HGG
Advances

ARTICLE

Multivariable Mendelian randomization with incomplete measurements on the exposure variables in the Hispanic Community Health Study/Study of Latinos

Yilun Li,[1] Kin Yau Wong,[2] Annie Green Howard,[1,3] Penny Gordon-Larsen,[3,4] Heather M. Highland,[5] Mariaelisa Graff,[5] Kari E. North,[5] Carolina G. Downie,[5] Christy L. Avery,[3,5] Bing Yu,[6] Kristin L. Young,[5] Victoria L. Buchanan,[5] Robert Kaplan,[7,10] Lifang Hou,[8] Brian Thomas Joyce,[8] Qibin Qi,[7] Tamar Sofer,[9] Jee-Young Moon,[7] and Dan-Yu Lin[1,11,*]

Summary

Multivariable Mendelian randomization allows simultaneous estimation of direct causal effects of multiple exposure variables on an outcome. When the exposure variables of interest are quantitative omic features, obtaining complete data can be economically and technically challenging: the measurement cost is high, and the measurement devices may have inherent detection limits. In this paper, we propose a valid and efficient method to handle unmeasured and undetectable values of the exposure variables in a one-sample multivariable Mendelian randomization analysis with individual-level data. We estimate the direct causal effects with maximum likelihood estimation and develop an expectation-maximization algorithm to compute the estimators. We show the advantages of the proposed method through simulation studies and provide an application to the Hispanic Community Health Study/Study of Latinos, which has a large amount of unmeasured exposure data.

Introduction

Mendelian randomization (MR) enables estimation of the total causal effect of an exposure on an outcome with observational data, using genetic variants as instrumental variables (IVs).[1] When there are multiple correlated exposures, multivariable MR (MVMR) is needed to simultaneously estimate the direct causal effects of the exposures on the outcome.[2] Specifically, by using genetic variants that satisfy the IV assumptions, which means that the variants are associated with the exposure, not associated with the confounder of the exposure-outcome relationship, and only affect the outcome through their effects on the exposure, the study participants can be divided into different genotypic subgroups that have a different level of the exposure but do not have systematically different levels of the confounders. The discrepancy in the outcome between different genotypic subgroups can then imply the causal effect of the exposure on the outcome, making it possible to infer causal effects even if unmeasured confounders exist.

In practice, data on the exposures may be available for only a subset of the study participants because measurements are costly and some values are beyond the detection limits of the assays. A reliable MVMR method that appropriately accounts for the incompleteness of the exposures resulting from different causes is needed.

Many studies excluded individuals with unmeasured exposures from MVMR analyses.[3–5] However, such complete-case analysis results in a loss of information and thus reduced statistical efficiency, especially when the proportion of missingness is large. Moreover, exclusion of individuals with incomplete data will result in biased effect estimators if data are not missing completely at random.[6]

When the exposures of interest are quantitative omics variables, measurements above or below certain values may not be detected. In univariable MR studies, it is common for researchers to impute undetectable values with a fixed quantity related to the detection limits.[7–9] However, single imputation can cause bias in effect estimation and inflated type I error in hypothesis testing.[10] In MVMR analysis, the issue of undetectable exposure values has received even less attention.

Existing methods for MVMR analysis with individual-level data are based on the two-stage least-squares (TSLS) estimator.[2,11] In the TSLS methods, however, estimation of

[1]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong; [3]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [4]Department of Nutrition, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [5]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [6]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [7]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [8]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; [9]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; [10]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA
[11]Lead contact
*Correspondence: lin@bios.unc.edu
https://doi.org/10.1016/j.xhgg.2024.100338.

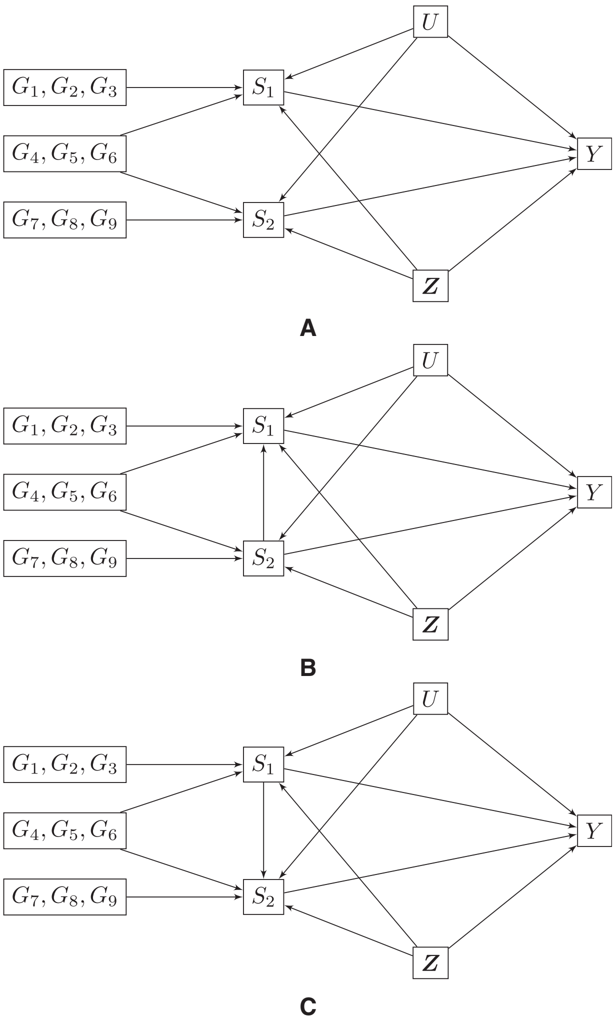

Figure 1. Direct acyclic graphs for the simulation studies where *Y* is the outcome, S_1 and S_2 are the exposure variables, G_1, G_2, …, and G_9 are nine IVs for S_1 and S_2, $\boldsymbol{Z}$ is a vector of measured covariates, and *U* is an unmeasured confounder

(A) Scenario where S_1 and S_2 are in independent pathways from the IVs to the outcome.

(B) Scenario where S_2 confounds the relationship between S_1 and *Y*.

(C) Scenario where S_2 mediates the relationship between S_1 and *Y*.

parameters in different stages is performed sequentially, and the correlation between the random errors of different models is not accounted for. In contrast, the (full-information) maximum likelihood method estimates all parameters simultaneously and takes into account the correlation between the error terms in the models, which can lead to greater efficiency.[12]

In this paper, we propose a valid and efficient method for MVMR analysis with a continuous outcome and two continuous exposure variables that are potentially unmeasured and undetectable. We construct a linear model between each exposure variable and the IVs and another linear model between the outcome and the exposure variables. We account for the unmeasured confounders and the potential correlation between exposure variables by allowing the error terms in the linear models to be correlated. We estimate the direct causal effects by the maximum likelihood estimators and use the expectation-maximization (EM) algorithm for computation. The proposed estimators are consistent and statistically efficient. We demonstrate the advantages of the proposed method over existing ones with simulated data. Lastly, we apply the proposed method to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).[13,14]

Material and methods

Let *Y* be a continuous outcome, S_1 and S_2 be two potentially correlated continuous exposure variables whose values can be unmeasured or undetectable, and $\boldsymbol{G}$ be a vector of IVs for S_1 and S_2. We also let $\boldsymbol{Z}$ be a vector of measured covariates, including the unit component, and let $\boldsymbol{X} = (\boldsymbol{G}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$. The number of IVs must be greater than or equal to the number of exposure variables.[2] In addition, according to existing literature on MVMR, each component of $\boldsymbol{G}$ should satisfy the following conditions: (1) it is associated with at least one exposure, (2) it does not affect the outcome except through its effects on the exposure variables, and (3) it is not associated with the confounders of any exposure-outcome relationships.[11,15] The exposure variables can be in independent pleiotropic pathways from the IVs to the outcome, or one exposure can confound or mediate the relationship between the other exposure and the outcome. Several examples are shown in Figure 1. Finally, each exposure should have at least one IV.

We consider the following linear models:

$$S_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X} + \epsilon_{S_1}, \qquad \text{(Equation 1)}$$

$$S_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{X} + \epsilon_{S_2}, \qquad \text{(Equation 2)}$$

and

$$Y = \gamma_1 S_1 + \gamma_2 S_2 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z} + \epsilon_Y, \qquad \text{(Equation 3)}$$

where $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\beta}$ are regression parameters; γ_1 and γ_2 represent the direct causal effects of S_1 and S_2 on *Y*, respectively; and $(\epsilon_{S_1}, \epsilon_{S_2}, \epsilon_Y)^{\mathrm{T}}$ is a zero-mean three-dimensional normal random vector, with $\mathrm{Var}(\epsilon_{S_j}) = \sigma_j^2$, $\mathrm{Var}(\epsilon_Y) = \sigma_Y^2$, $\mathrm{Corr}(\epsilon_{S_1}, \epsilon_{S_2}) = \rho_{12}$, and $\mathrm{Corr}(\epsilon_{S_j}, \epsilon_Y) = \rho_{jY}$ $(j = 1, 2)$. For model identifiability, we require that $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are linearly independent. In practice, measurements of the exposure variables are usually non-negative. However, we allow S_1 and S_2 to be negative to accommodate situations in which certain transformations (e.g., standardization and log-transformation) are performed. The joint density function of (Y, S_1, S_2) given $(\boldsymbol{X}, \boldsymbol{Z})$ is

$$f(Y, S_1, S_2|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(S_1 - \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}, S_2 - \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{X}, Y - \gamma_1 S_1 - \gamma_2 S_2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})\boldsymbol{\Sigma}^{-1}\begin{pmatrix} S_1 - \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X} \\ S_2 - \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{X} \\ Y - \gamma_1 S_1 - \gamma_2 S_2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z} \end{pmatrix}\right],$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\alpha}_2^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \gamma_1, \gamma_2, \sigma_1^2, \sigma_2^2, \sigma_Y^2, \rho_{12}, \rho_{1Y}, \rho_{2Y})^{\mathrm{T}}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $(\epsilon_{S_1}, \epsilon_{S_2}, \epsilon_Y)^{\mathrm{T}}$.

Let M_j indicate, by the values 1 versus 0, whether or not S_j is measured $(j = 1, 2)$. Let L_j and U_j be the intrinsic lower and upper detection limits of S_j, respectively. We define $R_j = M_j I(L_j \leq S_j \leq U_j)$, where I is the indicator function, such that $R_j = 1$ if S_j is measured and detectable and $R_j = 0$ otherwise. When $R_j = 0$, S_j is only known to lie in an interval C_j, where $C_j = (-\infty, L_j)$ if $S_j < L_j$ and $M_j = 1$, $C_j = (U_j, \infty)$ if $S_j > U_j$ and $M_j = 1$, and $C_j = (-\infty, \infty)$ if $M_j = 0$. We assume the missing-at-random mechanism, such that $(M_j, L_j, U_j)_{j=1,2}$ and (S_1, S_2) are independent given $(Y, \boldsymbol{G}, \boldsymbol{Z})$. We allow the detection limits to vary across individuals so as to accommodate multicenter studies. Then, for a sample with n individuals, the observed data consist of $\{Y_i, \boldsymbol{G}_i, \boldsymbol{Z}_i, M_{1i}, M_{2i}, R_{1i}, R_{2i}, R_{1i}S_{1i} + (1 - R_{1i})C_{1i}, R_{2i}S_{2i} + (1 - R_{2i})C_{2i}\}$ for $i = 1, \ldots, n$.

We assume that the joint distribution of R_{ji}, L_{ji}, and U_{ji} $(i = 1, \ldots, n; j = 1, 2)$ does not depend on $\boldsymbol{\theta}$. The observed-data likelihood for $\boldsymbol{\theta}$ is proportional to

$$\prod_{i=1}^{n}\left\{\begin{array}{l} f(Y_i, S_{1i}, S_{2i} | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})^{R_{1i}R_{2i}} \left[\int\limits_{s_1 \in C_{1i}} f(Y_i, s_1, S_{2i} | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) ds_1\right]^{(1-R_{1i})R_{2i}} \\ \times \left[\int\limits_{s_2 \in C_{2i}} f(Y_i, S_{1i}, s_2 | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) ds_2\right]^{R_{1i}(1-R_{2i})} \left[\int\limits_{s_2 \in C_{2i}} \int\limits_{s_1 \in C_{1i}} f(Y_i, s_1, s_2 | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) ds_1 ds_2\right]^{(1-R_{1i})(1-R_{2i})} \end{array}\right\}.$$

We use the EM algorithm to maximize this likelihood, treating the unobserved values of S_1 and S_2 as missing data. The complete-data log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\theta}) = & -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{3n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left[\sum_{k=1}^{2}\sum_{l=1}^{2}\sigma^{(kl)}\left(S_{ki} - \boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\right. \\ & \times\left(S_{li} - \boldsymbol{\alpha}_l^{\mathrm{T}}\boldsymbol{X}_i\right) + 2\sum_{k=1}^{2}\sigma^{(k3)}\left(S_{ki} - \boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\left(Y_i - \gamma_1 S_{1i} - \gamma_2 S_{2i}\right. \\ & \left.\left. - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i\right) + \sigma^{(33)}\left(Y_i - \gamma_1 S_{1i} - \gamma_2 S_{2i} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i\right)^2\right],\end{aligned} \quad \text{(Equation 4)}$$

where $\sigma^{(kl)}$ is the (k,l)th element of $\boldsymbol{\Sigma}^{-1}$ $(k,l = 1,2,3)$. In the E-step, we compute $\widehat{\ell}(\boldsymbol{\theta}) = \widehat{E}[\ell(\boldsymbol{\theta})]$, where $\widehat{E}$ denotes the conditional expectation given the observed data at the current parameter estimates. The conditional expectation $\widehat{\ell}(\boldsymbol{\theta})$ involves the first and second conditional moments of (S_{1i}, S_{2i}). In the M-step, we use the maximizer of the expected complete-data log-likelihood function to update the parameters. We iterate between the E-step and M-step until the Euclidean distance between the parameter estimates at two consecutive iterations is smaller than a pre-specified small positive constant. We use $\widehat{\boldsymbol{\theta}}$ to denote the resulting estimator of $\boldsymbol{\theta}$. Details about the EM algorithm are provided in the supplemental information.

We estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$ with the Louis formula.[6] First, we derive the complete-data information matrix $\mathbf{I}_{\mathrm{c}}$, which is the negative of the Hessian matrix of $\widehat{\ell}(\boldsymbol{\theta})$ evaluated at $\widehat{\boldsymbol{\theta}}$. Then, we compute the gradient of $\log f(Y_i, S_{1i}, S_{2i} | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, denoted by $\boldsymbol{U}_i$, and evaluate $\widehat{E}(\boldsymbol{U}_i)$ and $\widehat{E}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})$ at $\widehat{\boldsymbol{\theta}}$. The observed-data information matrix is

$$\mathbf{I}_{\mathrm{obs}} = \mathbf{I}_{\mathrm{c}} - \sum_{i=1}^{n}(1 - R_{1i}R_{2i})\left[\widehat{E}\left(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}\right) - \widehat{E}(\boldsymbol{U}_i)\widehat{E}(\boldsymbol{U}_i)^{\mathrm{T}}\right], \quad \text{(Equation 5)}$$

and the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $\mathbf{I}_{\mathrm{obs}}^{-1}$. The expressions of $\mathbf{I}_{\mathrm{c}}$ and $\boldsymbol{U}_i$ $(i = 1, \ldots, n)$ are provided in the supplemental information.

Results

We performed extensive simulation studies to compare the performance of the proposed method with that of existing methods under different scenarios. We let $Z_1 = 1$ and generated Z_2 from the standard uniform distribution, Z_3 from the Bernoulli distribution with 0.5 success probability, and Z_4 from the standard normal distribution. The variables Z_2, Z_3, and Z_4 corresponded to (normalized) age, gender, and the first principal component for ancestry, respectively. (Here, we let the age of individuals follow a uniform distribution, so the variable followed the standard uniform distribution after being normalized to the range between 0 and 1.)

We generated nine correlated IVs as follows. First, we generated $\boldsymbol{G}^* \equiv (G_1^*, \ldots, G_9^*)^{\mathrm{T}}$ from the nine-dimensional zero-mean normal distribution with a covariance matrix whose (k,l)th element was $0.2^{|k-l|}$ $(k,l = 1, \ldots, 9)$; $\boldsymbol{G}^*$ was independent of $\boldsymbol{Z}$. Second, for the kth genetic variant, whose minor allele frequency was p_k, we set $G_k^{(1)}$ to 1 if G_k^* was greater than the $(1 - p_k)$ quantile of the standard normal distribution and set $G_k^{(1)}$ to 0 otherwise $(k = 1, \ldots, 9)$. Since an individual inherits two alleles, one from each parent, we repeated the two steps above and generated $G_k^{(2)}$ for $k = 1, \ldots, 9$ with the same procedure. Then, we let $G_k = G_k^{(1)} + G_k^{(2)}$ $(k = 1, \ldots, 9)$. The marginal distribution of G_k was binomial(2, p_k), and the IVs were correlated. We allow the IVs to be correlated since $r^2 < 0.2$ has often been used as the criteria for linkage disequilibrium pruning, which means that the pruned variants can still be slightly correlated. We aim to show that the proposed method performs well even if the IVs are not strictly independent. We set $p_1 = p_4 = p_7 = 0.3$, $p_2 = p_5 = p_8 = 0.4$, and $p_3 = p_6 = p_9 = 0.5$.

We let $\boldsymbol{Z} = (Z_1, \ldots, Z_4)^{\mathrm{T}}$ and $\boldsymbol{X} = (G_1, \ldots, G_9, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$. We let S_1 be associated with G_1 to G_6 and S_2 be associated with G_4 to G_9. We generated the exposure variables and the outcome from the following equations:

$$S_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X} + \lambda_1 U + e_1, \quad \text{(Equation 6)}$$

$$S_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{X} + \lambda_2 U + e_2, \quad \text{(Equation 7)}$$

and

$$Y = \gamma_1 S_1 + \gamma_2 S_2 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z} + \lambda_Y U + e_Y, \quad \text{(Equation 8)}$$

where U, e_1, e_2, and e_Y were independent standard normal variables; the non-zero genetic association parameters (i.e., the first six components of $\boldsymbol{\alpha}_1$ and the last six components of $\boldsymbol{\alpha}_2$) were all set to 0.25; the intercept and the association parameters for the measured covariates were set to 0.15; γ_1 was set to 0 or 0.12; γ_2 was set to 0.12; and λ_1, λ_2, and λ_Y were set to 0.6. A causal diagram is shown in Figure 1A. We simulated the unmeasured confounder U to induce correlations (among S_1, S_2, and Y) that cannot be explained by the IVs and measured covariates, equivalent to simulating correlated residual errors. The selected values of γ_1 and γ_2 represent moderate direct causal effects of the exposures on the outcome. The choices of λ_1, λ_2, and λ_Y also reflect a moderate pairwise correlation among S_1, S_2, and Y.

We set the sample size to 9,000 and the probability of (S_1, S_2) being unmeasured to 2/3 for each individual, mimicking the real dataset. We assumed that the two exposures have equal lower detection limits, which varied from -0.5 to 0.5 with a 0.1 increment, and we assumed no upper detection limit.

The proportion of individuals with undetectable values increased from 0.54% to 8.56% as the lower detection limit increased, which covered the situations in the HCHS/SOL data. To evaluate the strength of the IVs, we calculated the partial F-test statistic and the Sanderson-Windmeijer conditional F-statistic using individuals with complete exposure data.[2] As the lower detection limit increased from -0.5 to 0.5, the number of complete cases became smaller, resulting in a decrease in the mean of the partial F-statistic from 39.36 to 21.36 and a decrease in the mean of the Sanderson-Windmeijer conditional F-statistic from 30.98 to 18.18. Nevertheless, the Sanderson-Windmeijer conditional F-statistics were greater than 10, suggesting that the IVs were sufficiently strong.[2]

We considered three existing methods to handle the datasets generated above, including the complete-case analysis, "imputation at limit," and "imputation at mid-point." For the complete-case analysis, we included only individuals with measured and detectable values for both exposure variables. For the imputation methods, we included only individuals with both exposure variables measured; we imputed values below the lower detection limit L by L for the imputation at limit method and by $L - \log 2$ for the imputation at mid-point method. Then, for all three methods, we used TSLS for estimation. In our simulation studies, we treated S_1 and S_2 as the log-transformation of their original measurements; thus, the imputed value for the imputation at mid-point method was the log of the mid-point between e^L (the lower detection limit on the original scale) and 0 (the smallest possible value of the exposure variable). For each method, we performed the Wald test on each direct causal effect at the nominal significance level of 0.001. We simulated 10,000 and 10 million replicates for $\gamma_1 = 0.12$ and $\gamma_1 = 0$, respectively.

Figure 2 shows the results for the scenario of $\gamma_1 = 0.12$. For both exposure variables, the proposed direct causal effect estimators are nearly unbiased, the proposed standard error estimators are accurate, and the 95% confidence intervals have correct empirical coverage probabilities. The complete-case analysis yields negatively biased direct causal effect estimators, and it has extremely low power in testing γ_1 and γ_2; as the lower detection limit increases, the estimators become more severely biased, the standard errors increase, and the empirical coverage probabilities of the nominal 95% confidence intervals decrease substantially. The two imputation methods yield estimators that are biased away from the null value, and the magnitude of bias increases as the lower detection limit becomes larger. In addition, the imputation methods yield much lower power in testing γ_1 and γ_2 than the proposed method.

Figure 3 shows the results for the scenario of $\gamma_1 = 0$. The results of the inference on γ_2 are similar to those in the previous scenario. For the inference on γ_1, the proposed method performs the best among all of the methods, yielding unbiased estimators with the smallest standard error, accurate standard error estimators, correct empirical coverage probabilities, and correct empirical type I error in testing γ_1. The complete-case estimator is negatively biased; as the lower detection limit increases, the bias becomes more severe, and the type I error becomes more inflated. The two imputation methods yield virtually unbiased estimators for γ_1 and correct type I errors in testing γ_1, but those estimators have much larger standard errors than the estimators from the proposed method.

Figures 2 and 3 show that as the lower detection limit increases, the power in testing γ_2 remains nearly a constant for the imputation methods, although the standard error increases. Our inspections show that the empirical distributions of the z-values over the replicates are almost unchanged as the lower detection limit increases, where the z-value is defined as the ratio of the causal effect estimate to the standard error estimate; this is also consistent with the results that both the magnitude of bias and the standard error increase with an increasing detection limit. As a result, the proportion of z-values that exceed the range between the 0.05% and the 99.95% quantiles of the standard normal distribution is almost unchanged, which explains why the power in testing γ_2 is nearly the same when the detection limit varies.

In previous simulation studies, we generated data from Equations 6, 7, and 8, processed the data using the complete-case or imputation methods, and estimated the parameters with the TSLS method to compare the performance with the proposed method. All the methods involved above use individual-level data. Here, we considered replacing the TSLS method with another parameter estimation method called "MVMR based on constrained maximum likelihood" (MVMR-cML) as a comparison; this MVMR-cML method is robust to the violation of IV assumptions and can accommodate one-sample and two-sample designs.[16] The data generation process and the specifications of parameters remained unchanged. Since the complete-case analysis and the imputation at limit method perform worse than the imputation at mid-point method, we only applied the MVMR-cML method to the dataset processed by the imputation at mid-point approach. Results from Tables S1 and S2 show that the TSLS and MVMR-cML methods had similar performances when γ_1 was set to 0.12, but MVMR-cML yielded inflated type I errors in testing γ_1 when the true effect was 0. Based on these results, we conclude that the proposed method performs better than imputation at mid-point with either TSLS or MVMR-cML.

We also conducted simulation studies where S_2 confounds or mediates the relationship between S_1 and Y (see Figures 1B and 1C). When S_2 is a confounder, the data were generated from Equations 7, 8, and 9:

$$S_1 = \boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X} + \lambda_1 U + \lambda_{21} S_2 + e_1, \qquad \text{(Equation 9)}$$

where λ_{21} was set to 0.2. When S_2 is a mediator, the data were generated from Equations 6, 8, and 10:

$$S_2 = \boldsymbol{\alpha}_2^{\mathrm{T}} \boldsymbol{X} + \lambda_2 U + \lambda_{12} S_1 + e_2, \qquad \text{(Equation 10)}$$

where λ_{12} was set to 0.2. The rest of the simulation setup remained the same. Figures S1 and S2 show the results when S_2 is a confounder, and Figures S3 and S4 show the results when S_2 is a mediator. In both situations, the proposed method performs the best among the four methods, yielding unbiased effect estimators with the smallest standard errors, accurate standard error estimators, and the highest power (under the alternative hypothesis) and correct type I error (under the null hypothesis) in testing the causal effects.

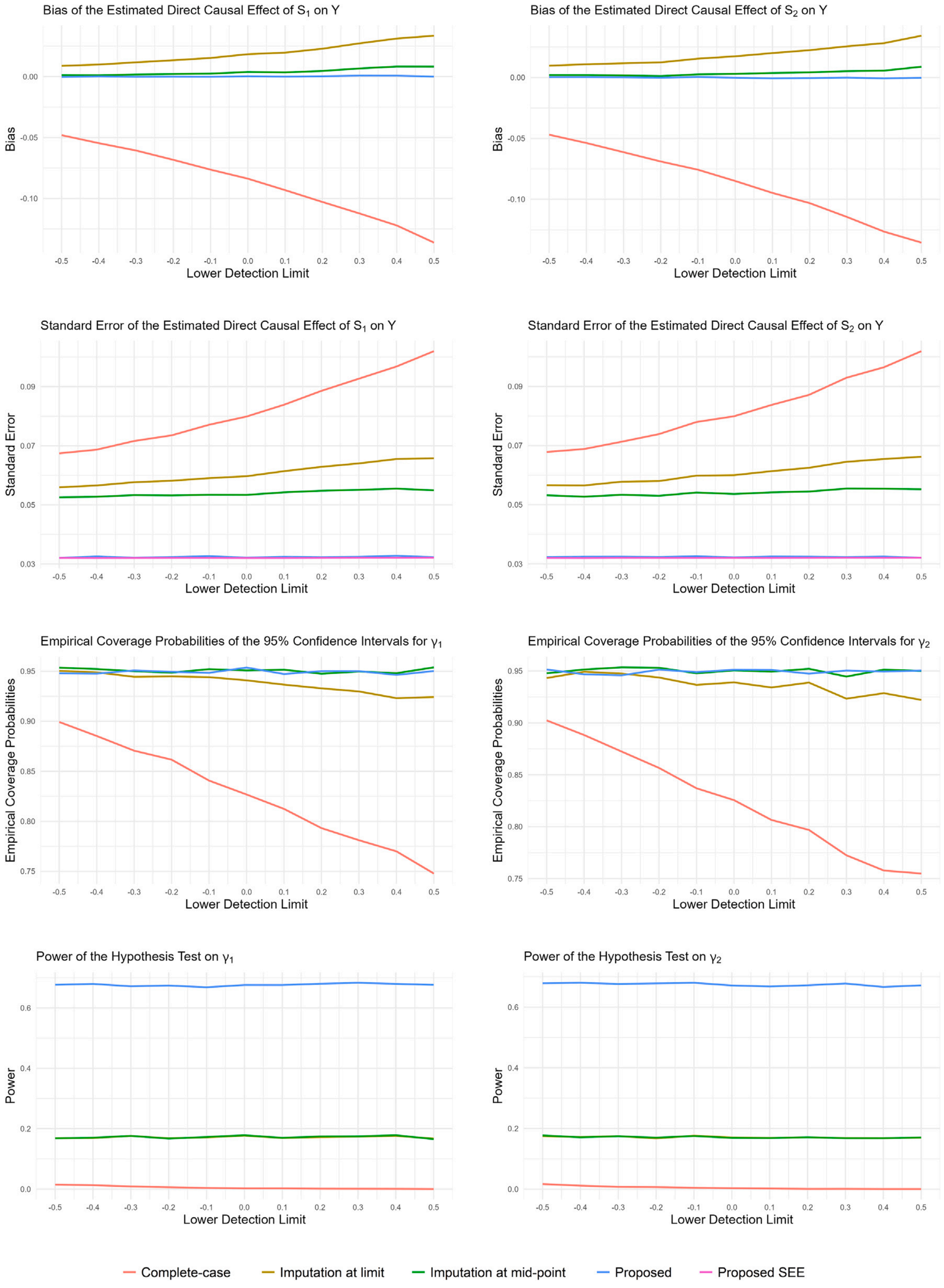

Figure 2. Simulation results for the scenario where S_1 and S_2 are in independent pathways from the IVs to the outcome, with γ_1 set to 0.12

The left and right images correspond to the inference on γ_1 and γ_2, respectively. The bias and standard error of the estimators, the empirical coverage probabilities of the 95% confidence intervals, and the empirical power of the hypothesis test on γ_1 and γ_2 are plotted against the lower detection limit of each exposure variable. The red, brown, green, and blue curves correspond to the complete-case analysis, the imputation at limit method, the imputation at mid-point method, and the proposed method, respectively. The pink curve represents the mean of the standard error estimator (SEE) given by the proposed method.

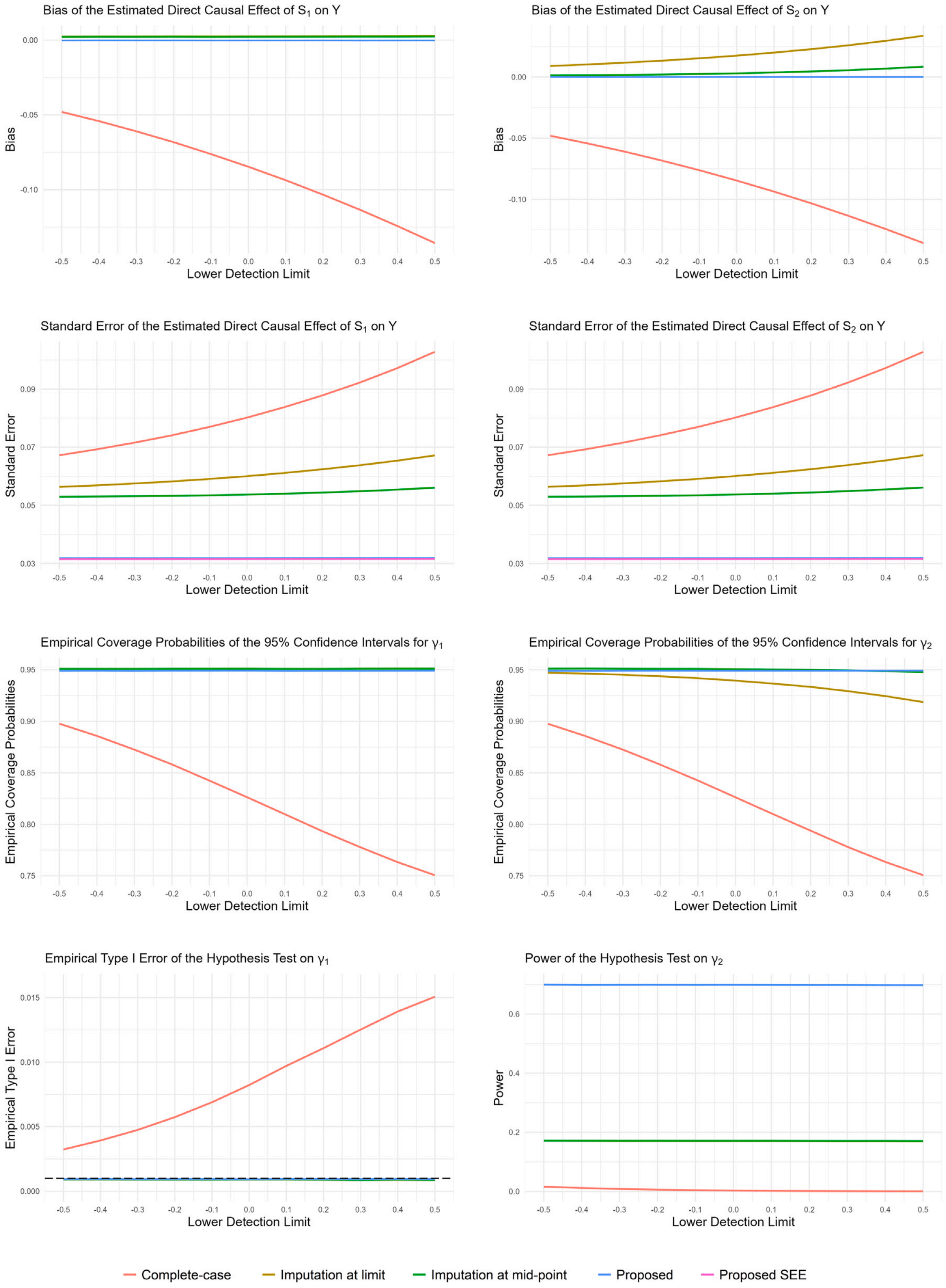

Figure 3. Simulation results for the scenario where S_1 and S_2 are in independent pathways from the IVs to the outcome, with γ_1 set to 0
The left and right images correspond to the inference on γ_1 and γ_2, respectively. The bias and standard error of the estimators, the empirical coverage probabilities of the 95% confidence intervals, the empirical type I error of the hypothesis test on γ_1, and the empirical power of the hypothesis test on γ_2 are plotted against the lower detection limit of each exposure variable. The red, brown, green, and blue curves correspond to the complete-case analysis, the imputation at limit method, the imputation at mid-point method, and the proposed method, respectively. The pink curve represents the mean of the SEE given by the proposed method. The black dashed line in the plot for empirical type I error represents the nominal level of 0.001.

To assess the performance of the proposed method when the exposure variables have lower heritability, we performed additional simulation studies in which we only simulated three IVs. Specifically, the genetic variants G_k ($k = 1, 2, 3$) were generated with the same approach as before except that we reduced the dimension from 9 to 3. We set the minor allele frequency of the three IVs to 0.3. We let S_1 be associated with G_1 and G_2 and S_2 be associated with G_2 and G_3. We also set γ_1 to 0 or 0.25 and set γ_2 to 0.25. We kept the other simulation parameters unchanged and then generated the simulated data with Equations 6, 7, and 8.

To maintain similar proportions of undetectable exposure values compared with previous simulation studies, we set the lower detection limit to vary from -1.5 to -0.5. As a result, when the lower detection limit increased, the proportion of individuals with undetectable exposure values increased from 0.96% to 6.90%, the mean of the partial F-statistic in the first-stage regression model decreased from 34.94 to 21.89, and the mean of the Sanderson-Windmeijer conditional F-statistic decreased from 24.54 to 16.35. The heritability of S_1 and S_2 was about 0.027, similar to the HCHS/SOL data.

Figure S5 shows the results for the scenario of $\gamma_1 = 0.25$. The results are similar to those in Figure 2 except for having higher levels of bias (for the complete-case and imputation methods) and larger standard errors due to decreased strength of the IVs. Figure S6 shows the results for the scenario of $\gamma_1 = 0$. The results on γ_2 are similar to those in Figure S5. As for the inference on γ_1, all the estimators are nearly unbiased except for the complete-case estimators, and the proposed method yields the least standard errors and provides accurate standard error estimators. The empirical type I errors for testing γ_1 are below the nominal level for all methods. By checking the histograms and quantile-quantile plots of the z-values, we attributed the deflation of type I errors to the thin-tailed distribution of the z-values. Figure S6 also shows a non-monotonic trend in the empirical type I errors of testing γ_1 for the complete-case analysis. When the lower detection limit increases from -1.5 to -0.7, the magnitude of bias increases, leading to higher levels of empirical type I errors; however, when the lower detection limit further increases, the tails of the distributions of the z-values are so thin that the proportion of the z-values that exceed the range between the 0.05% and 99.95% quantiles of the standard normal distribution decreases, leading to lower empirical type I errors.

The above simulation results show that the removal or single imputation of the undetectable values in the exposures can lead to biased direct causal effect estimators in MVMR analyses. In addition, complete-case analysis and the two imputation methods exclude individuals with unmeasured exposures, resulting in information loss and low statistical efficiency. The proposed method can overcome the limitations of the existing methods and yield unbiased estimators and substantially higher statistical power.

Application to the HCHS/SOL

We used the proposed method to assess the direct causal effects of two metabolites, 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) (a phosphatidylinositol) and 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) (a phosphatidylcholine), on total cholesterol (TC), triglycerides (TGs), and low-density lipoprotein cholesterol (LDL-C) in the HCHS/SOL participants. Phosphatidylinositol and phosphatidylcholine were previously found to be related to the secretion, transport, and excretion of cholesterol[17,18] and were potentially correlated since they are both involved in glycerophospholipid metabolism (according to the web resource available at https://www.genome.jp/pathway/hsa00564). In addition, a previous study of the HCHS/SOL participants showed that both 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) and 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) correlated with the three lipoprotein variables mentioned above, with each metabolite-lipoprotein pair having an absolute Pearson's correlation parameter greater than 0.15 ($p < 10^{-28}$).[19] We performed an MVMR analysis to assess the relevant direct causal effects and better understand the underlying causal relationships.

The HCHS/SOL is a multicenter longitudinal cohort study of Hispanics/Latinos in the United States. Through a stratified multistage area probability sampling strategy, a total of 16,415 individuals from different Hispanic/Latino backgrounds (Central American, Cuban, Dominican, Mexican, Puerto Rican, South American, and others) were recruited in the Bronx, Chicago, Miami, and San Diego.[14] The HCHS/SOL study was approved by institutional review boards at participating institutions. Written informed consent was obtained from all participants. Only data from the participants' baseline visit were used in our analysis.

Fasting blood was collected at the baseline visit. The process by which the lipoprotein variables were measured is described in the HCHS/SOL Manual 7 (Addendum), available at https://sites.cscc.unc.edu/hchs/manuals-forms. Non-fasting participants and individuals with missing outcome measurements were excluded from our analysis. In addition, the measurements for individuals using statins, a type of lipid-lowering medication, were adjusted based on the results of Wu et al.[20]

Measurements of the exposure variables were only available for a randomly selected subset of about 1/3 of the study participants with available genetic data at the baseline visit. The processes of metabolomic profiling and quantification were described in previous literature.[19] We considered the minimal measured and detectable value for each metabolite exposure as the intrinsic lower detection limit L_0. In addition, we treated outlying values as being beyond detection limits. Specifically, we set an upper detection limit of $\exp(m+2s)$ and redefined the lower detection limit as $\max\{\exp(m - 2s), L_0\}$, where m and s were the sample mean and sample standard deviation of the log-transformed measurements, respectively.

We obtained data on the genetic variants significantly associated with at least one exposure through an existing genome-wide association study (Table S3).[21] Details about the IV selection process are shown in the supplemental information. We performed the imputation with the TOPMed freeze 8 imputation reference panel[22–24] and derived the principal components for ancestry. To handle genetic relatedness among the HCHS/SOL participants, we used the software package of Pedigree Reconstruction and Identification of a Maximum Unrelated Set to obtain the maximum unrelated subset of participants, such that the estimated proportion of alleles shared identical by descent was no more than 0.2 for any individual pair.[25] All analyses were performed using unrelated individuals only.

For the demographic variables, age and gender information was collected during participants' baseline visits. The Hispanic/Latino background was derived using the method described in Conomos et al.[26]

We performed the inverse-normal transformation on each lipoprotein outcome and the measured values of each metabolite exposure. Then, we evaluated the direct causal effects of the transformed exposures on each transformed outcome using the genetic variants in Table S3 as the IVs and using age, gender, the center of recruitment, the Hispanic/Latino background, and the first five

Table 1. Estimated direct causal effects of the exposure variables on the outcomes

			Proposed method				Imputation at mid-point			
Outcome	*n*	Exposure variable	Est	SE	95% CI	*p* value	Est	SE	95% CI	*p* value
TC	9,608	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	0.416	0.088	(0.243, 0.590)	2.52E−06	0.351	0.121	(0.115, 0.588)	3.64E−03
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	0.010	0.031	(−0.050, 0.070)	0.736	0.012	0.046	(−0.079, 0.103)	0.795
TGs	9,608	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	0.608	0.093	(0.425, 0.791)	7.26E−11	0.655	0.119	(0.422, 0.888)	3.47E−08
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	−0.023	0.034	(−0.089, 0.043)	0.499	−0.014	0.045	(−0.104, 0.075)	0.750
LDL-C	9,428	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	0.239	0.085	(0.072, 0.406)	0.005	0.153	0.150	(−0.141, 0.448)	0.308
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	−0.039	0.031	(−0.099, 0.021)	0.201	−0.064	0.054	(−0.169, 0.041)	0.229

Abbreviations: *n*, sample size; Est, direct causal effect estimate; SE, standard error; CI, confidence interval; TC, total cholesterol; TGs, triglycerides; LDL-C, low-density lipoprotein cholesterol.

principal components for ancestry as the measured covariates. We performed the analysis with all methods described in the simulation studies. For both imputation methods, we imputed values above the upper detection limit with the upper detection limit. For the imputation at limit method, we imputed values below the lower detection limit by the lower detection limit; for the imputation at mid-point method, we imputed values below the lower detection limit by half of the lower detection limit on the original scale.

We computed the partial *F*-statistics and the Sanderson-Windmeijer conditional *F*-statistics to assess IV strength. In addition, we performed the Sargan test[2] to evaluate whether there was significant unmeasured horizontal pleiotropy.

Descriptive statistics are shown in Table S4, and the numbers of individuals with an unmeasured or undetectable exposure are presented in Table S5. In brief, only about one-third of the participants have measured values for the two metabolite exposures, and about 2.3% of the individuals (who have measured metabolites) have undetectable exposure values.

Statistical analysis results are shown in Table 1. The proposed method detected positive direct causal effects of 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) on TC, TGs, and LDL-C ($p < 0.005$). In addition, 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) had a positive effect on TC and a negative effect on TGs and LDL-C, although these effects were not statistically significant. The imputation at mid-point method yielded similar effect estimates but reported 28.0%–76.5% larger standard errors, wider confidence intervals, and larger *p* values than the proposed method. The imputation at mid-point method failed to detect a significant direct causal effect of 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) on LDL-C. Results from the complete-case analysis and the imputation at limit method were similar to those from the imputation at mid-point method since the proportions of undetectable values were small; these results are given in Table S6.

Table 2 shows the results relevant to the assessment of IV assumptions. The partial *F*-statistics and the Sanderson-Windmeijer conditional *F*-statistics were all greater than the rule-of-thumb value of 10,[27] indicating that the IVs were sufficiently strong.[2] In addition, the *p* values of the Sargan test were all greater than 0.05, showing no statistically significant unmeasured horizontal pleiotropy. Thus, the IVs in our analysis satisfied the assumptions required.

Discussion

In this paper, we present a maximum likelihood estimation method for MVMR analysis with two exposure variables whose values may be unmeasured and undetectable. Unlike the existing TSLS method, where the parameters in different model equations are estimated separately, the proposed method includes all parameters in the likelihood and conducts joint estimation. The proposed method performs well even when only a small proportion of the exposure values are measured and within detection limits. In addition, we can handle outliers by treating them as being beyond the detection limits, as shown in application to the HCHS/SOL, whereas the common practice of removing outliers or winsorizing may induce bias and reduce power. The proposed method is based on the condition that the IV assumptions are met, and our method is not robust to weak IVs or horizontal pleiotropy. For the scenarios where individual genetic variants (that do not induce horizontal pleiotropy) are weak, we can construct allele scores (e.g., polygenic risk scores and unweighted allele scores) to form stronger instruments for MR analysis.[28–30] We measured the IV strength with the Sanderson-Windmeijer conditional *F*-statistic, which should be greater than 10 for an IV to be considered strong.[2]

Besides the methods discussed in simulation studies, multiple imputation is another common approach to address missing data. However, most existing algorithms generate the imputed values from a specific distribution that requires proper estimation of the parameters or specification of a prior distribution for the parameters, which can be challenging since the data are incomplete. In addition, multiple imputation is computationally intensive, especially when there are missing values for multiple

Table 2. Results of assessing the instrumental variable assumptions

Outcome	Sargan test p value	Exposure variable	F	F_c
TC	0.156	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	39.032	15.470
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	161.130	103.668
TGs	0.103	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	39.032	15.470
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	161.130	103.668
LDL-C	0.105	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	32.131	13.044
		1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	154.388	98.707

Abbreviations: F, partial F-statistic; F_c, the Sanderson-Windmeijer conditional F-statistic; TC, total cholesterol; TGs, triglycerides; LDL-C, low-density lipoprotein cholesterol.

variables or the proportion of missing values is large.[6] Thus, we did not consider multiple imputation in this paper.

One limitation of the proposed method is that it can only handle two exposure variables. Lin et al. proposed a general framework to handle the unmeasured and undetectable values in more than two exposures.[10] The authors considered similar models to those in this paper, except that the residuals of the linear models are assumed uncorrelated in their models; as a result, the parameter estimates can be updated with explicit formulas in the M-step, and thus, the computational burden of the EM algorithm is not a concern. However, by assuming no correlation among the residuals, the unmeasured confounders are not considered, and the models can only infer associations rather than causality. When we incorporate the residual correlations into the models to estimate the causal effects, we need to run the M-step with the Newton-Raphson algorithm, which makes the computation more complicated and challenging when the number of exposures increases. The computation in the E-step and the derivation of the estimated covariance matrix will also become much more complex when we include more exposure variables. There may also be extra computational challenges when the sample size is large. As a future direction, we will develop a computationally efficient algorithm to accommodate more exposure variables and larger sample sizes.

For a general MVMR analysis, investigators can include the potential confounders as exposures to avoid horizontal pleiotropy. The proposed method is designed for the setting with two exposure variables, so we should be careful in the IV selection in order not to induce horizontal pleiotropic effects. For instance, a previous study found that the genetic variant rs174559 in the *FADS1* gene (MIM: 606148) on chromosome 11 was significantly associated with both 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) and 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) in the HCHS/SOL.[21] However, including this variant as an IV in our analysis led to significant unmeasured horizontal pleiotropy, evident by p values of the Sargan test being smaller than 0.003 in the analysis of TGs and LDL-C. For the analysis of TC, including rs174559 did not lead to unmeasured horizontal pleiotropy, and the corresponding results were close to those reported in Table 1.

A previous univariable MR study reports that the total causal effects of 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) on TC, TGs, and LDL-C are significantly positive and that 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) has significantly negative total effects on TGs and LDL-C.[31] Our MVMR analysis results suggest direct causal effects in the same directions; however, the direct causal effects of 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) on TGs and LDL-C are not significant. The difference in the MR and MVMR results implies that the effects of 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) on TGs and LDL-C may be operated through 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4), and more biological and statistical investigations may be needed.

While little attention has been paid to the biological links between the two metabolites and the levels of lipoproteins in the existing literature, our analysis provides an assessment from a statistical perspective. For example, the significant direct causal effects show that 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) can be a potential biomarker or therapeutic target for hyperlipidemia. Future biological or epidemiological research can further study the roles of these metabolites, which is important to drive more insights into the treatment of relevant diseases. In addition, our study focuses on individuals with Hispanic/Latino backgrounds, and it is worthwhile to extend the scope to consider individuals from other ethnic groups and compare the effects of the metabolites on lipoproteins in different populations; this may help develop population-specific strategies for the prevention and intervention of coronary artery disease, stroke, atherosclerosis, and other diseases of which lipoproteins are important risk factors.

The sampling scheme of the HCHS/SOL was complex, and the study participants in different sampling units had unequal selection probabilities.[14] Using sampling weights to handle the complex sampling design would enable us to generalize the effect estimates to the target population, providing more reliable inference on the effects of interest.[32] As a future direction, we will develop a weighted version of the proposed method to accommodate different sampling strategies.

Data and code availability

Data from the HCHS/SOL are available at https://sites.cscc.unc.edu/hchs/ upon request. The R package we developed for our proposed method is available at https://github.com/OSylli/MVMRIE.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2024.100338.

Acknowledgments

The authors are grateful for the support from the National Heart, Lung, and Blood Institute (NHLBI) under R01HL143885 and the National Human Genome Research Institute under R01HG009974 to develop this methodology. The authors are also grateful for funding from the National Institute on Aging under P30AG066615 and the Population Architecture using Genomics and Epidemiology Study under R01HG010297. The authors thank the staff and participants of HCHS/SOL for their important contributions (investigators' website: https://sites.cscc.unc.edu/hchs/). The HCHS/SOL is a collaborative study supported by contracts from the NHLBI to the University of North Carolina (HHSN268201300001I/N01-HC-65233), University of Miami (HHSN268201300004I/N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I/N01-HC-65235), University of Illinois at Chicago (HHSN268201 300003I/N01-HC-65236 Northwestern University), and San Diego State University (HHSN268201300005I/N01-HC-65237). The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Support for HCHS/SOL metabolomics data was graciously provided by the JLH Foundation (Houston, Texas). The authors also thank the Trans-Omics for Precision Medicine (TOPMed) program imputation panel (v.TOPMed-r2), which is supported by the NHLBI (see www.nhlbiwgs.org). The panel was constructed and implemented by the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1; contract HHSN268201800002I). The TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I) provided additional data management, sample identity checks, and overall program coordination and support. We gratefully acknowledge the participants who provided biological samples and the studies that provided data for TOPMed.

Author contributions

D.Y.L. conceived the proposed method. Y.L. derived the mathematical formulas and developed the statistical programs to implement the method. Y.L., K.Y.W., and D.Y.L. designed and performed the simulation studies and verified the results. Y.L., A.G.H., P.G.L., H.M.H., M.G., K.E.N., C.G.D., C.L.A., B.Y., K.L.Y., V.L.B., R.K., L.H., B.T.J., Q.Q., T.S., and J.Y.M., acquired, processed, and analyzed the data and discussed the results. Y.L., K.Y.W., and D.Y.L. wrote the manuscript. All authors reviewed and commented on the manuscript.

Declaration of interests

The authors declare no competing interests.

Received: March 9, 2024
Accepted: July 27, 2024

Web resources

HCHS/SOL Manual 7, https://sites.cscc.unc.edu/hchs/manuals-forms

KEGG Pathway Database, https://www.genome.jp/pathway/hsa00564

OMIM, http://www.omim.org

References

1. Burgess, S., and Thompson, S.G. (2015). Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation (CRC Press).
2. Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int. J. Epidemiol. *48*, 713–727.
3. Chadeau-Hyam, M., Bodinier, B., Vermeulen, R., Karimi, M., Zuber, V., Castagné, R., Elliott, J., Muller, D., Petrovic, D., Whitaker, M., et al. (2020). Education, biological ageing, all-cause and cause-specific mortality and morbidity: UK biobank cohort study. EClinicalMedicine *29–30*, 100658.
4. Carter, A.R., Sanderson, E., Hammerton, G., Richmond, R.C., Davey Smith, G., Heron, J., Taylor, A.E., Davies, N.M., and Howe, L.D. (2021). Mendelian randomisation for mediation analysis: current methods and challenges for implementation. Eur. J. Epidemiol. *36*, 465–478.
5. Hartley, A., Sanderson, E., Granell, R., Paternoster, L., Zheng, J., Smith, G.D., Southam, L., Hatzikotoulas, K., Boer, C.G., van Meurs, J., et al. (2022). Using multivariable Mendelian randomization to estimate the causal effect of bone mineral density on osteoarthritis risk, independently of body mass index. Int. J. Epidemiol. *51*, 1254–1267.
6. Little, R.J., and Rubin, D.B. (2020). Statistical Analysis with Missing Data, 3rd Edition (John Wiley & Sons).
7. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Stat. Med. *27*, 1133–1163.
8. Nowak, C., Sundström, J., Gustafsson, S., Giedraitis, V., Lind, L., Ingelsson, E., and Fall, T. (2016). Protein biomarkers for insulin resistance and type 2 diabetes risk in two large community cohorts. Diabetes *65*, 276–284.
9. Li, L., Huang, L., Huang, S., Luo, X., Zhang, H., Mo, Z., Wu, T., and Yang, X. (2020). Non-linear association of serum molybdenum and linear association of serum zinc with nonalcoholic fatty liver disease: Multiple-exposure and Mendelian randomization approach. Sci. Total Environ. *720*, 137655.

10. Lin, D.Y., Zeng, D., and Couper, D. (2020). A general framework for integrative analysis of incomplete multiomics data. Genet. Epidemiol. *44*, 646–664.
11. Burgess, S., and Thompson, S.G. (2015). Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am. J. Epidemiol. *181*, 251–260.
12. Davidson, R., and MacKinnon, J.G. (1993). Estimation and Inference in Econometrics, *63*.
13. Sorlie, P.D., Avilés-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 629–641.
14. LaVange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 642–649.
15. Burgess, S., and Thompson, S.G. (2021). Mendelian Randomization: Methods for Causal Inference Using Genetic Variants (CRC Press).
16. Lin, Z., Xue, H., and Pan, W. (2023). Robust multivariable Mendelian randomization based on constrained maximum likelihood. Am. J. Hum. Genet. *110*, 592–605.
17. Stamler, C.J., Breznan, D., Neville, T.A., Viau, F.J., Camlioglu, E., and Sparks, D.L. (2000). Phosphatidylinositol promotes cholesterol transport *in vivo*. J. Lipid Res. *41*, 1214–1221.
18. Cole, L.K., Vance, J.E., and Vance, D.E. (2012). Phosphatidylcholine biosynthesis and lipoprotein metabolism. Biochim. Biophys. Acta *1821*, 754–761.
19. Feofanova, E.V., Chen, H., Dai, Y., Jia, P., Grove, M.L., Morrison, A.C., Qi, Q., Daviglus, M., Cai, J., North, K.E., et al. (2020). A genome-wide association study discovers 46 loci of the human metabolome in the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *107*, 849–863.
20. Wu, J., Province, M.A., Coon, H., Hunt, S.C., Eckfeldt, J.H., Arnett, D.K., Heiss, G., Lewis, C.E., Ellison, R.C., Rao, D.C., et al. (2007). An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. BMC Genet. *8*, 60–69.
21. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
22. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
23. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.
24. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.
25. Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., University of Washington Center for Mendelian Genomics, Nickerson, D.A., and Below, J.E. (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. *95*, 553–564.
26. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *98*, 165–184.
27. Stock, J., and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Identification and Inference for Econometric Models, D.W. Andrews, ed. (Cambridge University Press).
28. Pierce, B.L., Ahsan, H., and VanderWeele, T.J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. Int. J. Epidemiol. *40*, 740–752.
29. Palmer, T.M., Lawlor, D.A., Harbord, R.M., Sheehan, N.A., Tobias, J.H., Timpson, N.J., Davey Smith, G., and Sterne, J.A.C. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. Stat. Methods Med. Res. *21*, 223–242.
30. Burgess, S., and Thompson, S.G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. Int. J. Epidemiol. *42*, 1134–1144.
31. Li, Y., Wong, K.Y., Howard, A.G., Gordon-Larsen, P., Highland, H.M., Graff, M., North, K.E., Downie, C.G., Avery, C.L., Yu, B., et al. (2024). Mendelian Randomization with Incomplete Measurements on the Exposure in the Hispanic Community Health Study/Study of Latinos. HGG Adv. *5*, 100245.
32. Lin, D.Y., Tao, R., Kalsbeek, W.D., Zeng, D., Gonzalez, F., 2nd, Fernández-Rhodes, L., Graff, M., Koch, G.G., North, K.E., and Heiss, G. (2014). Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *95*, 675–688.