Reciprocal causation mixture model for mendelian randomization analysis

Sham P [1,2,3], Liu Z [1], Qin Y [1], Wu T [1,2], Mak T [1], Zhang Y [1,4], Li M [2,5,6]

[1] Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China
[2] Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China
[3] State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China
[4] Department of Statistics & Actuarial Science, Faculty of Science, The University of Hong Kong, Hong Kong, China
[5] Zhongshan School of Medicine, Center for Precision Medicine, Sun Yat-sen University, Guangzhou, Guangdong, China
[6] Key Laboratory of Tropical Disease Control (SYSU), Ministry of Education, Guangzhou, Guangdong, China

BACKGROUND

Mendelian randomization (MR) using GWAS summary data is a useful method for inferring causal relationships between risk factors and diseases. However, standard MR requires stringent assumptions which are often implausible due to the widespread pleiotropy of single nucleotide polymorphisms (SNPs) effects on multiple phenotypes. Recent methods has been developed to address this issue either by removing pleiotropic SNPs (e.g. MR-Egger [1] and MR-PRESSO [2] under the assumption of InSIDE (instrument strength independent of direct effect)) or explicitly modelling pleiotropic effects (e.g. MRMix [3] assuming a normal-mixture model to consider horizontal pleiotropic effect). But they still require the selection of valid and independent instrumental variables (IVs), which may lead to spurious inferences concerning causation if invalid IVs were used, as well as loss of information due to the exclusion of the majority of SNPs from GWAS summary data. Additionally, current MR methods require a separate analysis to examine the causal effect in the reverse direction. In this study, we propose a novel strategy to estimate the reciprocal causation between two phenotypes simultaneously using whole-genome scale GWAS summary data.

METHOD

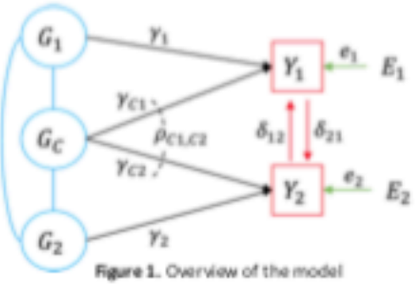

Figure 1. Overview of the model

We partition all available SNPs into mutually exclusive categories: trait-specific (π_1, π_2), pleiotropic (π_c) and null SNPs (π_0). In the context of a reciprocal causation, the joint linear model for a pair of phenotype Y_1 and Y_2 is $\boldsymbol{Y} = [\boldsymbol{I}-\boldsymbol{\Delta}]^{-1}\sum_{k=1}^{K}\boldsymbol{\gamma_k}G_k + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, $\boldsymbol{\gamma_k} = \begin{pmatrix} \gamma_{k1} \\ \gamma_{k2} \end{pmatrix}$, $\boldsymbol{\Delta} = \begin{pmatrix} 0 & \delta_{12} \\ \delta_{21} & 0 \end{pmatrix}$, γ_{k1} and γ_{k2} are direct effect sizes of the k-th SNP for Y_1 and Y_2, δ_{12} and δ_{21} are casual direction for $Y_2 \rightarrow Y_1$ and $Y_1 \rightarrow Y_2$ respectively (Figure 1).

Next, we define $\boldsymbol{\beta_k} = [\boldsymbol{I}-\boldsymbol{\Delta}]^{-1}\boldsymbol{\gamma_k}$ as a 2×1 vector of the joint effect size, following a bivariate mixture distribution in the form $\boldsymbol{\beta_k} \sim \sum \pi_h N(\boldsymbol{0}, \boldsymbol{\Sigma}_h) + \pi_0 N(\boldsymbol{0}, \boldsymbol{0})$, where π_0 is the mixing proportion of null SNPs, π_h and $\boldsymbol{\Sigma}_h$ are the mixing proportions and a 2×2 variance-covariance matrix of effect size respectively for the non-null SNPs belonging to the corresponding categories (i.e. $h = 1,2,c$). Then we could assume a bivariate normal distribution of the summary statistic for the k-th SNP ($\hat{\boldsymbol{\gamma}}_{\boldsymbol{k}}$ is the estimate of effect size in GWAS summary statistics):

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{k}} = \begin{pmatrix} \hat{\gamma}_{k1} \\ \hat{\gamma}_{k2} \end{pmatrix} \sim \sum_{\mathbb{N}_k} Pr_\ell(\mathbb{N}_k) N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} var(\hat{\gamma}_{k1}) + a_1 + 1/n_1 & cov(\hat{\gamma}_{k1}, \hat{\gamma}_{k2}) + \rho_0 \\ cov(\hat{\gamma}_{k1}, \hat{\gamma}_{k2}) + \rho_0 & var(\hat{\gamma}_{k2}) + a_2 + 1/n_2 \end{pmatrix}\right],$$

- $Pr_\ell(\mathbb{N}_k)$ can be calculated from the standard multinomial distribution with $\mathbb{N}_k = (N_k^{\ell 1}, N_k^{\ell 2}, N_k^{\ell C}, N_k^{\ell 0})$ and total counts $N_k^\ell = N_k^{\ell 1} + N_k^{\ell 2} + N_k^{\ell C} + N_k^{\ell 0}$.
- a_1 and a_2 are additional inflation factors accounting for systematic bias in variance estimates for phenotype Y_1 and Y_2 respectively.
- ρ_0 is a factor accounting for bias in the covariance estimates due to effects such as sample overlapping. n_1 and n_2 are the sample size for the two GWASs.
- n_1 and n_2 are sample sizes for GWAS Y_1 and Y_2 respectively.

Then, the composite log-likelihood (CL) function is in the form:

$$CL(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_{\boldsymbol{k}}) = \sum_{k=1}^{K} log\, L(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_{\boldsymbol{k}}) = \sum_{k=1}^{K} log\left[\sum_{\mathbb{N}_k} Pr_\ell(\mathbb{N}_k) f(\hat{\gamma}_{k1}, \hat{\gamma}_{k2})\right],$$

where $f(\hat{\gamma}_{k1}, \hat{\gamma}_{k2})$ is the density function of bivariate normal distribution.

Thus, the maximum composite likelihood estimator can be given by $\hat{\boldsymbol{\theta}} = argmax\, CL(\boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}_{\boldsymbol{k}})$. The reciprocal causal paths, together with nuisance parameters, are then estimated by an EM algorithm.

SIMULATION RESULTS

In simulations with correlated pleiotropy, our method (referred to as "JointModel") obtains well-controlled type I error rates ($\alpha = 0.05$) (Table 1) under the null hypothesis and adequate power under non-null hypothesis (Table 2). In terms of causal estimates, our method obtains nearly unbiased estimates of causation in both directions (Figure 2). After comparing with existing MR-based methods, we noticed that the estimates from most of the MR-based methods vary greatly when correlated pleiotropy exists while our method gives more accurate estimates (Figure 2).

Table 1. Type I error rate under null hypothesis in 100 simulations

Method	δ_{12}=0.0		δ_{21}=0.0	
	Mean $\hat{\gamma}^2$ (SD)	Type I error rate	Mean $\hat{\gamma}^2$ (SD)	Type I error rate
JointModel	1.11 (1.55)	0.06	0.91 (1.17)	0.08

Table 2. Power under non-null hypothesis in 100 simulations

Method	δ_{12}=0.1		δ_{21}=0.05	
	Mean $\hat{\gamma}^2$ (SD)	Power	Mean $\hat{\gamma}^2$ (SD)	Power
JointModel	76.52 (63.58)	0.97	18.9 (17.76)	0.85

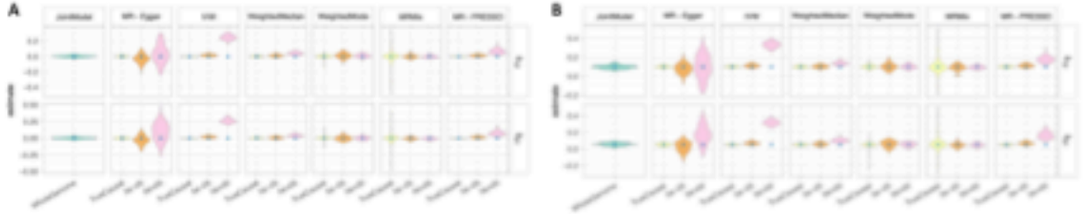

Figure 2. Comparison of estimates from different MR-based methods and JointModel. For 100 simulations, pleiotropic effects are correlated ($\rho_{C1,C2} = 0.1$). To investigate how the selection of IVs affects the MR estimation, we have tried: (1) using the true causal SNPs in the simulation as IVs; (2) using significant exposure-associated SNPs (P-val < 5×10^{-8}) which also satisfy the exclusion restriction. We set two levels of exclusion according to the SNP P-values associated with outcome (P-val > 5×10^{-3}/0.01). Next, the selected SNPs are clumped ($r^2 < 0.01$) to obtain the nearly-independent IVs for MR analysis. x-axis shows the IVs selection methods (JointModel considers whole-genome scale SNPs). The blue points are the true values of δ_{12}/δ_{21}. **(A)** Both true δ_{12} and δ_{21} were set as zero. **(B)** True causal effects were set as $\delta_{12} = 0.1$ and $\delta_{21} = 0.05$.

APPLICATION ON LDL-CAD

We applied the method to a pair of real GWAS phenotypes: low-density lipoprotein (LDL) [4] and coronary artery disease (CAD) [5]. Results show a significant causal effects from LDL to CAD (0.35, P-val=5.8×10^{-6}) but no significant causal effects from CAD to LDL (-0.11, P-val=0.10). Comparison between fitted GWAS and observed GWAS also suggest the estimates are close to the real situation (Figure 3).

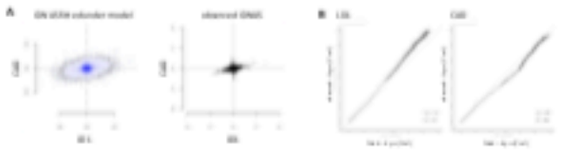

Figure 3. Comparison between fitted GWAS and observed GWAS. (A) Scatterplot for the fitted GWAS based on the estimated parameters (left) and the observed GWAS (right). **(B)** QQplot for LDL (left) and CAD (right). x-axis is the $-log_{10}P$ of fitted GWAS and y-axis is the $-log_{10}P$ of observed GWAS.

SUMMARY

In summary, we have developed a method for the causal inference among complex phenotypes. This method could simultaneously estimate reciprocal causal relationships between two phenotypes using GWAS summary statistics of all SNPs on the two phenotypes while accounting for LD correlations between SNPs. Simulations under various scenarios, including strong pleiotropy, show that the method gives nearly unbiased estimates of the reciprocal causal paths, and correct type I error rates under the null hypothesis. Using real GWAS summary data from LDL and CAD, we detected a significant causal path from LDL to CAD, and non-significant causation in the reverse direction.

REFERENCES

[1] Bowden J, *et al.* ***Int J Epidemiol***. 2015. [2] Verbanck M, *et al.* ***Nat Genet***. 2018. [3] Qi G, *et al.* ***Nat Commun***. 2019. [4] Willer C. J., et al. ***Nat Genet***. 2013. [5] Nelson C.P., et al. ***Nat Genet***. 2017