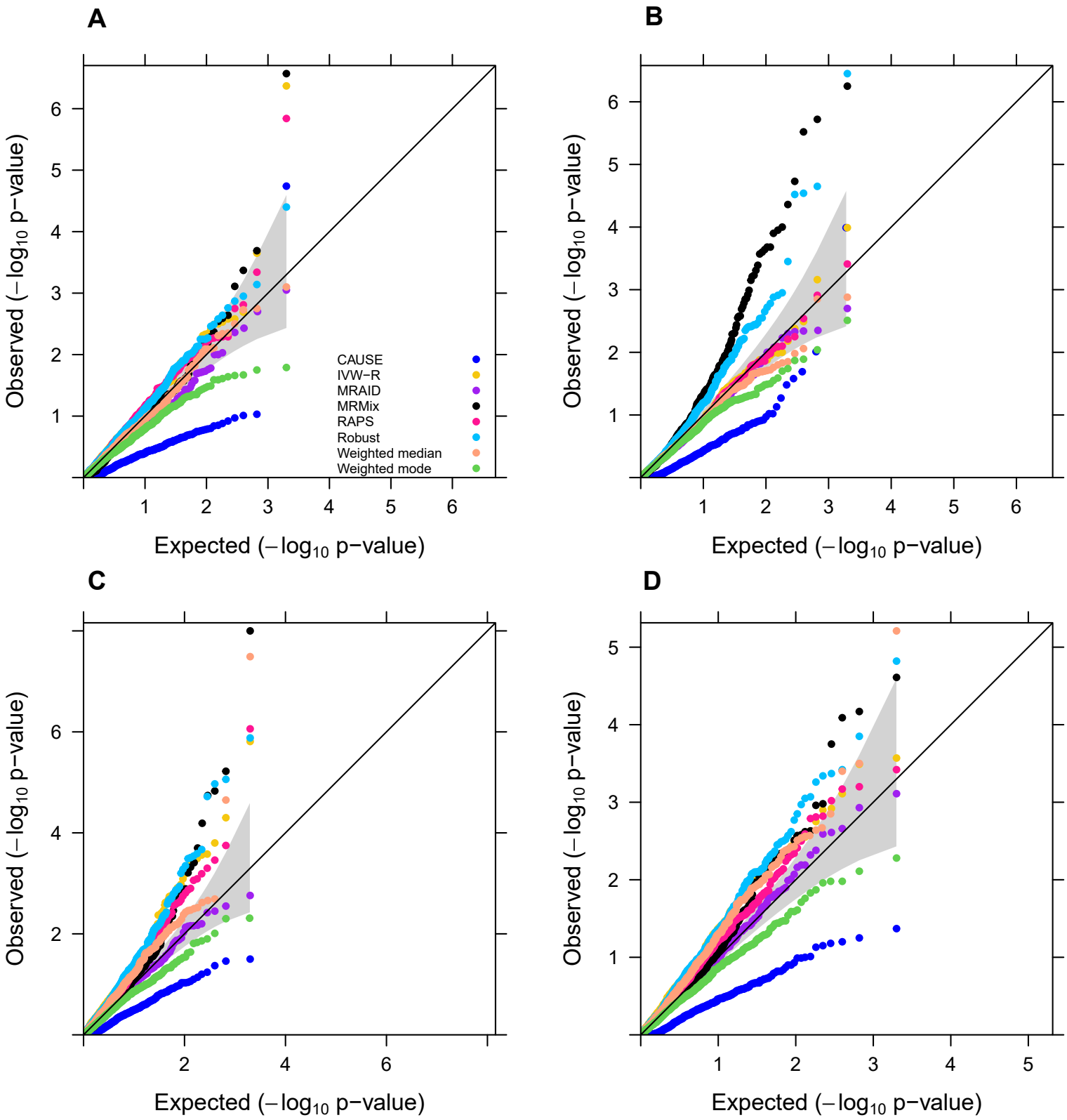

Fig. 2 Type I error control of different MR methods in simulations. Type I error is evaluated by quantile-quantile plots of -log10 p-values from different MR methods on testing the causal effect under the null simulations. Compared methods include CAUSE (blue), IVW-R (gold), MRAID (purple), MRMix (black), RAPS (deep pink), Robust (deep sky blue), Weighted median (light salmon), Weighted mode (green). Four null simulation scenarios are examined. (**A**) Null simulations in the absence of both correlated and uncorrelated horizontal pleiotropic effects. We simulated 100 instrumental SNPs with their effect sizes drawing from a normal distribution. (**B**) Null simulations in the absence of both correlated and uncorrelated horizontal pleiotropic effects. We simulated 1,000 instrumental SNPs with their effect sizes drawing from a BSLMM distribution with 1% SNPs having large effects and 99% SNPs having small effects. (**C**) Null simulations in the absence of correlated horizontal pleiotropic effect but in the presence of uncorrelated horizontal pleiotropic effect (PVE_u=5%). We simulated 100 instrumental SNPs and set the proportion of instrumental SNPs having uncorrelated horizontal pleiotropy to be 20%. (**D**) Null simulations in the presence of both correlated (π_c=5%, ρ=$\sqrt{0.05}$) and uncorrelated horizontal pleiotropic effects (PVE_u=5%). We simulated 100 instrumental SNPs and set the proportion of instrumental SNPs having the uncorrelated horizontal pleiotropy effect to be 20%.

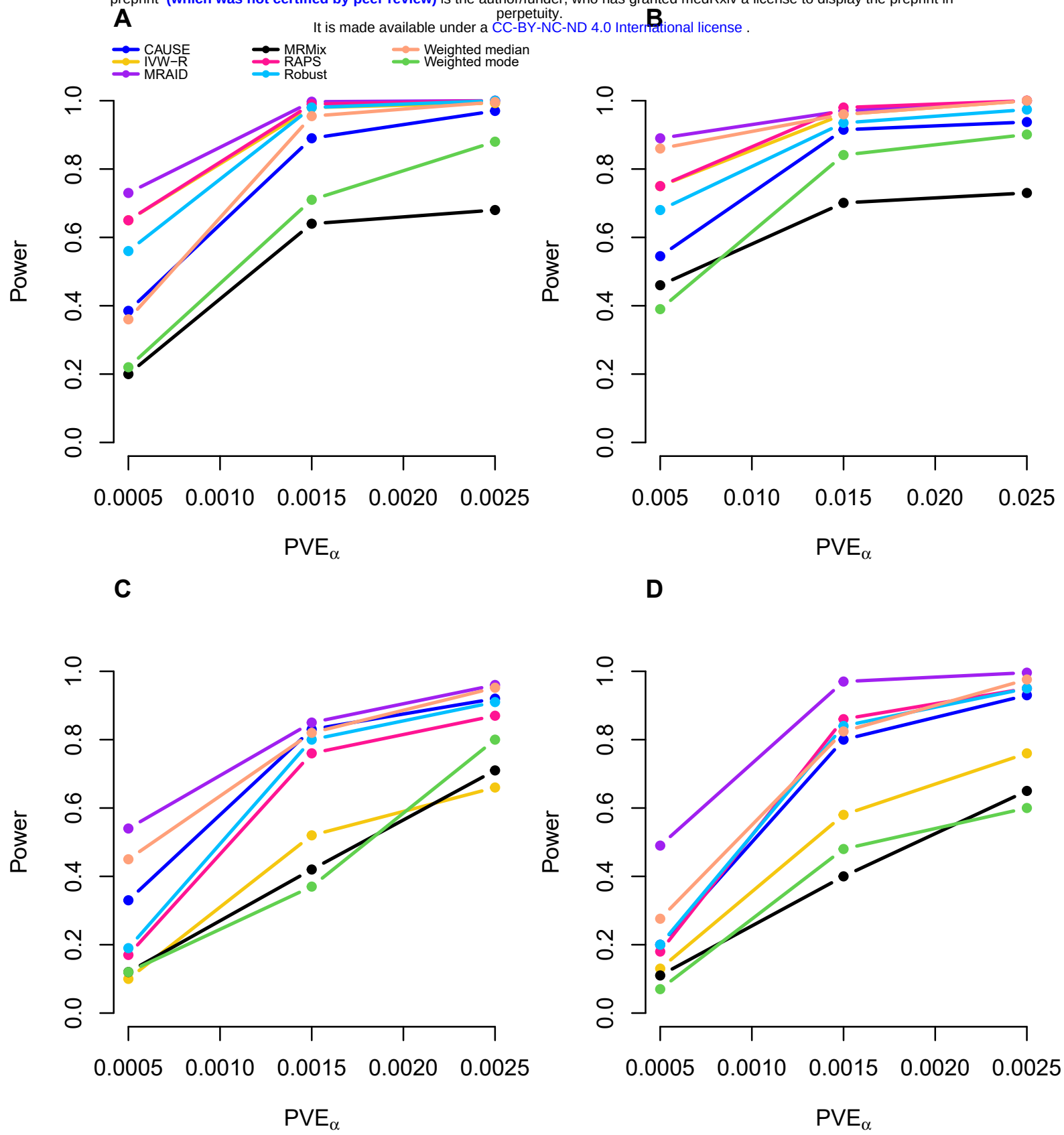

Fig. 3 Power of different MR methods in simulations. Power (y-axis) at a false discovery rate of 0.05 to detect the causal effect is plotted against different causal effect size characterized by PVE_α (x-axis). Compared methods include CAUSE (blue), IVW-R (gold), MRAID (purple), MRMix (black), RAPS (deep pink), Robust (deep sky blue), Weighted median (light salmon), Weighted mode (green). Four alternative simulation scenarios are examined. (**A**) Simulations in the absence of both correlated and uncorrelated horizontal pleiotropic effects. We simulated 100 instrumental SNPs with their effects size drawing from a normal distribution. (**B**) Simulations in the absence of both correlated and uncorrelated horizontal pleiotropic effects. We simulated 1,000 instrumental SNPs with their effects size drawing from a BSLMM distribution with 1% SNPs having large effects and 99% SNPs having small effects. (**C**) Simulations in the absence of correlated horizontal pleiotropic effect but in the presence of uncorrelated horizontal pleiotropic effect (PVE_u=5%). We simulated 100 instrumental SNPs and set the proportion of instrumental SNPs having the uncorrelated horizontal pleiotropy effect to be 30%. (**D**) Simulations in the presence of both correlated (π_c=5%, ρ=$\sqrt{0.05}$) and uncorrelated horizontal pleiotropic effects (PVE_u=5%). We simulated 100 causal instrumental SNPs and set the proportion of instrumental SNPs having the uncorrelated horizontal pleiotropy effect to be 20%.

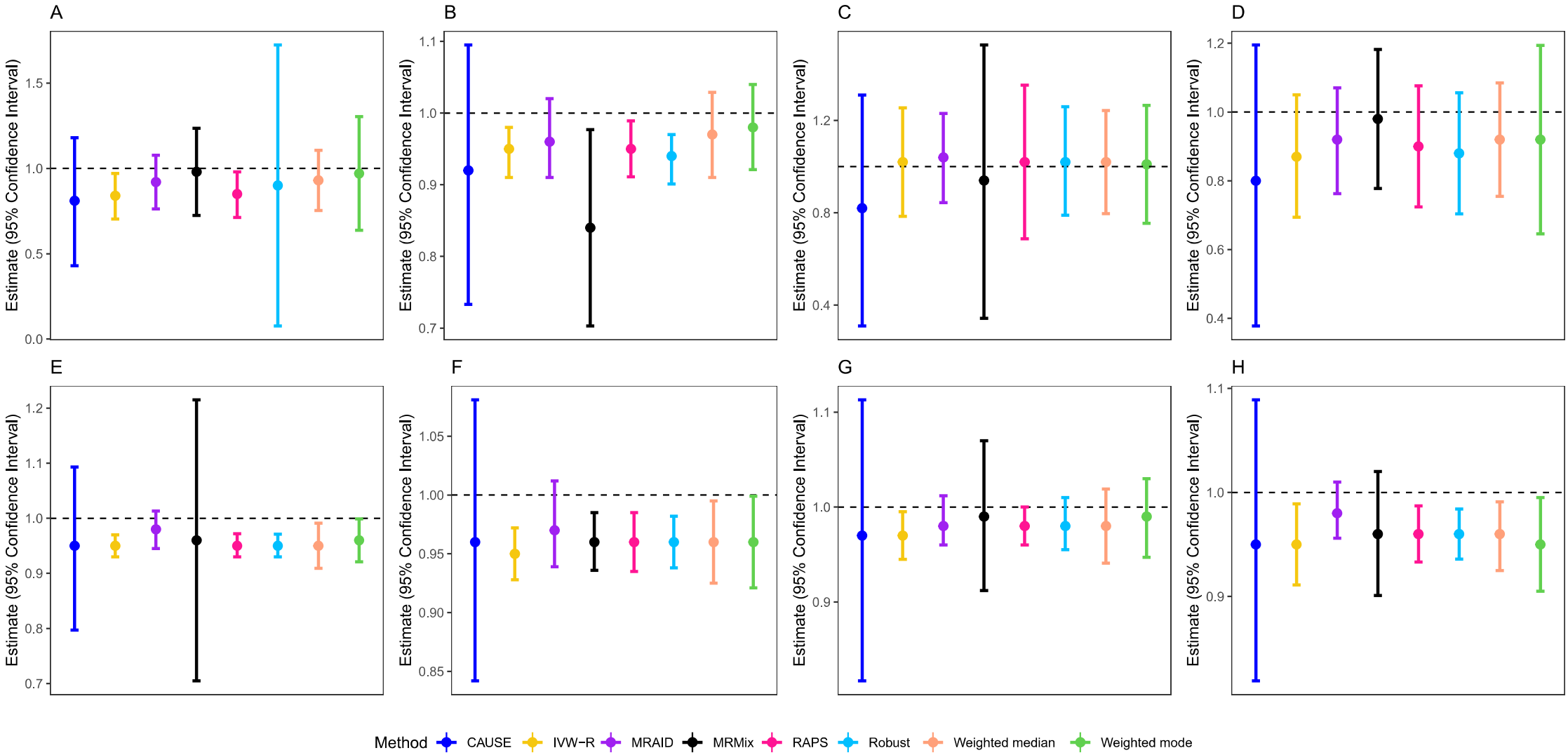

Fig. 4 Point estimates and 95% confidence intervals from different MR methods in the trait on itself analysis in the real data. Compared methods include CAUSE (blue), IVW-R (gold), MRAID (purple), MRMix (black), RAPS (deep pink), Robust (deep sky blue), Weighted median (light salmon), Weighted mode (green). Analyzed trait pairs include SBP-SBP (**A**), BMI-BMI (**B**), DBP-DBP (**C**), Pulse rate-Pulse rate (**D**), TC-TC (**E**), LDL-LDL (**F**), TG-TG (**G**), and HDL-HDL (**H**). The horizontal black dashed line in each panel represents the true causal effect size of α=1. Both MRAID and CAUSE can produce 95% confidence intervals that cover the true causal effects of all trait pairs, with CAUSE producing much larger confidence intervals than MRAID.

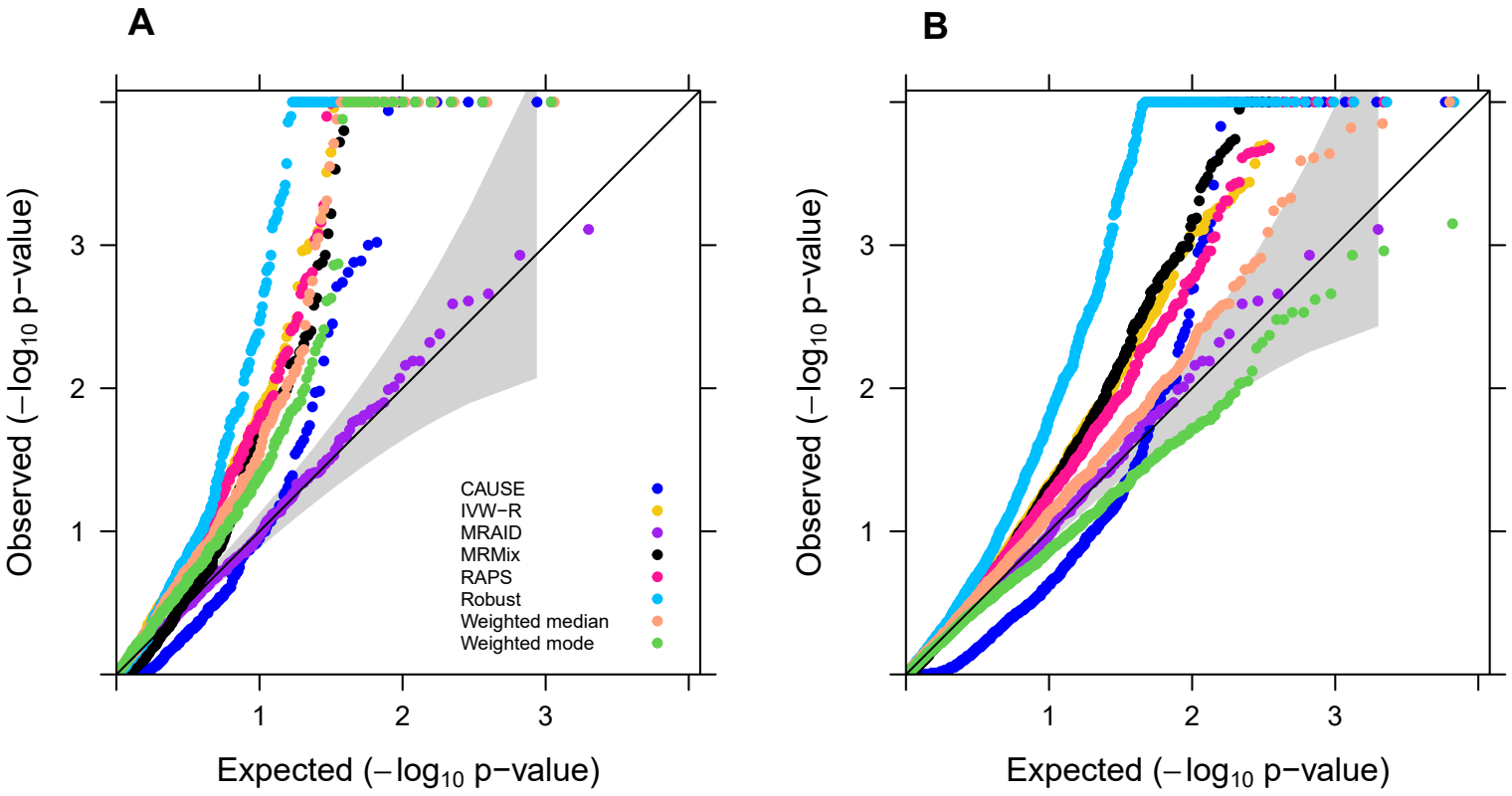

Fig. 5 Quantile-quantile plot of -log10 p-values from different MR methods on testing the causal relationship between lifestyle risk factors and CVD-related traits in UK Biobank. Compared methods include CAUSE (blue), IVW-R (gold), MRAID (purple), MRMix (black), RAPS (deep pink), Robust (deep sky blue), Weighted median (light salmon), Weighted mode (green). The results are shown for all 645 analyzed trait pairs (**A**) and the empirical null where we permuted the outcome ten times in the MR analysis of lifestyle traits on CVD- related traits (**B**).