A Hierarchical Approach Using Marginal Summary Statistics for Multiple Intermediates in a Mendelian Randomization or Transcriptome Analysis

Lai Jiang[1], Shujing Xu[1], Nicholas Mancuso[1, 2, 3], Paul J. Newcombe[4], David V. Conti[1, 2, 3 *]

[1] *Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California*

[2] *Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, California*

[3] *Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California*

[4] *MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom*

* Corresponding author. Department of Preventive Medicine, 2001 N. Soto St, SSB 202, Los Angeles CA 90032. Email: dconti@usc.edu

Abstract

Background: Previous research has demonstrated the usefulness of hierarchical modeling for incorporating a flexible array of prior information in genetic association studies. When this prior information consists of effect estimates from association analyses of single nucleotide polymorphisms (SNP)-intermediate or SNP-gene expression, a hierarchical model is equivalent to a two-stage instrumental or transcriptome-wide association study (TWAS) analysis, respectively.

Methods: We propose to extend our previous approach for the joint analysis of marginal summary statistics (JAM) to incorporate prior information via a hierarchical model (*hJAM*). In this framework, the use of appropriate effect estimates as prior information yields an analysis

similar to Mendelian Randomization (MR) and TWAS approaches such as FUSION and S-PrediXcan. However, *hJAM* is applicable to multiple correlated SNPs and multiple correlated intermediates to yield conditional estimates of effect for the intermediate on the outcome, thus providing advantages over alternative approaches.

Results: We investigate the performance of *hJAM* in comparison to existing MR approaches (inverse-variance weighted MR and multivariate MR) and existing TWAS approaches (S-PrediXcan) for effect estimation, type-I error and empirical power. We apply *hJAM* to two examples: estimating the conditional effects of body mass index and type 2 diabetes on myocardial infarction and estimating the effects of the expressions of gene *NUCKS1* and *PM20D1* on the risk of prostate cancer.

Conclusions: Across numerous causal simulation scenarios, we demonstrate that *hJAM* is unbiased, maintains correct type-I error and has increased power.

Key words: Mendelian randomization, transcriptome-wide association studies, hierarchical model, joint analysis of marginal summary statistics (JAM)

Key Messages:

- Mendelian randomization and transcriptome-wide association studies (TWAS) can be viewed as similar approaches via a hierarchical model.
- The hierarchal joint analysis of marginal summary statistics (*hJAM*) is a multivariate Mendelian randomization approach which offers a simple way to address the pleiotropy bias that is introduced by genetic variants associated with multiple risk factors or expressions of genes.

- *hJAM* incorporates the linkage disequilibrium structure of the single nucleotide polymorphism (SNPs) in a reference population to account for the correlation between SNPs.

- In addition to Mendelian randomization and TWAS, *hJAM* offers flexibility to incorporate functional or genomic annotation or information from metabolomic studies.

Introduction

Instrumental variable analysis with genetic variants has been widely used as a general framework for estimating effects of risk factors and gene expression on an outcome (Figure 1)[1-4]. Within this framework using single-nucleotide polymorphisms (SNPs) as instrumental variables, the intermediates X can be modifiable risk factors, expression of genes, or other potential intermediates such as methylation, metabolites or proteomics. To be a valid instrumental variable and to yield a causal effect of a risk factor, the genetic variants selected as the instruments must satisfy three assumptions: (1) they must not be associated with the outcome except through the intermediate, (2) they must be at least moderately associated with the intermediate, and (3) they must be independent of potential confounders of the association between the intermediate and the outcome (Figure 1). The violation of the first assumption results in a bias estimate due to pleiotropy. Weak instrument bias will be introduced if the second assumption is violated since the random error may mask the effect of the intermediate on the outcome[5]. Previous work has demonstrated that weak instruments may lead to a large bias in estimators even though the first assumption is only slightly violated[6]. Finally, the law of independent assortment of genetic variants within a homogeneous population or the ability to adequately control for potential

confounding due to population structure, often leads genetic variants fulfilling the third assumption.

Figure 1 The direct acyclic graph (DAG) for instrumental variable analysis with genetic variants.

This DAG describes the framework for several approaches. Arrow denotes a causal effect in the direction of the arrow. Solid line refers to moderate or strong association and dashed line refers to uncertain association.

Mendelian randomization (MR) and transcriptome-wide association studies (TWAS) are the two major approaches within the instrumental variable analysis framework using genetic variants. MR approaches focus on the modifiable risk factors while TWAS approaches adopt gene expression as the intermediate. One advantage of using these tools is the ubiquity of publicly-available genome-wide association studies (GWAS), such as UK Biobank[7], facilitates researchers to initiate investigation of complex traits and diseases nearly immediately[8]. The existing approaches differ in their strategies to combine the summary data from GWAS or RNA sequencing data.

In this paper, we propose an approach that leverages the joint analysis of marginal summary statistics (*JAM*)[9], a scalable algorithm designed to analyze published marginal summary statistics from GWAS under a joint multi-SNPs model to identify causal genetic variants for fine mapping. Here, we extend *JAM* with a hierarchical model to incorporate SNP-intermediate association estimates and unify the framework of MR and TWAS approaches when multiple intermediates and/or correlated SNPs exist.

Methods

Unify the framework of Mendelian Randomization and TWAS

Instrumental variable analysis with individual-level genotype data can be viewed as a two-stage hierarchical model. Using linear regression, the first stage models the outcome as a function of the genetic variants:

$$Y = G\beta + \delta. \quad (1)$$

Here, Y denotes a n-length vector of a continuous outcome, G denotes an $n \times P$ genotype matrix with P SNPs and n individuals and δ denotes the residuals. The second stage models the conditional effect estimates β as a function of prior information[10-13], $\hat{A} \in \mathbb{R}^{P \times m}$:

$$\beta = \hat{A}\pi + \epsilon. \quad (2)$$

where $\pi \in \mathbb{R}^{m \times 1}$ denotes the parameter of interest, the vector of effects for the intermediates X on outcome Y and m is the number of intermediates X. We can join these two-stage models into a single linear mixed model by substituting Eq. 2 into Eq. 1[14]:

$$Y = G\hat{A}\pi + G\epsilon + \delta = G\hat{A}\pi + \delta, \quad (3)$$

assuming there is no direct effect from the genetic variants to the outcome (i.e. $\epsilon = 0$). The estimate of π from Eq. 3 is equivalent to the result from the two-stage least square (2SLS) regression, which is employed by PrediXcan[15] and others[16]. The prior information $\hat{A}$ is the association estimates between the genetic variants and the intermediate and can be applied to impute the intermediate with the genetic variants:

$$\hat{X} = G\hat{A}. \quad (4)$$

Note that Eq. 4 is the stage-2 in the 2SLS regression and that MR approaches with summary data are developed based on Eq. 2. One key aspect of the instrumental variable analysis with genetic variants is that the $\hat{A}$ matrix is computed from a separate data, i.e. $\hat{A} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m)^T$, where $\hat{\alpha}_i$ denotes the vector of association estimates between genetic variants and i^{th} intermediate from external data. Two different $\hat{\alpha}$ vectors have been used by previous methods. Marginal estimates $\hat{\alpha}$ are widely employed by MR where marginal summary statistics from GWAS are being used[17-19]. Conditional estimates of $\hat{\alpha}$, which adjust for the correlations between estimates, can also be incorporated into the framework. One way to construct a conditional estimate $\hat{\alpha}$ is to apply regularized regression in individual-level data, such as the PredictDB developed for PrediXcan[15]. Another way is to convert the marginal estimates $\hat{\alpha}$ into conditional ones by incorporating the linkage disequilibrium (LD) block among the SNPs using the *JAM* approach[20]. To model multiple intermediates, we construct an $\hat{A}$ matrix by combining the vectors of effect estimates of the SNPs on each intermediate, $\hat{\alpha}_i$, into a matrix with the number of columns equal the number of intermediates (i.e. m):

$$\hat{A}_{P\times m} = \begin{bmatrix} \hat{\alpha}_{11} & \cdots & \hat{\alpha}_{1m} \\ \vdots & \ddots & \vdots \\ \hat{\alpha}_{p1} & \cdots & \hat{\alpha}_{pm} \end{bmatrix}.$$

Hierarchical JAM (hJAM) with summary data

We can employ the same hierarchical model to marginal summary data. Following Newcombe et al.[9], we use the marginal summary statistics, $\hat{b}$, which are obtained from a GWAS

and the minor allele frequency (MAF) of the genetic variants, $\hat{p}$, to construct a vector z representing the genotype weighted effect for each genetic variant i:

$$z_i = 2N_Y\hat{p}_i(1-\hat{p}_i)\hat{b}_{i,}$$

assuming Hardy-Weinberg Equilibrium. The MAF can be extracted from the same GWAS or using external populations such as 1000 Genomes Project[21] as reference data. Using standard linear algebra, we can express the distribution of z as

$$z \sim MVN_p\,(G_0'G_0\beta, \sigma_Y^2 G_0'G_0),$$

where $G_0'G_0$ denotes the $P \times P$ genotype variance-covariance from a centered reference data set (e.g. 1000 Genome[21]) to obtain the conditional effects of SNPs on the outcome, β. Details are described in Newcombe et al.[9]. To simplify the likelihood, we perform a Cholesky decomposition transformation $L'L = G_0'G_0$. Then, we transform z into z_L with the inverse of L' as $z_L = L'^{-1}z$. When L is positive semi-definite, we add a ridge term, i.e. a small positive element, on the diagonal to enforce it to be a positive definite matrix. The regularization term has a very small effect on the estimates while guaranteeing the invertibility of the L matrix. Then, the z_L is a vector of independent statistics that can be expressed as

$$z_L \sim MVN_p\left(L\beta, \sigma_Y^2 I_p\right). \quad (5)$$

Similar to above, we then fit a hierarchical model by incorporating the second-stage model (Eq. 2) into Eq.5 and construct the *hJAM* model as

$$z_L \sim MVN_p\left(L\hat{A}\pi, \sigma_Y^2 I_p\right), \quad (6)$$

assuming no direct effect from genetic variants to the outcome. Here, $\hat{\pi}$ denotes the association parameter of interest between the intermediate and outcome and is estimated using maximum likelihood and the statistical significance is given by a Wald test. The estimate of $\hat{\pi}$ and corresponding variance are

$$\hat{\pi} = \left(\left(L\hat{A}\right)'\left(L\hat{A}\right)\right)^{-1}\left(L\hat{A}\right)'z_L$$

and

$$\mathrm{Var}(\hat{\pi}) = \left(\left(L\hat{A}\right)'\left(L\hat{A}\right)\right)^{-1}\sigma_Y^2.$$

Egger-type approaches can be implemented in this framework by allowing an intercept in Eq. 6 by adding a column of ones to $\hat{A}$ matrix, which is analogue to MR Egger regression[22].

Simulation studies

To assess the performance of *hJAM*, we performed an extensive set of simulation studies. For each simulation, we simulated three standardized individual genotype matrices G_X, G_Y, and G_L, an intermediate matrix X, and an outcome vector Y. We then generated the summary statistics, including marginal effects $\hat{b}$, $\hat{A}$, and the reference LD structure, from the individual-level data.

For each genotype matrix, we had two inter-block relationships: no LD and moderate LD ($R = 0.6$). Each SNP block (i.e. G_1, , G_2 and G_3 in Figure 2) contains 10 SNPs, in which we set 3 SNPs to be causal to the intermediate with $R^2_{G,X} = 0.1$. The MAF was sampled from a uniform distribution (0.05, 0.3). Sample size for each genotype data set was set to be N_{G_X}= 1000, N_{G_Y} =

5000, and N_{G_L} = 500, respectively. We simulated two X's and four scenarios representing different causal models for the two intermediates likely to be encountered in epidemiologic studies (Figure 2). For scenario A, X_1 and X_2 were independent. For scenarios B and D, X_1 and X_2 were correlated through a shared SNPs set *G3*. The coefficient λ in the causal scenarios (Figure 2 (C) and (D)) was simulated by $R^2_{X_1,X_2} = 0.2$. These simulation scenarios are similar to those described in Sanderson et al.[23].

The primary objective was to estimate $\hat{\pi}$ with each true π_i being set to null ($\pi_i = 0$) or a positive effect ($\pi_i = 0.1$). To mimic applied applications and to ensure selection of at least two or more SNPs, a forward selection on $\hat{A}$ was performed to exclude the noninformative variants with a threshold $P < 0.2$ in the analysis step. We compared the performance of our approach to inverse-variance weighted MR (IVW MR)[17], multivariate inverse-variance weighted MR (MVIVW MR)[18], and S-PrediXcan[24] (see Appendix). All simulation analyses were performed in R version 3.4.0. Results were calculated from 1000 replications for each scenario. All tests were two-sided with a type-I error of 0.05.

Figure 2 Simulation scenarios of different relationships between X's.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated.

Simulation results

The average estimate and standard error for π across the four simulation scenarios are presented in Figure 3 and supplementary Table 1 to Table 4, respectively. Supplementary Figure 1 (independent SNPs) and Figure 4 (correlated SNPs) present results for empirical power.

Results from the base scenario A, where X_1 and X_2 were independent, demonstrate that the estimates from most methods were unbiased. However, when IVW MR and MVIVW MR do not incorporate the LD structure, there is a slightly inflated type-I error under simulation scenarios with correlated SNPs. IVW MR with and without correlation had a less precise estimate and lower power compared to the other methods in scenario A (Table 1). When a pleiotropic effect was simulated for each intermediate (scenario B to D), the estimates from *hJAM* and MVIVW MR with LD were unbiased and had a correct type-I error for the corresponding intermediate (Figure 4). The estimates from MVIVW MR without LD were unbiased but showed an inflated type-I error due to a smaller estimated standard error in scenarios in which SNPs were correlated (Figure 4). IVW MR and S-PrediXcan had a biased estimate and an inflated type-I error regardless of the correlation structure of the SNPs. The results for MVIVW MR and IVW MR reflect specification of the LD structure for the instruments when using the *MedelianRandomization*[25] package. Results without the LD structure showed a poor performance as indicated by increased type-I errors.

Supplementary Table 1 The estimate and its standard error of simulation scenario A: independent *X*'s.

Supplementary Table 2 The estimate and its standard error of simulation scenario B: correlated *X*'s.

Supplementary Table 3 The estimate and its standard error of simulation scenario C: X_1 causes X_2.

Supplementary Table 4 The estimate and its standard error of simulation scenario D: X_1 causes X_2 and X_1 and X_2 are correlated.

Figure 3 Average estimates and 95% confidence intervals of the correlated SNPs scenarios across 1000 replications.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated. The black solid line refers to the default Type-I error, $\alpha = 0.05$.

Figure 4 Empirical Power of the correlated SNPs scenarios across 1000 replications.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated. The black solid line refers to the default Type-I error, $\alpha = 0.05$.

Supplementary Figure 1 Empirical Power of the independent SNPs scenarios across 1000 replications.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated. The black solid line refers to the default Type-I error, $\alpha = 0.05$

Data Examples

To demonstrate *hJAM* on real data, we applied various methods to two examples: 1) for body mass index (BMI) and type 2 diabetes (T2D) on myocardial infarction (MI); and 2) gene expression and prostate cancer risk. As the study populations for both examples include individuals of European ancestry, we used the 503 European-ancestry subjects from the 1000 Genomes Project[21] as our reference data for the LD structure.

Causal effect of BMI and T2D on myocardial infarction

Previous studies have shown that obesity[26, 27] and T2D[28, 29] are two important risk factors for MI. In addition, the association between obesity and T2D is well-established[30, 31]. A directed acyclic graph (DAG) shows the relationships between the two risk factors and MI (Figure 5).

Figure 5 Direct acyclic graph (DAG) of the relationship between BMI, type 2 diabetes and myocardial infarction.

To examine the conditional effects of the two risk factors, we extracted the summary statistics for MI, BMI, and T2D from the UK Biobank (n = 459,324)[7], GIANT consortium (n =

339,224)[32], and DIAGRAM+GERA+UKB (n = 659,316)[33], respectively. In total, 75 SNPs and 136 SNPs were identified as genome-wide significant for BMI and T2D, respectively (supplementary Figure 2 and supplementary Figure 3). In this set of SNPs, there was one overlapping SNP in both the instrument sets for BMI and T2D (rs7903146, α_{BMI} = -0.016, P_{BMI} = 1.4×10^{-14}, α_{T2D} = 0.319, and P_{T2D} = 1.7×10^{-204}). This SNP is a well-known T2D associated SNP and has being identified as a BMI-associated hit in GIANT. Additionally, four correlated pairs of SNPs exist between the two sets (supplementary Table 5). We re-orientated the effects of all SNPs but one (except the effect of the overlapping SNP rs7903146 on BMI) to have a positive effect and we used MR Egger regression[22] and *hJAM* Egger to detect a potential directional pleiotropy effect.

Results are shown in Table 1. All methods suggested a significantly increasing risk of MI with an increased BMI and the presence of T2D. This agrees with previous studies[26, 28]. The magnitude of *hJAM* and MVIVW MR were similar while IVW MR and S-PrediXcan showed larger estimated values. The odds ratio (OR) from *hJAM* for the risk of MI was 1.38 (95% CI=1.22, 1.56) and 1.16 (95% CI=1.12, 1.20) for per one unit increase in BMI and having T2D, respectively. MVIVW MR with LD has similar estimates with 1.37 (95% CI=1.22, 1.54) and 1.15 (95% CI=1.11, 1.19) for BMI and having T2D, respectively. The difference in estimates between the multivariate approaches and the univariate MR/TWAS approaches may be attributed to potential pleiotropy not accounted for in the analyses that do not model the intermediates jointly. When modeled jointly, results from *hJAM* Egger suggested that there was no residual pleiotropy detected when we incorporated both BMI- and T2D-associated instruments in the analysis ($P = 0.57$). In contrast, the MR-Egger approach applied univariately to T2D resulted in

a significant test for the intercept, suggesting the presence of pleiotropy, potentially due to association of some of the SNPs to the outcome via BMI.

Table 1 Causal odds ratios (95% confidence interval) for myocardial infarction per unit in body mass index and having type 2 diabetes.

Methods	Odds ratios (95% CI)	P
BMI		
hJAM	1.38 (1.22, 1.56)	3.19E-07
MVIVW MR	1.37 (1.22, 1.54)	1.94E-07
MVIVW MR (w/o LD)	1.34 (1.20, 1.49)	1.65E-07
IVW MR	1.54 (1.32, 1.79)	2.07E-08
IVW MR (w/o LD)	1.53 (1.32, 1.77)	1.45E-08
S-PrediXcan	1.66 (1.58, 1.74)	9.88E-96
Egger-intercept	0.005 (-0.003, 0.013)	2.00E-01
T2D		
hJAM	1.16 (1.12, 1.20)	4.12E-11
MVIVW MR	1.15 (1.11, 1.19)	8.34E-12
MVIVW MR (w/o LD)	1.16 (1.11, 1.20)	1.29E-11
IVW MR	1.15 (1.11, 1.20)	1.77E-14
IVW MR (w/o LD)	1.15 (1.11, 1.20)	1.98E-14
S-PrediXcan	1.14 (1.11, 1.16)	9.43E-109
Egger-intercept	0.007 (0.000, 0.013)	0.017

Abbreviation: w/o LD, without linkage disequilibrium adjustment; s.e., standard error; 95% CI, 95% confidence interval.
Note: * For MR Egger (intercept), we showed log odds ratio and its 95% CI.

Supplementary Table 5 Four correlated pairs of SNPs in the instrument sets of BMI and type 2 diabetes.

Supplementary Figure 2 Scatter plots for the univariate effect estimates $\hat{\alpha}$ vs. $\hat{\beta}$ for BMI (A) and type 2 diabetes (B).

Supplementary Figure 3 Heatmap of the Pearson correlation between the 210 instrumental SNPs in data example 1: BMI and type 2 diabetes of myocardial infarction.

Causal effect of PM20D1 and NUCKS1 on prostate cancer risk

To further illustrate the benefit of *hJAM*, we next considered the gene-prostate cancer risk association of two genes on chromosome 1q32.1, gene *PM20D1* (Peptidase M20 Domain Containing 1) and gene *NUCKS1* (Nuclear Casein Kinase and Cyclin Dependent Kinase Substrate 1). Both *PM20D1* and *NUCKS1* are protein coding genes and previous transcriptome studies have found a significant effect of both *PM20D1* and *NUCKS1* on the risk of prostate cancer among a European-ancestry population[34, 35]. Due to the close proximity of the two genes along the genome, there is a potential for a univariate approach to result in biased estimates. To examine the effects jointly, we applied *hJAM* to this research question.

We constructed the $\hat{A}$ matrix with 114 marginally significant expression quantitative trait loci (eQTL) estimates for the two genes from GTEx v7[36]. Among the 114 eQTLs, one eQTL has significant associations with both *PM20D1* and *NUCKS1*. To limit the correlation between the eQTLs ($|R| > 0.9$), we used priority pruner[37] to prune the eQTLs by limited the absolute pairwise correlation coefficient $|R| < 0.7$ and using the magnitude of the eQTLs association for each gene as the priority criteria. After pruning, we had 12 eQTLs in the analysis set (supplementary Figure 3). The genome-wide summary statistics for the risk of prostate cancer was taken from a published GWAS with more than 140,000 European-ancestry men[38].

Table 2 presents results from *hJAM* and the competing approaches. *hJAM* and MVIVW MR with LD yield non-significant results for both *PM20D1* and *NUCKS1* for the risk of prostate cancer ($P_{\mathrm{PM20D1}} = 0.90$ and $P_{\mathrm{NUCKS1}} = 0.21$ for hJAM, and $P_{\mathrm{PM20D1}} = 0.90$ and $P_{\mathrm{NUCKS1}} = 0.17$ for MVIVW MR with LD). However, univariate models, including IVW MR and S-PrediXcan, results in a significant positive effect for prostate cancer risk for *PM20D1* and *NUCKS1* ($P_{\mathrm{PM20D1}} = 0.024$ and $P_{\mathrm{NUCKS1}} = 3.53 \times 10^{-15}$ for IVW MR without correlation, and

$P_{\mathrm{PM20D1}} = 0.003$ and $P_{\mathrm{NUCKS1}} = 2.84 \times 10^{-10}$ for S-PrediXcan). We consider the significance in the univariate models was due to the correlation between the two genes and the LD between the eQTLs, which could be adjusted for by the *hJAM* and MVIVW MR with LD models.

Table 2 Causal odds ratios (95% confidence interval) for prostate cancer risk per unit increasing in gene expression reads.

Methods	Odds ratio (95% CI)	P
PM20D1		
hJAM	0.10 (0.92, 1.08)	0.91
MVIVW MR	0.10 (0.93, 1.07)	0.90
MVIVW MR (w/o LD)	0.99 (0.94, 1.04)	0.66
IVW MR	1.02 (0.96, 1.10)	0.49
IVW MR (w/o LD)	1.03 (1.00, 1.05)	0.02
S-PrediXcan	1.01 (1.00, 1.01)	0.003
NUCKS1		
hJAM	1.12 (0.93, 1.36)	0.21
MVIVW MR	1.12 (0.95, 1.33)	0.17
MVIVW MR (w/o LD)	1.15 (0.10, 1.33)	0.06
IVW MR	1.16 (1.10, 1.21)	5.03E-10
IVW MR (w/o LD)	1.16 (1.12, 1.20)	3.53E-15
S-PrediXcan	1.10 (1.07, 1.13)	2.84E-10

Abbreviation: w/o LD, without linkage disequilibrium adjustment; s.e., standard error; 95% CI, 95% confidence interval.

Supplementary Figure 4 Heatmap of the Pearson correlation between the 12 instrumental SNPs in data example 2: the effect of two gene expressions on the risk of prostate cancer.

Discussion

In this paper, we have proposed a two-stage hierarchical model which unifies the framework of Mendelian randomization and transcriptome-wide association tools and can be applied to correlated instruments and multiple intermediates. We have implemented the method in an R package which is available on Github (https://github.com/lailylajiang/hJAM).

When only one intermediate or multiple independent intermediates present, *hJAM* yield an equivalent estimate and standard error to alternative approaches (see Appendix). However, when intermediates are correlated, only MVIVW MR showed a comparable performance with *hJAM* under the independent SNPs scenarios. For correlated SNPs scenarios, when the LD structure is specified, *hJAM*'s estimates are empirically equivalent to MVIVW MR although the two approaches use slightly different weighted matrices – *hJAM* uses the adjusted variance-covariance matrix of SNPs from a reference panel while MVIVW MR uses an inverse-variance matrix. Nevertheless, we believe that the *hJAM* formulation offers several advantages in flexibility to specify the $\hat{A}$ matrix. As in TWAS, this matrix can specify eQTL estimates or as in more classical MR approaches this can specify SNP-intermediate associations. Moreover, it can incorporate other types of prior information such as functional or genomic annotation or information from metabolomic studies[39]. Inclusion of this type of annotation information can offer potential advantages for characterization of SNP effects as demonstrated in the hierarchical modeling context[10, 11, 40]. Future research needs to be performed on how best to construct this matrix for various types of intermediates.

Although *hJAM* provides an overall improvement over most existing MR methods, it is also susceptible to the caveats of these types of approaches. Firstly, it may subject to the bias in estimation due to unknown horizontal pleiotropy. *hJAM* can be extended to include a column of ones to the $\hat{A}$ matrix to allow for estimating an intercept term to formally test for directional pleiotropy, analogous to MR Egger[22]. This *hJAM*-Egger version showed a similar performance to the univariate MR-Egger regression with unbiased estimates under simulations in which the horizontal pleiotropy is balanced, but bias estimates in the presence of unbalanced pleiotropy (results not shown)[22]. *hJAM*-Egger can be applied as a sensitivity analysis of a multivariable

framework MR analysis[41]. An extension of the current *hJAM* approach could include variable selection to assess the pleiotropy assumption before incorporating the $\hat{A}$ matrix into the model. Several approaches have been proposed, such as JAM MR[42] and MR-presso[43]. Secondly, the effects of the SNPs on the intermediates and the outcome, and the causal effect of intermediates on the outcome may be non-linear (e.g. interactions). One way to address such limitation is to use summary data from stratified GWAS; however, it may attenuate the power due to a smaller sample size of the subset GWAS.

In applied applications, population structure may introduce potential difficulties for *hJAM*, as is similar for all MR and TWAS approaches using summary statistics. First, there is the reliance that the association statistics are unbiased due to potential confounding by population structure. This includes summary data for the SNPs to intermediate associations in $\hat{A}$ matrix, as well as the marginal SNP-outcome associations using within the *hJAM* model. However, given that modern techniques to account for population structure are often sufficient[44, 45], this is a fair assumption. Additionally, to account for the correlation structure between SNPs, *hJAM* assumes that the LD structure estimated from the reference data is the same as the study data used to generate the summary statistics. Since *hJAM* and MVIVW MR incorporate the correlation structure of SNPs in a slightly different weight matrices, there is the potential for this to impact these methods differently. Although, in a limited set of simulations we found that both methods are fairly robust to scenarios in which the reference data and the association data have modest differences in LD structures (results not shown).

In contrast to most current methods that rely on independent SNPs or analyze intermediates in isolation, we propose a two-stage hierarchical model to jointly model summary statistics (*hJAM*) for correlated SNPs and multiple intermediates within Mendelian

Randomization and transcriptome-wide association studies. As technology expands the potential use of these types of studies to proteomic, methylation and metabolomic data, such flexible approaches will be needed to account for the potential increase in complexity in underlying relationships between factors.

Funding

This work was supported by National Cancer Institute at the National Institutes of Health [grant P01CA196569 and R01CA140561]. Paul J. Newcombe was funded by the UK Medical Research Council [Unit Programme number MC_UU_00002/9] and also acknowledges support from the NIHR Cambridge Biomedical Research Centre.

Acknowledgement

The authors thank Drs. Duncan C. Thomas, William Gauderman, and Juan Pablo Lewinger for valuable discussions and comments throughout development.

References

1. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. *International journal of epidemiology* 2004; **33**: 21-5.
2. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research* 2007; **16**: 309-30.
3. Greenland S. An introduction to instrumental variables for epidemiologists. *International journal of epidemiology* 2000; **29**: 722-9.
4. McKeigue PM, Campbell H, Wild S, et al. Bayesian methods for instrumental variable analysis with genetic instruments ('Mendelian randomization'): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome. *International journal of epidemiology* 2010; **39**: 907-18.
5. Newhouse JP, McClellan M. Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health* 1998; **19**: 17-34.
6. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables application and limitations. *Epidemiology* 2006; **17**: 260-7.
7. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**: e1001779.
8. Burgess S, Davey Smith G. How humans can contribute to Mendelian randomization analyses. Oxford University Press; 2019.
9. Newcombe PJ, Conti DV, Richardson S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet Epidemiol* 2016; **40**: 188-201.
10. Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrum: genetic structure and spatial relations. *The American Journal of Human Genetics* 2003; **72**: 351-63.
11. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2007; **31**: 871-82.
12. Thomas DC, Conti DV, Baurley J, Nijhout F, Reed M, Ulrich CM. Use of pathway information in molecular epidemiology. *Human genomics* 2009; **4**: 21.
13. Greenland S. Principles of multilevel modelling. *International journal of epidemiology* 2000; **29**: 158-67.
14. Witte JS, Greenland S, Kim L-L, Arab L. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology* 2000; **11**: 684-8.
15. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015; **47**: 1091-8.
16. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016; **48**: 245-52.
17. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013; **37**: 658-65.

18. Burgess S, Thompson SG. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *American Journal of Epidemiology* 2015; **181**: 251-60.
19. Burgess S, Bowden J. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods. *arXiv preprint arXiv:151204486* 2015.
20. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44**: 369-75, S1-3.
21. Consortium GP. A global reference for human genetic variation. *Nature* 2015; **526**: 68.
22. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015; **44**: 512-25.
23. Sanderson E, Smith GD, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single sample and two-sample summary data settings. *bioRxiv* 2018: 306209.
24. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 2018; **9**: 1825.
25. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* 2017; **46**: 1734-9.
26. Yusuf S, Hawken S, Ounpuu S, et al. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: a case-control study. *The Lancet* 2005; **366**: 1640-9.
27. Lauer MS, Anderson KM, Kannel WB, Levy D. The impact of obesity on left ventricular mass and geometry: the Framingham Heart Study. *Jama* 1991; **266**: 231-6.
28. Manson JE, Colditz GA, Stampfer MJ, et al. A prospective study of maturity-onset diabetes mellitus and risk of coronary heart disease and stroke in women. *Archives of internal medicine* 1991; **151**: 1141-7.
29. Barrett-Connor EL, Cohn BA, Wingard DL, Edelstein SL. Why is diabetes mellitus a stronger risk factor for fatal ischemic heart disease in women than in men?: the Rancho Bernardo Study. *Jama* 1991; **265**: 627-31.
30. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* 2006; **444**: 840.
31. Group LAR. Long term effects of a lifestyle intervention on weight and cardiovascular risk factors in individuals with type 2 diabetes: four year results of the Look AHEAD trial. *Archives of internal medicine* 2010; **170**: 1566.
32. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**: 197-206.
33. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* 2018; **9**: 2941.
34. Wu L, Wang J, Cai Q, et al. Identification of novel susceptibility loci and genes for prostate cancer risk: A transcriptome-wide association study in over 140,000 European descendants. *Cancer research* 2019: canres. 3536.2018.

35. Mancuso N, Gayther S, Gusev A, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature communications* 2018; **9**: 4079.
36. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. *Nature genetics* 2013; **45**: 580.
37. Edlund CK, Anker M, Schumacher FR, Gauderman WJ, Conti DV. Priority Pruner.
38. Schumacher FR, Al Olama AA, Berndt SI, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* 2018; **50**: 928-36.
39. Gusev A, Shi H, Kichaev G, et al. Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nature communications* 2016; **7**: 10979.
40. Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *The American Journal of Human Genetics* 2007; **81**: 397-404.
41. Rees JM, Wood AM, Burgess S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in medicine* 2017; **36**: 4705-18.
42. Gkatzionis A, Burgess S, Conti D, Newcombe PJ. Bayesian variable selection with a pleiotropic loss function in Mendelian randomization. *BioRxiv* 2019: 593863.
43. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* 2018; **50**: 693-+.
44. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 2006; **38**: 904.
45. Runcie DE, Crawford L. Fast and flexible linear mixed models for genome-wide genetics. *PLoS genetics* 2019; **15**: e1007978.

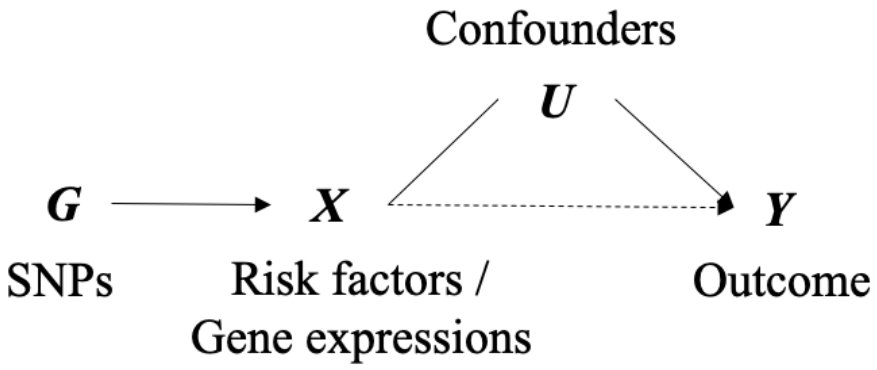

Figure 1 The direct acyclic graph (DAG) for instrumental variable analysis with genetic variants.

This DAG describes the framework for several approaches. Arrow denotes a causal effect in the direction of the arrow. Solid line refers to moderate or strong association and dashed line refers to uncertain association.

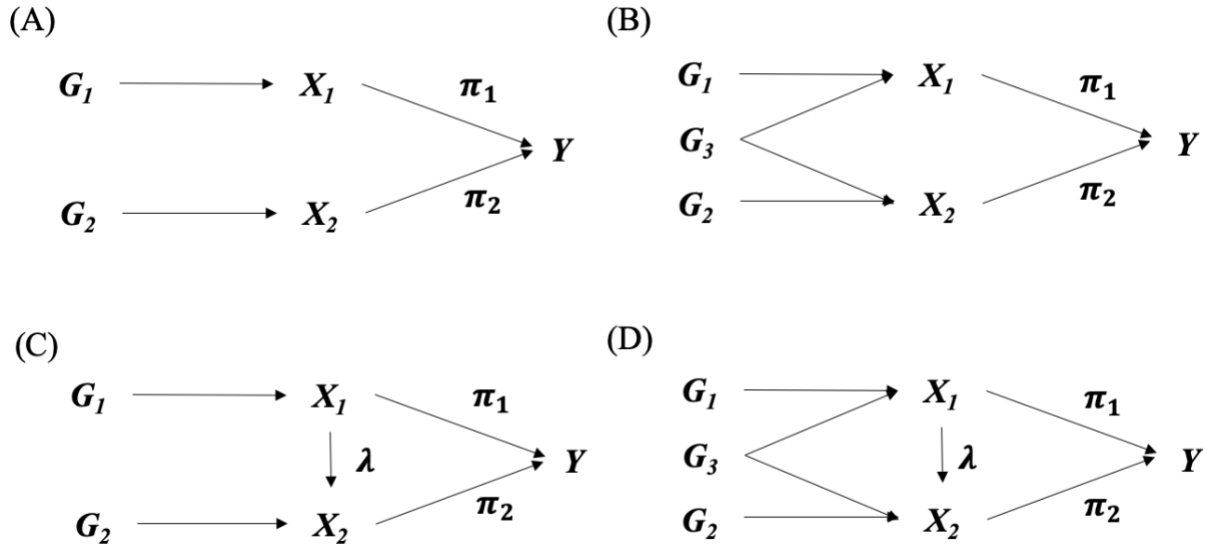

Figure 2 Simulation scenarios of different relationships between X's.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated.

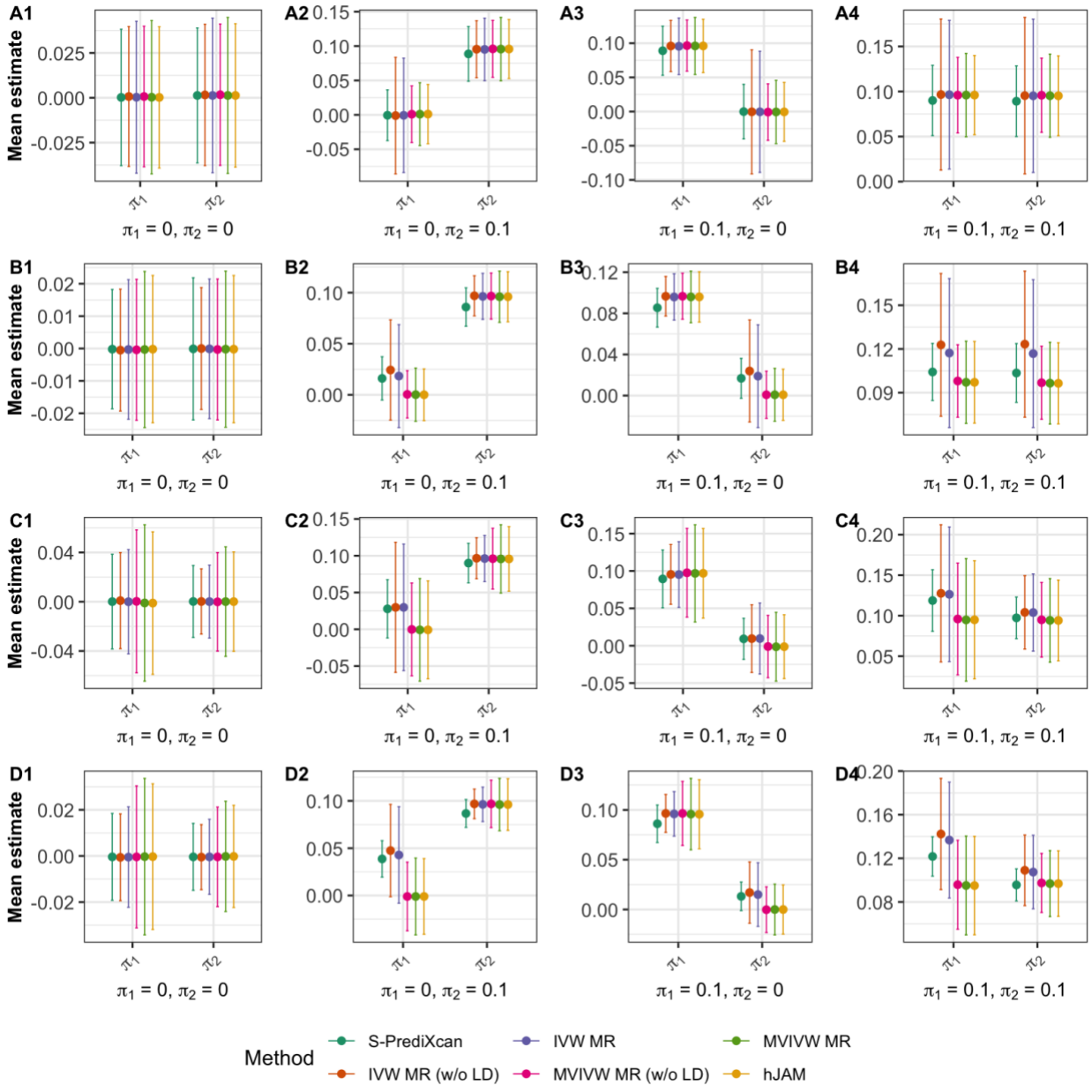

Figure 3 Average estimates and 95% confidence intervals of the correlated SNPs scenarios across 1000 replications.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated. The black solid line refers to the default Type-I error, $\alpha = 0.05$.

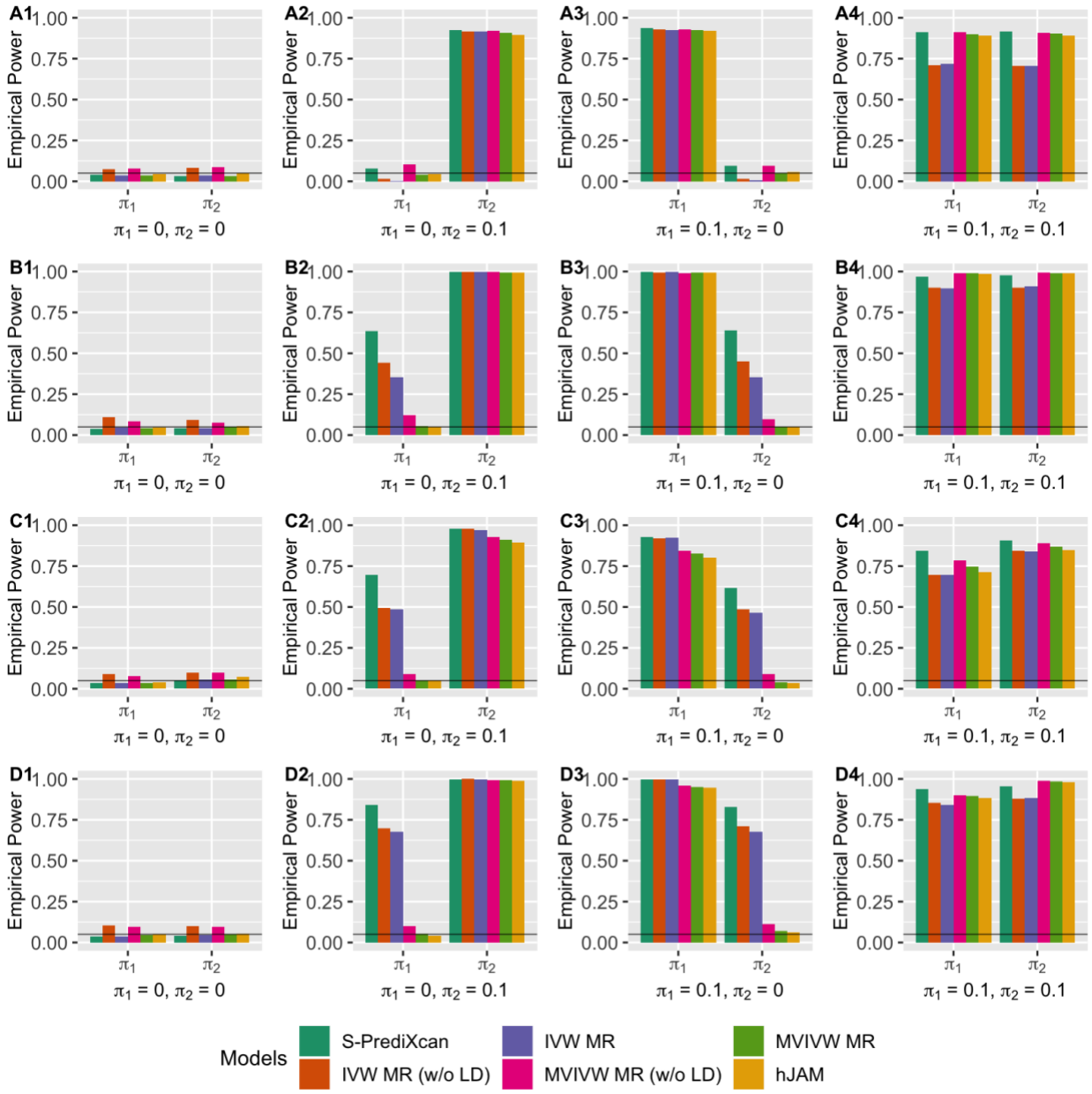

Figure 4 Empirical Power of the correlated SNPs scenarios across 1000 replications.

(A) X_1 and X_2 are independent. (B) X_1 and X_2 are correlated. (C) X_1 causes X_2. (D) X_1 causes X_2 and correlated. The black solid line refers to the default Type-I error, $\alpha = 0.05$.

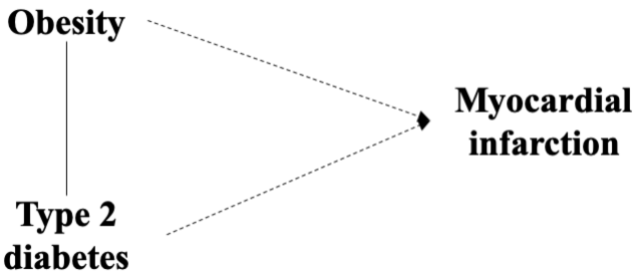

Figure 5 Direct acyclic graph (DAG) of the relationship between BMI, type 2 diabetes and myocardial infarction.