A Two-stage Linear Mixed Model (TS-LMM) for Summary-data-based Multivariable Mendelian Randomization

Ming Ding

Department of Emergency Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Correspondence to:

Ming Ding

Department of Emergency Medicine

University of North Carolina at Chapel Hill

Email: ming_ding@med.unc.edu

ABSTRACT

Multivariable Mendelian randomization (MVMR) methods provide a strategy for applying genome-wide summary statistics to assess simultaneous causal effects of multiple risk factors on a disease outcome. In contrast to univariate MR methods that assumes no horizonal pleiotropy (genetic variants only associate with one risk factor), MVMR allows for genetic variants associate with multiple risk factors and models pleiotropy by including summary statistics with risk factors as multiple variables into the regression model. Here, we propose a two-stage linear mixed model (TS-LMM) for MVMR that accounts for variance of summary statistics not only in outcome, but also in all of the risk factors. In stage I, we apply linear mixed model to treat variance in summary statistics of disease as fixed-/random-effects, while accounting for covariance between genetic variants due to linkage disequilibrium (LD). Particularly, we use an iteratively re-weighted least squares algorithm to obtain estimates for the random-effects. In stage II, we account for variance in summary statistics of multiple risk factors simultaneously by applying measurement error correction methods that take into consideration LD between genetic variants and correlation between summary statistics of risk factors. We compared our MVMR approach to other approaches in a simulation study. When most of the instrumental variables (IVs) were strong, our model generated the highest coverage of true causal associations, the highest power of detecting significant causal associations, and the lowest false positive rate of identifying null causal effect for a range of scenarios that varied correlation (weak, strong) between summary statistics of risk factors and LD among genetic variants (weak LD [$\gamma^2 \leq 0.1$], moderate LD [$0.1 < \gamma^2 \leq 0.5$]). When the proportion of strong IVs was reduced, our model showed performances comparable to MVMR-Egger and MVMR-IVW. The more accurate inference of our model in the presence of correlation among risk factors supports potential wide application to -omics data that are commonly multi-

dimensional and correlated, as shown in application to determinants of longevity, where our method nominated a specific significant lipoprotein subfraction for causal association from a panel of 10 lipoprotein cholesterol measures. The robustness of our model to correlation structure suggests that in practice we can allow moderate LD in selection of IVs, thereby potentially leveraging genome-wide summary data in a more effective manner. Our model is implemented in 'TS_LMM' macro in R.

1. INTRODUCTION

In recent years, genome-wide association studies (GWAS) with numerous phenotypes have been conducted, and the summary statistics generated provide us invaluable opportunity to assess causality between phenotypes and disease using Mendelian randomization (MR). A foundational concern about inference by MR is the potential for confounding by horizontal pleiotropy, where one variant may have independent effects on multiple risk factors. Univariate MR approaches that examine association of a single risk factor with an outcome typically acknowledge horizontal pleiotropy as a limitation or try to infer its presence based on distributional properties. For example, MR-IVW averages causal estimates across all variant instruments using inverse-variant weighted meta-analysis, while recognizing that pleiotropy may be manifest in heterogeneity.[1] Other methods are designed to be robust to some specific forms of pleiotropy (e.g. MR-Egger[2]), or try to detect and exclude instruments that are inferred to have potential pleiotropic influences (e.g. MR-median, [1] MR-PRESSO, [3] GSMR with HEIDI[4,5]). Still other methods focusing on univariate associations distinguish possible direct and indirect effects (e.g. MRMix[6]), or regression approach that accounts for random error in both risk factor and disease (e.g. MR-RAPS, [7] MR-MtRobin[8]), or estimate the proportion of instrumental effects that are causal as opposed to pleiotropic (e.g. LCV[9]), the latter class including Bayesian approaches to separate causal from confounding effects (e.g. CAUSE[10]). However, toward the goal of understanding mechanisms of disease, an intrinsic limitation of all univariate MR methods is that they do not explicitly model pleiotropy, even when arising from recognized association of genetic instrument variables (IVs) with multiple correlated risk factors, an increasingly important concern given technologies that allow high-dimensional deep phenotyping, e.g. metabolomics.

To account for recognized pleiotropy due to IVs associated with ≥1 risk factors, multivariable MR (MVMR) methods have been proposed regressing point estimates from summary statistics of disease on those of multiple risk factors, with coefficients indicating causal associations between risk factors and disease.[11,12] For example, leveraging point estimates, the MR-Egger, MR-IVW, median-based MR, and MR-PRESSO methods have been extended to multivariable case. However, as summary statistics generated from genome-wide association studies include both point estimates and variances, there is an opportunity to improve MVMR methods by including the latter. Therefore, we propose a two-stage linear mixed model (TS-LMM) for MVMR using both point estimates and variances available from summary-data: In stage I, by treating variation in summary statistics of disease in a mixed effects framework (i.e. fixed-/random-effects), we use a weighted linear mixed model to obtain estimates for fixed-effects and an iteratively re-weighted least squares algorithm to obtain estimates for random-effects. In stage II, we treat variation in summary statistics of risk factors as measurement error and apply measurement error correction to obtain an estimate of a causal association. As we take into consideration LD between genetic variants and correlation between summary statistics of risk factors in stage II, our method allows for correlated genetic variants as IVs and can be applied to correlated summary statistics of risk factors.

We compared our method to other existing MVMR methods in simulations and applied it to real data. In simulations, when most of the instrumental variables (IVs) were strong, our model generated the highest coverage of true causal associations, the highest power of detecting significant causal associations, and the lowest false positive rate of identifying null causal effect in most of the scenarios varying correlations (weak, strong) between summary statistics of risk factors and linkage disequilibrium (LD) of genetic variants (weak, moderate). When the

percentage of strong IVs was reduced, our model performed comparably to MVMR-Egger and MVMR-IVW. In an application, we examined causal associations between 10 correlated lipoprotein subfraction cholesterol biomarkers and longevity, demonstrating that our model can be applied to multi-dimensional and correlated -omics data.

2. TWO-STAGE MULTIVARIABLE LINEAR MIXED MODEL (TS-LMM)

2.1. Model overview (Figure 1).

As shown in **Figure 1a**, we use genetic variants G as IVs to examine causal association between risk factor X and disease Y, allowing for pleiotropy due association of the IVs with ≥1 risk factor or directly associated with disease Y. Suppose we include N genetic variants and M -omics biomarkers in our analysis. Let b_i be the summary statistics between genetic variant i and disease, with coefficient β_i and standard deviation (SD) σ_i^2. Let b_{mi} be the summary statistics between variant i and -omics marker m, with coefficient β_{mi} and SD σ_{mi}^2. Summary statistics, which are on the log scale for binary traits and identity scale for continuous traits, are normally distributed (**Figure 1b**). An illustration of the workflow of TS-LMM work is in **Figure 1c**. We fit a multivariable linear mixed model, where k_m represents the causal association between risk factor m and disease Y, and k_0, the intercept, is the average direct effect across all genetic variants on Y.

$$b_i = \sum_{m=1}^{M} k_m * b_{mi} + k_0 + \varepsilon_i,\ b_i \sim N(\beta_i, \sigma_i^2),\ b_{mi} \sim N(\beta_{mi}, \sigma_{mi}^2),\ \varepsilon_i \sim N\ (0, \delta^2) \quad \text{— model 1}$$

As b_i and b_{mi} belong to normal distributions, we need to account for randomness in summary statistics of risk factors and disease. Thus, we use a two-stage weighted regression model to estimate k_m. In stage I, we model β_i as outcome and treat its variance, σ_i^2, as random effect γ_i (model 2). This is similar to meta-regression used in meta-analysis, where γ_i can be treated as within study variation and ε_i can be treated as between study variation. Given that b_{mi} has a

normal distribution with mean value β_{mi} and variance σ_{mi}^2, this can be treated as random measurement error ρ_{mi}, as shown in model 2. In stage II, we correct for this error in risk factors simultaneously using multiple measurement error correction method.

$\beta_i = \sum_{m=1}^{M} k'_m * \beta_{mi} + k'_0 + \gamma_i + \varepsilon_i$, $b_{mi} = \beta_{mi} + \rho_{mi}$, $\rho_{mi} \sim N(0, \sigma_{mi}^2)$, $\gamma_i \sim N(0, \sigma_i^2)$,

$\varepsilon_i \sim N\ (0, \delta^2)$ — model 2

2.2. Stage I of TS-LMM.

In stage I, we use weighted linear mixed model to obtain estimates of k'_m, by treating variation in summary statistics of disease (σ_i^2) in a mixed effects framework (i.e. fixed-/random-effects).

2.2.1. Estimation of covariance matrix of disease ($\sum$). Given that γ_i and ε_i are independent, for genetic variant i, the variance of outcome $var(\beta_i)$=σ_i^2+δ^2. Covariance between genetic variants 1 and 2 can be calculated as cov(β_1, β_2) = corr(β_1, β_2)*$\sqrt{var(\beta_1)}\sqrt{var(\beta_2)}$

$\approx h_{12}\sqrt{(\sigma_1^2 + \delta^2)(\sigma_2^2 + \delta^2)}$. Corr($\beta_1$, β_2) can be approximated from squared root of linkage disequilibrium (LD) measured by Pearson correlation (h_{12}) between genetic variants using LDmatrix Tool in LDlink.[13]

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \delta^2 & h_{12}\sqrt{(\sigma_1^2 + \delta^2)(\sigma_2^2 + \delta^2)} & h_{13}\sqrt{(\sigma_1^2 + \delta^2)(\sigma_3^2 + \delta^2)} & \dots \\ \dots & \sigma_2^2 + \delta^2 & h_{23}\sqrt{(\sigma_2^2 + \delta^2)(\sigma_3^2 + \delta^2)} & \dots \\ \dots & \dots & \sigma_3^2 + \delta^2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

2.2.2. Model estimation. To determine whether fixed- or random- effects model should be used, we assess heterogeneity of summary statistics with disease across genetic variants by calculating Cochran's Q statistic=$(\boldsymbol{Y} - \boldsymbol{BK'})^{\boldsymbol{T}}\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{BK'})$ and $I^2 = \frac{\mathbf{100*(Q-n-1)}}{\mathbf{Q}}$, with I^2>50% indicating significant heterogeneity.[14] If no heterogeneity is found, model 2 is equivalent to a fixed-effects model with δ=0. We can obtain $\boldsymbol{K'}$ as $(\boldsymbol{B^T W B})^{\mathbf{-1}}\boldsymbol{B^T W Y}$, where **B** is a N×(M+1) vector of $(1, \beta_{1,\dots,}\beta_m)$, **Y** is a N×1 vector of β, $\mathbf{K'}$ is a (M+1)×1 vector of $(k'_0, k'_1, \dots, k'_m)$, and **W** is the

weight calculated as $\sum^{-1}$. Var($\boldsymbol{K'}$) is $\frac{(\boldsymbol{Y}-\boldsymbol{BK'})^{\boldsymbol{T}}\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{BK'})}{n-m-1} * (\boldsymbol{B^T W B})^{-1}$. We use Moore-Penrose pseudoinverse if $\boldsymbol{B^T W B}$ is singular. In the presence of heterogeneity in summary statistics with disease across genetic variants, we apply a random-effects model. Given that δ is unknown in a random-effects setting, we use an iteratively re-weighted least squares algorithm to estimate $\boldsymbol{K'}$. Initially, we assume $\delta^{2(0)}$=0 and use WLS to obtain an initial estimate of $\boldsymbol{K'}^{(1)}$. $\delta^{2(1)}$ can be estimated as $(\sum_{i=1}^{n} W_{ii}((\frac{n}{n-m-1})(\beta_i - \sum_1^m k'_m * \beta_{mi})^2 - \sigma_i^2))/\sum_{i=1}^{n} W_{ii}$,[15] where absolute value of $\delta^{2(1)}$ is used when the estimate is negative. We estimate δ^2 and $\boldsymbol{K'}$ iteratively until the estimate of δ^2 converges.

2.3. Stage II of TS-LMM.

In stage II, we treat variance in summary statistics of risk factors as random measurement error and apply measurement error correction to obtain causal association.[16-18] Specifically, we simulate an observed dataset belonging to a multivariate normal distribution, where the means and variances are the means and variances of summary statistics of risk factors and the covariance taking into consideration correlations between summary statistics of risk factors and LD between genetic variants. We perform multivariate regression analysis between simulated and observed summary statistics of risk factors to obtain regression coefficients, which are used to obtain corrected causal association. Given that we account for variance in summary statistics of multiple risk factors simultaneously, our method can be applied to correlated summary statistics of risk factors, such as correlated lipidomic data.

2.3.1. Generate a simulated dataset of summary statistics for N genetic variants with M risk factors. We randomly sample β'_{mi} from multivariate normal distributions, where $\beta'_{mi} \sim N(\beta_{mi}, \sigma_{mi}^2)$. Covariance between β_{mi} and β_{kj} is calculated as cov(β_{mi}, β_{kj}) = corr(β_{mi},

$\beta_{kj})*\sqrt{var(\beta_{mi})}\sqrt{var(\beta_{kj})} \approx h_{ij}r_{mk}\sqrt{\sigma_{mi}^2\sigma_{kj}^2}$, where h_{ij} is the Pearson correlation between genetic variants i and j, (a measurement of LD as aforementioned) and r_{mk} is the correlation between summary statistics of risk factors m and k.

2.3.2. Perform multivariate regression analysis to obtain coefficient between simulated and observed summary statistics of risk factors. We correct for measurement error of all $\boldsymbol{\beta}$ (or variance of summary statistics of all risk factors) simultaneously by performing a multivariate linear regression of $\boldsymbol{\beta'}$ on $\boldsymbol{\beta}$**,** where $\boldsymbol{\beta'}$ is $c(\beta_1',\ldots,\beta_m')$ and $\boldsymbol{\beta}$ is $c(\beta_1,\ldots,\beta_m)$. $\boldsymbol{\lambda}$ can be calculated as $(\boldsymbol{\beta^T\beta})^{-1}\boldsymbol{\beta^T\beta'}$ and var($\boldsymbol{\lambda}$) would be $(\boldsymbol{\beta^T\beta})^{-1} \otimes\sum$, where $\otimes$ denotes Kronecker product and $\sum$ is $\boldsymbol{\beta'}^{\boldsymbol{T}}(\boldsymbol{I_n}-\boldsymbol{H})\boldsymbol{\beta'}/(\boldsymbol{n}-\boldsymbol{p}-\boldsymbol{1})$ with $\boldsymbol{I_n}$ as identity matrix, $\boldsymbol{H}$ as hat matrix, and p as degree of freedom of regression model.

2.3.3. Apply coefficient correction to obtain causal association**.** The true coefficient $\boldsymbol{K'}=\boldsymbol{K\lambda}^{-1}$, and thus observed coefficient $\boldsymbol{K}$ would be $\boldsymbol{K'}*\boldsymbol{\lambda}$, where $\boldsymbol{K}$ is $c(k_{1,\ldots,}k_m)$ and $\boldsymbol{K'}$ is $c(k_{1\ ,\ldots,}'k_m')$. var$(\boldsymbol{K})_{p,q}$ can be calculated using the following formula:[19]

$$\mathrm{var}(\boldsymbol{K})_{p,q} = \sum_{i=1}^{m}\sum_{j=1}^{m}(cov_{\boldsymbol{K'}}(\boldsymbol{K'}_i,\boldsymbol{K'}_j)+E_{\boldsymbol{K'}}[\boldsymbol{K'}_i]E_{\boldsymbol{K'}}[\boldsymbol{K'}_j])(cov_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_{pi},\boldsymbol{\lambda}_{qj})+E_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}_{ip}]E_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}_{jq}])$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{m}E_{\boldsymbol{K'}}[\boldsymbol{K'}_i]E_{\boldsymbol{K'}}[\boldsymbol{K'}_j]E_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}_{ip}]E_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}_{jq}]$$

2.4. Software of TS-LMM.

We have developed our model into a macro named 'TS_LMM', which can be accessed publicly in Github (https://github.com/mingding-hsph/TS_LMM).

3. SIMULATION STUDY.

3.1. Simulate datasets.

We performed a simulation study to validate the utility of our model. For each dataset, we generated data on 20 genetic variants (G) from 1000 participants. Specifically, genetic variants were simulated from multivariate binary distributions with marginal probabilities of 0.1-0.5 representing genetic variants with minor allele frequency 0.1-0.5 and covariance (γ^2) representing LD between genetic variants. For participant i, we generated risk factor m (A_{mi}) and outcome Y from the following models.

$A_{mi}=\sum_1^k(ll_k + \mathrm{D}_{mk}) * (G_{ki}+\mathrm{W}_{mki}) + u_{mi} + \varepsilon_{mi}$, $\varepsilon_{mi} \sim N\ (0,1)$, $\mathrm{W}_{mki} \sim \mathrm{N}\ (0, \sigma^2)$, $\mathrm{D}_{mk} \sim \mathrm{N}\ (0, \xi^2)$

$Y_i=\sum_1^k l_k * (G_{ki} + \mathrm{W}_{ki}) + u_{yi} + \sum_1^m p_m * A_{mi} + \varepsilon_i$, $\varepsilon_i \sim N\ (0,1)$, $\mathrm{W}_{ki} \sim \mathrm{N}\ (0, \sigma^2)$

To generate unmeasured confounders between risk factors and outcome, we simulated $u_1 - u_m$ and u_y from multivariate normal distributions, with mean values 1, variance 1, and moderate to high covariance (0.5-0.7). l_k and ll_k are coefficients of genetic variant k with disease Y and risk factors, respectively, which were simulated from uniform distributions [-1, 1]. D_{mk} is a random variable that is used to regulate correlation between summary statistics of risk factors. Simulating ll_k and D_{mk} allows for pleiotropy due to genetic variants associated with ≥1 risk factors, and simulating l_k allows for pleiotropy due to genetic variants directly associated with Y. p_m is the causal effect of risk factor A_m on risk of Y. W is a random variable that is used to regulate association between genetic variant and risk factor.

To test the validity of our model in correctly estimating causal effects, we simulated three risk factors, A_1, A_2, A_3, with positive (p_1=1), null (p_2=0), and negative (p_3=-1) effects on Y. To test the robustness of our model, we simulated scenarios with high (σ=0) and low (σ=1) proportions of strong IVs; scenarios with low (γ^2 between 0-0.1) and moderate LD (γ^2 between 0.1-0.5) between genetic variants by changing covariance of multivariate binary distributions; scenarios with weak (ξ^2=1) and strong (ξ^2=0.1) correlations between summary statistics of risk

factors. To show that our model can be applied to the univariate MR case, we simulated only one risk factor (A) each time and conducted simulations multiple times with A=1, A=0, and A=-1.

We simulated 1000 datasets for each scenario. We examined the causal effects using TS-LMM, and compared our model to MRMV-IVW, MRMV-Egger, MRMV-median method, and MR-PRESSO.[3,20] We calculated the average of estimated causal effects and its standard error, coverage, power, and false-positive rate over the 1000 datasets. Coverage was defined as the proportion of 95% CIs including the true causal effect over all simulated datasets. Power was estimated by the proportion of 95% CIs excluding zero in settings with significant causal effect. False-positive rate was defined as the proportion of significant causal estimates in settings with no causal effect.

3.2. Simulation results.

In simulation scenarios aiming for high proportion of strong IVs, 60%-70% genetic variants were strong IVs (i.e., F-statistics > 10) in the majority of the 1000 simulated datasets (**Figure S1**), while in scenarios of low proportion of strong IVs, only 20%-30% genetic variants were strong IVs in most of the simulated datasets (**Figure S2**), showing that the simulation of IVs with different strength is successful. As shown in **Table S1,** Pearson correlation between summary statistics of risk factors was ~0.1 in scenarios with weak correlation between risk factors and ~0.9 in scenarios with strong correlation, indicating that the simulation of summary statistics of risk factors with weak and strong correlations was successful. The correlation matrix between low and moderate LD are shown in **Figure S3**, showing that the simulation of low and moderate LD is successful.

When most of the IVs were strong, the estimated causal associations obtained using our method are the closest to true values in comparation to other methods for positive, null, or negative causal effects in most of the scenarios with different correlations (weak, strong)

between summary statistics of risk factors and different (weak and moderate) LD between genetic variants (**Table 1**). Moreover, our model generated the highest coverage of true causal association, the highest power of detecting significant causal association, and the lowest false positive rate of identifying null causal effect in most of the scenarios. When the proportion of strong IVs was reduced, our model showed performances comparable to MVMR-Egger and MVMR-IVW (**Table 2**). We further applied our model to univariable MR, and as expected, our model showed high coverage, high power, and low false positive rate under different strength of IV and different LD between IVs, demonstrating the generalizability of our model to univariable case (**Table 3**). As to the performance of other models, MR-PRESSO showed very good performances with low LD between genetic variants, however, with moderate LD between genetic variants, coverage and power decreased greatly and false positive rate increased drastically using MR-PRESSO (Tables 1-3).

4. APPLICATION STUDY TO LIPOPROTEIN CHOLESTEROL BIOMARKERS AND LONGEVITY.

We applied our model to examine causal associations of blood lipoprotein cholesterol biomarkers with longevity for two reasons. First, state-of-the-art technologies have allowed for high-throughput profiling of -omics markers, and these emerging data have been powerful tools to understand complex biological systems.[21] We utilized blood lipoprotein data to show the potential wide application of our model to -omics data. Second, genome-wide summary statistics with lipoprotein cholesterol biomarkers and longevity are open resources, which provides an excellent opportunity to investigate this question.

4.1. Methods of application study.

Summary statistics with longevity was obtained from a longevity genomics consortium,[22] for which longevity was defined as age >90th survival percentile. The summary statistics were generated from 11,262 cases surviving at or beyond this age threshold and 25,483 controls whose age at death or at last contact was at or below the age corresponding to the 60th survival percentile. Summary statistics with lipoprotein cholesterol biomarkers were from the MAGNETIC consortium, which were generated from GWAS with 24,925 participants (https://www.ebi.ac.uk/gwas/publications/27005778).[23]

The lipoprotein cholesterol biomarkers were measured using nuclear magnetic resonance (NMR) metabolomics platform (Nightingale Health Ltd., Helsinki, Finland), and 10 cholesterol biomarkers were considered, namely, cholesterol in very large high-density lipoprotein (HDL) (XL.HDL.C), cholesterol in large HDL (L.HDL.C), cholesterol in medium HDL (M.HDL.C), cholesterol in large very low-density lipoprotein (VLDL) (L.VLDL.C), cholesterol in medium VLDL (M.VLDL.C), cholesterol in small VLDL (S.VLDL.C), cholesterol in large low-density lipoprotein (LDL) (L.LDL.C), cholesterol in medium LDL (M.LDL.C), cholesterol in small LDL (S.LDL.C), cholesterol in intermediate-density lipoprotein (IDL) (IDL.C).

To evaluate the performance of our model, we selected genetic variants that were strongly associated with at least one of the lipoprotein cholesterol biomarkers as IVs. We first used the conventional criteria by including IVs with P value $<5*10^{-8}$ and then extended to P value$<5*10^{-5}$, which is still considered as strong IV.[24] These genetic variants were first pruned to moderate LD ($\gamma^2 \leq 0.5$) and then to low LD ($\gamma^2 \leq 0.1$) using SNPclip tool in LDlink.[13] To implement TS-LMM, correlation among selected genetic variants was calculated as the r^2 measure of LD using LDmatrix tool in LDlink,[13] and the correlation between summary statistics with lipoprotein cholesterol biomarkers was obtained by calculating pairwise Pearson correlation. We applied TS-LMM to

examine causal associations between cholesterol biomarkers and longevity and compared the findings to those obtained using MVMR-IVW, MVMR-Egger, MVMR-median method, and MR-PRESSO.

4.2. Results of application study.

Our study included 47 and 211 genetic variants with low ($\gamma^2 \leq 0.1$) and moderate LD ($\gamma^2 \leq 0.5$) at the significance level of $5*10^{-8}$, respectively. We did not identify cholesterol biomarkers significantly causally associated with longevity, either using TS-LMM or MR-PRESSO. By extending the significance level to $5*10^{-5}$, we included 739 and 1213 genetic variants with low and moderate LD, respectively. Using the 739 genetic variants as IVs, TS-LMM, as well as MR-PRESSO, identified M.HDL.C significantly positively associated with longevity (**Table 4**). By further including genetic variants with moderate LD as IVs, TS-LMM additionally identified L.VLDL.C and L.LDL.C inversely causally associated with longevity. However, for MR-PRESSO that does not take LD between genetic variants into consideration (assumes independence between genetic variance), narrower 95% CI were generated and more significant associations were identified by including IVs with moderate LD, with S.VLDL.C and IDL.C positively associated with longevity and L.LDL.C inversely associated. As MVMR-IVW, MVMR-Egger, and MVMR-median methods were inestimable due to singular matrix, we applied the models by assuming genetic variants were independent and presented the results in **Table S2**. The MVMR-median method generated far more significant associations than the other methods, even using genetic variants with low LD as IVs.

5. DISCUSSION

In this study, we propose a summary-data-based MVMR model that accounts for variance in summary statistics of risk factors and disease outcome. Our model can be used to estimate causal associations of multiple risk factors with disease and is particularly suitable to -omics data that are multi-dimensional and correlated. Also, the robustness of our model to correlated genetic variants allows for inclusion of genetic variants with higher LD as IVs, which is a more efficient utilization of genome-wide summary data. We have developed our method into R macros that can be accessed publicly by the research community.

Given the variance in summary statistics with not only disease but also risk factors, how to account for them is an importance and challenging question for MVMR. If risk factors are independent, it is possible to apply univariate methods for each risk factor to obtain causal estimates, such as using MR-RAPS[7] and MR-MtRobin[8] that account for random error in both risk factor and disease. However, if risk factors are correlated, it is necessary to account for the variance in summary statistics of risk factors simultaneously to generate less biased estimates. The key innovation of our model is the account for variance in summary statistics of multiple risk factors simultaneously using measurement error correction method in stage II. Specifically, we simulated summary statistics of risk factors from multivariate normal distribution that accounted for correlation between genetic variants and correlation between summary statistics of risk factors, obtained regression coefficients between observed and simulated summary statistics of risk factors, and applied multiple measurement error correction to obtain corrected estimates of causal association. Our model generates causal estimates of high validity, as indicated by the high coverage, high power, and low false positive rate in simulation study under weak and strong correlations between risk factors. The robustness to correlated summary statistics of risk factors shows that our model is highly suitable to the emerging -omics data that are multi-dimensional and correlated.[21,25]

Another advantage of our model is the robustness to correlation of IVs, as demonstrated by the high coverage, high power, and low false positive rate in simulation study under weak and moderate LD between genetic variants. This is due to that we carefully take into consideration correlation between genetic variants when accounting for variance in summary statistics of disease in stages I and II. In stage I, in the calculation of the covariance matrix of disease ($\sum$), variance is from summary statistics of disease, and covariance accounts for correlation between genetic variants. This is important as the inverse of the covariance matrix is used as weight in calculation of the weighted Cochran's Q statistic to decide whether to use fixed- or random-effects model,[14] and this weight is also used to obtain causal estimates under fixed-effects model. Considering the correlation between genetic variants, we adopted an iteratively re-weighted least squares algorithm to obtain estimates for random-effects.[15] In stage II, we simulate a multivariate normally distributed dataset based on summary statistics of risk factors, with covariance structure taking into consideration LD between genetic variants and correlation between summary statistics of risk factors.

Given the robustness of our model to correlated IVs, we propose to extend the LD from low to moderate in selection of IVs using summary data. For now, MR studies tend to choose IVs with low LD, which does not fully utilize information from genome-wide summary statistics. Our model allows to extend to moderate LD and include more genetic variants as IVs, which boosts power of MVMR and utilizes genome-wide information in a more effective manner. This is evident from the results of the application study: We only identified one biomarker using genetic variants with low LD, and identified a few more when further include variants with moderate LD. However, by including IVs with moderate LD, we need to choose MRMVs that account for LD such as our model. The reason is that models such as MR-PRESSO does not

account for LD and tends to generate more false positive results, as shown in the simulation study and the application study.

In the application study, our model identified cholesterol markers causally associated with longevity. M.HDL.C. is identified as a significant candidate causal protective factor of longevity. As a major vehicle for lipid transport, HDL takes up cholesterols from peripheral tissues through reverse cholesterol transport and deliver it to the liver for elimination. HDL is a collection of subtype particles with different sizes.[26] The particle with the smallest volume is nascent HDL secreted by the liver and small intestine, and it has the highest capacity of taking cholesterols from peripheral tissues. As cholesterols are taken and esterified, HDL particles increase in size, with larger size composing higher cholesterol.[27] Thus, HDL with smaller size exert stronger cholesterol efflux capacity from peripheral tissues and is more effective in offloading cholesterol.[28] This may explain why M.HDL.C., rather than L.HDL.C or XL.HDL.C, was identified as a protective factor of longevity. Our model also identified L.VLDL.C and L.LDL.C inversely associated with longevity. VLDL is synthesized in the liver and carries triglycerides to muscle and adipose tissue for metabolism and storage.[29] When triglycerides carried in VLDL are absorbed by peripheral tissues, they become smaller and denser and IDL are formed. The IDL are further metabolized to LDL when triglycerides are removed from VLDL. LDL contains a high proportion of cholesterol, which are taken up by the LDL receptor in numerous tissues including the liver.[30] Therefore, L.VLDL.C, an indicator of high plasma triglycerides, and LDL.C, an indicator of high plasma cholesterol, are risk factors of longevity.

In summary, we have developed a TS-LMM method for summary-data-based MVMR. Our model accounts for variance in summary statistics of correlated risk factors and disease, and potentially have wide application to -omics data that are multi-dimensional and correlated. Also,

given the robustness of our model to correlated genetic variants, we propose to extend the LD in selection of IVs to leverage genome-wide summary data in a more effective manner. We have developed a 'TS_LMM' macro in R that is convenient to implement and open to the whole research community.

ACKNOWLEDGMENTS

This work was supported by grant R21 HG012365 from the National Institutes of Health. Dr. Daniel I. Chasman (Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA) provided highly constructive comments to the manuscript. The author is very grateful to Dr. Chasman's valuable comments and suggestions on the manuscript.

DATA AVAILABILITY STATEMENT

'TS_LMM' R macro and the code used for the simulations and the application are available at the Github repository (https://github.com/mingding-hsph/TS_LMM). Summary statistics with longevity is openly available from Longevity Genomics website (https://www.longevitygenomics.org/downloads), and summary statistics with lipoprotein cholesterol biomarkers is also openly available from GWAS Catalog (https://www.ebi.ac.uk/gwas/publications/27005778).

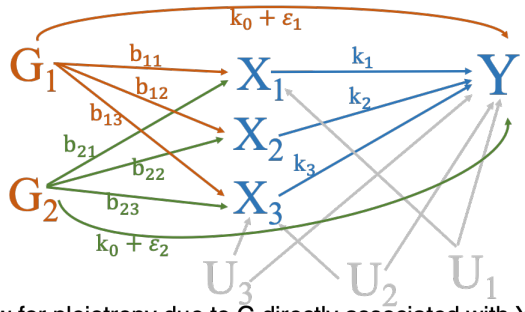
Reference

1. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol 2016;40(4):304-14. DOI: 10.1002/gepi.21965.
2. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32(5):377-389. DOI: 10.1007/s10654-017-0255-x.
3. 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. Nat Genet 2018;50(5):693-698. DOI: 10.1038/s41588-018-0099-7.
4. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun 2018;9(1):224. DOI: 10.1038/s41467-017-02317-2.
5. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481-7. DOI: 10.1038/ng.3538.
6. Qi G, Chatterjee N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. Nat Commun 2019;10(1):1941. DOI: 10.1038/s41467-019-09432-2.
7. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv:180109652 2018.
8. Gleason KJ, Yang F, Chen LS. A robust two-sample transcriptome-wide Mendelian randomization method integrating GWAS with multi-tissue eQTL summary statistics. Genet Epidemiol 2021. DOI: 10.1002/gepi.22380.
9. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nat Genet 2018;50(12):1728-1734. DOI: 10.1038/s41588-018-0255-0.
10. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat Genet 2020. DOI: 10.1038/s41588-020-0631-4.
11. Burgess S, Thompson DJ, Rees JMB, Day FR, Perry JR, Ong KK. Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. Genetics 2017;207(2):481-487. DOI: 10.1534/genetics.117.300191.
12. Porcu E, Rueger S, Lepik K, et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun 2019;10(1):3300. DOI: 10.1038/s41467-019-10936-0.
13. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31(21):3555-7. DOI: 10.1093/bioinformatics/btv402.
14. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-58. DOI: 10.1002/sim.1186.

15. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Stat Med 1995;14(4):395-411. DOI: 10.1002/sim.4780140406.
16. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990;132(4):734-45. DOI: 10.1093/oxfordjournals.aje.a115715.
17. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. Annu Rev Public Health 1993;14:69-93. DOI: 10.1146/annurev.pu.14.050193.000441.
18. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014;33(12):2137-55. DOI: 10.1002/sim.6095.
19. Proof Verification: Joint variance of the product of a random matrix with a random vector. https://stats.stackexchange.com/questions/420032/proof-verification-joint-variance-of-the-product-of-a-random-matrix-with-a-rand. Accessed May, 2021.
20. Broadbent JR, Foley CN, Grant AJ, Mason AM, Staley JR, Burgess S. MendelianRandomization v0.5.0: updates to an R package for performing Mendelian randomization analyses using summarized data. Wellcome Open Res 2020;5:252. DOI: 10.12688/wellcomeopenres.16374.2.
21. Joshi A, Rienks M, Theofilatos K, Mayr M. Systems biology in cardiovascular disease: a multiomics approach. Nat Rev Cardiol 2021;18(5):313-330. DOI: 10.1038/s41569-020-00477-1.
22. Deelen J, Evans DS, Arking DE, et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. Nat Commun 2019;10(1):3669. DOI: 10.1038/s41467-019-11558-2.
23. Kettunen J, Demirkan A, Wurtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun 2016;7:11122. DOI: 10.1038/ncomms11122.
24. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37(7):658-65. DOI: 10.1002/gepi.21758.
25. Ding M, Zeleznik OA, Guasch-Ferre M, et al. Metabolome-Wide Association Study of the Relationship Between Habitual Physical Activity and Plasma Metabolite Levels. Am J Epidemiol 2019;188(11):1932-1943. DOI: 10.1093/aje/kwz171.
26. Freedman DS, Otvos JD, Jeyarajah EJ, Barboriak JJ, Anderson AJ, Walker JA. Relation of lipoprotein subclasses as measured by proton nuclear magnetic resonance spectroscopy to coronary artery disease. Arterioscler Thromb Vasc Biol 1998;18(7):1046-53. DOI: 10.1161/01.atv.18.7.1046.
27. Khera AV, Cuchel M, de la Llera-Moya M, et al. Cholesterol efflux capacity, high-density lipoprotein function, and atherosclerosis. N Engl J Med 2011;364(2):127-35. DOI: 10.1056/NEJMoa1001689.
28. Ding M, Rexrode KM. A Review of Lipidomics of Cardiovascular Disease Highlights the Importance of Isolating Lipoproteins. Metabolites 2020;10(4). DOI: 10.3390/metabo10040163.
29. Tiwari S, Siddiqi SA. Intracellular trafficking and secretion of VLDL. Arterioscler Thromb Vasc Biol 2012;32(5):1079-86. DOI: 10.1161/ATVBAHA.111.241471.
30. Wadhera RK, Steen DL, Khan I, Giugliano RP, Foody JM. A review of low-density lipoprotein cholesterol, treatment strategies, and its impact on cardiovascular disease morbidity and mortality. J Clin Lipidol 2016;10(3):472-89. DOI: 10.1016/j.jacl.2015.11.010.

Figure 1. An illustration of the workflow of two-stage linear mixed model (TS-LMM).

a. Causal directed acyclic graph (DAG) between genetic variants (Z), risk factor (X), and disease Y.



We allow for pleiotropy due to G directly associated with Y and pleiotropy due to G associated with ≥1 risk factor.

b. Data structure of existing GWAS summary statistics with risk factors and disease

	Summary statistics with risk factors				Summary statistics with disease
	$X_1\,(b_{i1})$	$X_2\,(b_{i2})$	$X_3\,(b_{i3})$	$X_m\,(b_{im})$	$Y\,(b_i)$
Genetic variants G_1	$(\beta_{11}, \sigma^2_{11})$	$(\beta_{12}, \sigma^2_{12})$	$(\beta_{13}, \sigma^2_{13})$	$(\beta_{1m}, \sigma^2_{1m})$	(β_1, σ^2_1)
G_2	$(\beta_{21}, \sigma^2_{21})$	$(\beta_{22}, \sigma^2_{22})$	$(\beta_{23}, \sigma^2_{23})$	$(\beta_{2m}, \sigma^2_{2m})$	(β_2, σ^2_2)
G_i	$(\beta_{i1}, \sigma^2_{i1})$	$(\beta_{i2}, \sigma^2_{i2})$	$(\beta_{i3}, \sigma^2_{i3})$	$(\beta_{im}, \sigma^2_{im})$	(β_3, σ^2_3)

c. Two-stage linear mixed model (TS-LMM)

Model: $b_i = \sum_1^m k_m * b_{mi} + k_0 + \varepsilon_i$

$b_i \rightarrow N(\beta_i, \sigma_i^2)$; $b_{mi} \rightarrow N(\beta_{mi}, \sigma_{mi}^2)$; k_m → Causal association between risk factor X_m and Y

Stage I: Apply linear mixed model to account for variance in summary statistics of disease (b_i) and the covariance due to linkage disequilibrium (LD) of genetic variants.

Access heterogeneity of b_i using weighted Cochran's Q statistic and I^2 →

- No heterogeneity ($I^2 \leq 50\%$)

Fixed-effects model: $K' = (\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{Y}$

- Significant heterogeneity ($I^2 > 50\%$)

Random-effects model: Use an iteratively re-weighted least squares algorithm to estimate K'

Stage II. Treat variance in summary statistics of risk factors b_{mi} as measurement error and correct for multiple risk factors simultaneously.

- Generate a simulated dataset of summary statistics of risk factors: $\beta'_{mi} \sim N(\beta_{mi}, \sigma^2_{mi})$
- Perform a multivariate linear regression to obtain coefficient $\boldsymbol{\lambda}$: $\boldsymbol{\beta}' \sim \boldsymbol{\lambda}\boldsymbol{\beta}$, where $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}$ are simulated and observed risk factors.
- Apply coefficient correction to obtain causal association: $\boldsymbol{K}' = \boldsymbol{K}\boldsymbol{\lambda}^{-1}$.

Table 1. A simulation study estimating causal associations between positive (A_1), null (A_2), and negative (A_3) risk factors and outcome (Y) under different correlation patterns among risk factors and different linkage disequilibrium (LD) between genetic variants when most of the instrumental variables (IVs) are strong.

	Low LD ($r^2 \leq 0.1$)						Moderate LD ($0.1 < r^2 \leq 0.5$)					
	Weak correlation between risk factors			Strong correlation between risk factors			Weak correlation between risk factors			Strong correlation between risk factors		
	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1
Estimated p_1, p_2, p_3 (mean value over 1000 simulated datasets)												
TS-LMM	1.08	0.08	-0.92	1.33	0.31	-0.68	1.08	0.1	-0.94	1.32	0.33	-0.69
MVMR-Egger	1.13	0.09	-0.95	1.35	0.32	-0.69	1.14	0.11	-0.99	1.34	0.33	-0.69
MVMR IVW	1.13	0.08	-0.95	1.35	0.31	-0.69	1.14	0.11	-0.99	1.35	0.34	-0.69
MVMR median-based	1.14	0.08	-0.96	1.34	0.32	-0.67	1.15	0.13	-0.97	1.34	0.33	-0.69
MR-PRESSO	1.13	0.08	-0.94	1.34	0.32	-0.67	1.13	0.12	-0.96	1.34	0.33	-0.69
Estimated standard error of p_1, p_2, p_3 (mean value over 1000 simulated datasets)												
TS-LMM	0.26	0.25	0.25	0.26	0.25	0.26	0.26	0.26	0.25	0.26	0.26	0.26
MVMR-Egger	0.40	0.25	0.25	0.26	0.24	0.23	0.28	0.26	0.25	0.24	0.23	0.24
MVMR IVW	0.24	0.24	0.24	0.23	0.23	0.23	0.25	0.26	0.25	0.23	0.23	0.23
MVMR median-based	0.16	0.16	0.16	0.25	0.26	0.25	0.16	0.17	0.16	0.24	0.25	0.25
MR-PRESSO	0.24	0.24	0.24	0.23	0.23	0.23	0.25	0.25	0.25	0.23	0.23	0.23
Coverage[#] (%)												
TS-LMM	91	91	90	62	62	61	83	82	85	69	64	62
MVMR-Egger	90	89	89	58	58	61	80	77	80	63	61	65
MVMR IVW	87	89	90	54	56	59	74	75	78	60	59	63
MVMR median-based	56	58	63	67	67	65	38	42	44	55	56	58
MR-PRESSO	84	86	86	63	64	64	61	61	65	54	52	54
Power[&] (%)												
TS-LMM	96	NA	92	97	NA	69	94	NA	88	98	NA	73
MVMR-Egger	78	NA	93	94	NA	74	91	NA	87	99	NA	77
MVMR IVW	98	NA	94	98	NA	75	94	NA	87	99	NA	78
MVMR median-based	98	NA	97	99	NA	72	94	NA	92	95	NA	70
MR-PRESSO	95	NA	91	99	NA	73	86	NA	81	94	NA	70

False positive rate*(%)												
TS-LMM	NA	9	NA	NA	38	NA	NA	18	NA	NA	36	NA
MVMR-Egger	NA	11	NA	NA	43	NA	NA	23	NA	NA	39	NA
MVMR IVW	NA	11	NA	NA	44	NA	NA	25	NA	NA	41	NA
MVMR median-based	NA	42	NA	NA	33	NA	NA	58	NA	NA	45	NA
MR-PRESSO	NA	10	NA	NA	32	NA	NA	36	NA	NA	45	NA

p_1, p_2, p_3 are the true causal associations of A_1, A_2, A_3 with Y, respectively.
#Coverage is defined as the proportion of 95% CIs including the true causal effect over all simulated datasets.
&Power is estimated by the proportion of 95% CIs excluding zero in settings with significant causal effect.
*False positive rate is defined as the proportion of significant causal estimates in settings with no causal effect.

Table 2. A simulation study estimating causal associations between positive (A_1), null (A_2), and negative (A_3) risk factors and outcome (Y) under different correlation patterns among risk factors and different linkage disequilibrium (LD) between genetic variants when most of the instrumental variables (IVs) are weak.

	Low LD ($r^2 \leq 0.1$)						Moderate LD ($0.1 < r^2 \leq 0.5$)					
	Weak correlation between risk factors			Strong correlation between risk factors			Weak correlation between risk factors			Strong correlation between risk factors		
	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1	p_1=1	p_2=0	p_3= -1
Estimated p_1, p_2, p_3 (mean value over 1000 simulated datasets)												
TS-LMM	1.1	0.07	-0.92	1.31	0.32	-0.69	1.08	0.1	-0.93	1.29	0.32	-0.71
MVMR-Egger	1.08	0.07	-0.94	1.32	0.32	-0.69	1.08	0.1	-0.94	1.31	0.32	-0.7
MVMR IVW	1.09	0.06	-0.93	1.32	0.32	-0.69	1.08	0.09	-0.94	1.31	0.32	-0.7
MVMR median-based	1.09	0.06	-0.94	1.31	0.33	-0.69	1.09	0.1	-0.94	1.31	0.33	-0.69
MR-PRESSO	1.09	0.06	-0.93	1.32	0.32	-0.7	1.08	0.1	-0.94	1.31	0.33	-0.69
Estimated standard error of p_1, p_2, p_3 (mean value over 1000 simulated datasets)												
TS-LMM	0.28	0.28	0.27	0.3	0.3	0.3	0.27	0.27	0.27	0.3	0.29	0.3
MVMR-Egger	0.4	0.25	0.25	0.34	0.28	0.29	0.28	0.25	0.25	0.29	0.29	0.29
MVMR IVW	0.24	0.24	0.24	0.28	0.28	0.28	0.24	0.24	0.24	0.28	0.28	0.28
MVMR median-based	0.28	0.28	0.28	0.37	0.37	0.37	0.28	0.28	0.28	0.36	0.37	0.37
MR-PRESSO	0.24	0.24	0.24	0.23	0.23	0.23	0.24	0.24	0.24	0.23	0.23	0.23
Coverage[#] (%)												
TS-LMM	91	90	93	79	73	72	84	87	86	74	70	69
MVMR-Egger	93	90	92	86	83	82	87	81	84	84	78	84
MVMR IVW	90	90	91	81	81	82	82	80	84	81	78	82
MVMR median-based	88	88	88	91	90	88	71	71	72	81	78	81
MR-PRESSO	86	88	89	68	68	70	67	66	66	58	56	59
Power[&] (%)												
TS-LMM	96	NA	91	98	NA	62	93	NA	89	96	NA	63
MVMR-Egger	76	NA	92	97	NA	72	92	NA	90	99	NA	71
MVMR IVW	97	NA	93	100	NA	75	94	NA	91	100	NA	73
MVMR median-based	94	NA	87	96	NA	50	87	NA	81	90	NA	49
MR-PRESSO	95	NA	90	100	NA	78	87	NA	83	96	NA	72

False positive rate*(%)												
TS-LMM	NA	10	NA	NA	27	NA	NA	13	NA	NA	30	NA
MVMR-Egger	NA	10	NA	NA	17	NA	NA	19	NA	NA	22	NA
MVMR IVW	NA	10	NA	NA	19	NA	NA	20	NA	NA	22	NA
MVMR median-based	NA	12	NA	NA	10	NA	NA	30	NA	NA	22	NA
MR-PRESSO	NA	10	NA	NA	28	NA	NA	32	NA	NA	42	NA

p_1, p_2, p_3 are the true causal associations of A_1, A_2, A_3 with Y, respectively.
#Coverage is defined as the proportion of 95% CIs including the true causal effect over all simulated datasets.
&Power is estimated by the proportion of 95% CIs excluding zero in settings with significant causal effect.
*False positive rate is defined as the proportion of significant causal estimates in settings with no causal effect.

Table 3. A simulation study estimating causal association between univariate risk factor (A) and outcome (Y) with positive, null, or negative effect on disease, different linkage disequilibrium (LD) between genetic variants, and different strength of instrumental variables (IVs).

	Most IVs strong, low LD			Most IVs strong, moderate LD			Most IVs weak, low LD			Most IVs weak, moderate LD		
	A=1	A=0	A= -1	A=1	A=0	A= -1	A=1	A=0	A= -1	A=1	A=0	A= -1
Estimated mean value of risk factor A (mean value over 1000 simulated datasets)												
TS-LMM	1.00	-0.02	-0.99	1.01	0.00	-1.01	1.01	0.00	-1.01	1.02	-0.01	-1.01
MVMR-Egger	1.16	-0.02	-1.10	1.16	-0.01	-1.15	1.03	0.00	-1.02	1.03	0.00	-1.02
MVMR IVW	1.11	-0.02	-1.09	1.17	-0.01	-1.17	1.02	0.00	-1.02	1.03	-0.01	-1.03
MR-PRESSO	1.11	-0.01	-1.08	1.09	-0.02	-1.09	1.02	0.00	-1.02	1.03	0.02	-1.02
Estimated standard error of risk factor A (mean value over 1000 simulated datasets)												
TS-LMM	0.23	0.22	0.23	0.23	0.22	0.23	0.25	0.23	0.25	0.24	0.23	0.24
MVMR-Egger	0.44	0.43	0.43	0.27	0.27	0.26	0.42	0.42	0.41	0.27	0.27	0.27
MVMR IVW	0.22	0.22	0.22	0.22	0.23	0.22	0.22	0.22	0.22	0.22	0.22	0.22
MR-PRESSO	0.22	0.22	0.22	0.24	0.23	0.22	0.22	0.22	0.22	0.23	0.22	0.22
Coverage[#] (%)												
TS-LMM	84	81	83	77	76	76	87	86	86	82	77	79
MVMR-Egger	87	84	85	71	68	70	88	88	88	82	79	80
MVMR IVW	82	78	81	63	60	61	84	83	82	74	72	73
MR-PRESSO	79	74	78	44	39	42	81	81	78	43	43	44
Power[&] (%)												
TS-LMM	98	NA	98	95	NA	94	98	NA	97	95	NA	94
MVMR-Egger	72	NA	67	91	NA	90	64	NA	65	89	NA	87
MVMR IVW	100	NA	99	95	NA	95	98	NA	98	95	NA	93
MR-PRESSO	97	NA	96	80	NA	83	97	NA	95	82	NA	81
False positive rate*(%)												
TS-LMM	NA	19	NA	NA	24	NA	NA	14	NA	NA	23	NA
MVMR-Egger	NA	16	NA	NA	32	NA	NA	12	NA	NA	22	NA
MVMR IVW	NA	22	NA	NA	40	NA	NA	18	NA	NA	28	NA
MR-PRESSO	NA	23	NA	NA	58	NA	NA	16	NA	NA	55	NA

#Coverage is defined as the proportion of 95% CIs including the true causal effect over all simulated datasets.
&Power is estimated by the proportion of 95% CIs excluding zero in settings with significant causal effect.
*False positive rate is defined as the proportion of significant causal estimates in settings with no causal effect.

Table 4. Causal associations between lipoprotein cholesterol biomarkers and longevity using multivariable Mendelian randomization (MVMR).

	TS-LMM	MR-PRESSO	TS-LMM	MR-PRESSO
Lipoprotein cholesterol biomarkers	Effect size (95% CI); P value	Effect size (95% CI); P value	Effect size (95% CI); P value	Effect size (95% CI); P value
Genetic variants with P value < $5*10^{-8}$ with any of the 10 cholesterol biomarkers				
	IVs with low LD ($r^2 \leq 0.1$) (No of IVs=47)		IVs with moderate LD ($r^2 \leq 0.5$) (No of IVs=211)	
XL.HDL.C	-0.38 (-3.86, 3.10); 0.830	-0.68 (-2.09, 0.74); 0.355	-0.36 (-1.55, 0.82); 0.548	-0.39 (-1.05, 0.26); 0.244
L.HDL.C	0.59 (-5.68, 6.87); 0.853	0.65 (-0.96, 2.25); 0.434	-0.44 (-1.60, 0.72); 0.456	0.28 (-0.45, 1.01); 0.456
M.HDL.C	2.38 (-5.99, 10.75); 0.577	0.27 (-0.51, 1.05); 0.508	1.12 (-0.54, 2.77); 0.186	0.53 (0.15, 0.92); 0.007
L.VLDL.C	-0.58 (-5.80, 4.64); 0.828	1.66 (-0.41, 3.73); 0.125	0.11 (-0.91, 1.12); 0.838	0.20 (-0.77, 1.17); 0.683
M.VLDL.C	-6.56 (-18.26, 5.15); 0.272	-1.65 (-4.07, 0.76); 0.188	0.71 (-0.71, 2.13); 0.327	-0.92 (-2.09, 0.25); 0.124
S.VLDL.C	-6.16 (-17.85, 5.53); 0.301	0.66 (-1.87, 3.19); 0.612	1.43 (-0.12, 2.98); 0.071	1.18 (0.12, 2.24); 0.030
L.LDL.C	-8.04 (-22.48, 6.39); 0.275	-0.92 (-5.48, 3.65); 0.696	-1.29 (-3.74, 1.17); 0.304	-1.74 (-3.73, 0.24); 0.086
M.LDL.C	-2.23 (-8.24, 3.78); 0.467	2.51 (-2.01, 7.02); 0.283	0.81 (-1.70, 3.33); 0.525	0.36 (-1.65, 2.37); 0.727
S.LDL.C	-9.10 (-20.35, 2.15); 0.113	-2.69 (-5.69, 0.31); 0.087	0.06 (-0.91, 1.03); 0.906	-0.37 (-1.64, 0.89); 0.564
IDL.C	-5.71 (-14.15, 2.73); 0.185	0.76 (-1.82, 3.33); 0.568	0.60 (-0.40, 1.60); 0.242	1.05 (-0.07, 2.18); 0.067
Genetic variants with P value < $5*10^{-5}$ with any of the 10 cholesterol biomarkers				
	IVs with low LD ($r^2 \leq 0.1$) (No of IVs= 739)		IVs with moderate LD ($r^2 \leq 0.5$) (No of IVs=1213)	
XL.HDL.C	-0.01 (-0.56, 0.53); 0.959	0.30 (0.07, 0.53); 0.010	0.13 (-0.08, 0.34); 0.224	0.16 (-0.04, 0.36); 0.108
L.HDL.C	0.03 (-0.37, 0.43); 0.880	-0.23 (-0.51, 0.05); 0.107	0.09 (-0.19, 0.38); 0.521	-0.13 (-0.37, 0.11); 0.278
M.HDL.C	**0.56 (0.24, 0.88); 0.001**	**0.32 (0.15, 0.49); 0.000**	**0.53 (0.26, 0.81); 0.000**	**0.34 (0.20, 0.48); 0.000**
L.VLDL.C	-0.45 (-0.80, -0.11); 0.011	-0.31 (-0.61, 0.00); 0.049	**-0.73 (-1.02, -0.44); 0.000**	-0.30 (-0.57, -0.03); 0.031
M.VLDL.C	0.01 (-0.54, 0.56); 0.968	0.14 (-0.27, 0.56); 0.497	-0.19 (-0.79, 0.40); 0.525	0.01 (-0.34, 0.37); 0.949
S.VLDL.C	0.17 (-0.30, 0.63); 0.479	0.43 (0.03, 0.82); 0.035	0.03 (-0.39, 0.45); 0.891	**0.50 (0.17, 0.84); 0.004**
L.LDL.C	-1.62 (-2.88, -0.35); 0.012	-1.01 (-1.72, -0.29); 0.006	**-1.68 (-2.80, -0.55); 0.004**	**-1.27 (-1.88, -0.65); 0.000**
M.LDL.C	-0.11 (-0.40, 0.17); 0.438	0.34 (-0.33, 1.01); 0.318	-0.21 (-0.51, 0.09); 0.164	0.37 (-0.21, 0.94); 0.215

S.LDL.C	-0.86 (-1.75, 0.04); 0.060	-0.19 (-0.61, 0.22); 0.363	-0.79 (-1.64, 0.05); 0.067	-0.17 (-0.53, 0.20); 0.368
IDL.C	-0.10 (-0.53, 0.33); 0.656	0.36 (-0.02, 0.75); 0.065	-0.07 (-0.43, 0.30); 0.718	**0.53 (0.19, 0.86); 0.002**

We corrected for multiple testing using Bonferroni correction, and defined significance as P value<0.05/10, where 10 is the number of cholesterols included in the model.

XL.HDL.C: Cholesterol in very large high-density lipoprotein (HDL); L.HDL.C: Cholesterol in large HDL; M.HDL.C: Cholesterol in medium HDL; L.VLDL.C: Cholesterol in large very low-density lipoprotein (VLDL); M.VLDL.C: Cholesterol in medium VLDL; S.VLDL.C: Cholesterol in small VLDL; L.LDL.C: Cholesterol in large low-density lipoprotein (LDL); M.LDL.C: Cholesterol in medium LDL; S.LDL.C: Cholesterol in small LDL; IDL.C: Cholesterol in intermediate-density lipoprotein (IDL).