HGG
Advances

ARTICLE

Mendelian randomization with incomplete measurements on the exposure 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

Mendelian randomization has been widely used to assess the causal effect of a heritable exposure variable on an outcome of interest, using genetic variants as instrumental variables. In practice, data on the exposure variable can be incomplete due to high cost of measurement and technical limits of detection. In this paper, we propose a valid and efficient method to handle both unmeasured and undetectable values of the exposure variable in one-sample Mendelian randomization analysis with individual-level data. We estimate the causal effect of the exposure variable on the outcome using maximum likelihood estimation and develop an expectation maximization algorithm for the computation of the estimator. Simulation studies show that the proposed method performs well in making inference on the causal effect. We apply our method to the Hispanic Community Health Study/Study of Latinos, a community-based prospective cohort study, and estimate the causal effect of several metabolites on phenotypes of interest.

Introduction

Mendelian randomization (MR) is a technique of using genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome with observational data. Unlike conventional association analysis, MR analysis with properly selected IVs can provide valid causal inference even in the presence of unmeasured confounders and reverse causation.[1] In practice, MR analysis is often complicated by incomplete data on the exposure variable. In particular, the exposure data may be collected only on a subset of study subjects due to high cost of measurement or degraded samples. In addition, measurements of quantitative omics features, such as metabolite levels, may be subject to detection limits, such that values beyond certain thresholds are undetectable. As one of the major goals of omic studies is to assess the causal effects of quantitative omic variables on phenotypes of interest, a rigorous MR approach that properly accounts for incomplete exposure data is needed.

It is a common practice to exclude individuals with unmeasured values of the exposure variable from MR analysis.[2,3] However, this approach results in a loss of information and causes bias in parameter estimation when data are not missing completely at random.[4] Pierce and Burgess[5] proposed a subsample estimator for MR analysis when data on the exposure are collected only for a subset of participants. Specifically, the causal effect estimator is the ratio between the estimated effect of a genetic variant on the outcome (based on the entire sample) and the estimated effect of the genetic variant on the exposure (based on the complete cases only). This approach requires the missing-completely-at-random assumption and is not applicable when the exposure variable is subject to detection limits.

It is common to remove subjects with undetectable values or impute the undetectable values.[6–8] However, complete-case analysis and single imputation methods can result in biased effect estimators and inflation of type I error in hypothesis tests.[9] Multiple imputation has been used to impute unmeasured and undetectable values in association analysis,[10–12] but not in MR analysis. This approach is computationally intensive and is sensitive to the distributional assumption.[13]

In this article, we present a valid and powerful approach to MR analysis with a continuous outcome and a

[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.2023.100245.

continuous exposure, where the latter is potentially unmeasured or undetectable. We consider a linear model between the exposure and the IVs and another linear model between the outcome and the exposure. We accommodate unmeasured confounders of the exposure-outcome relationship by allowing a nonzero correlation between the error terms in the two regression models. We estimate the causal effect using maximum likelihood estimation (MLE) and develop a computationally efficient expectation maximization (EM) algorithm. Under the normality assumption, the proposed estimator is virtually unbiased and statistically efficient. We show the advantages of the proposed method over the existing methods through simulation studies. We apply the proposed method to data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).[14,15]

Material and methods

Let Y be a continuous outcome variable, S be a continuous exposure variable that is potentially unmeasured and subject to detection limits, and $\boldsymbol{G}$ be a vector of IVs for S. To be a valid IV, each component of $\boldsymbol{G}$ should satisfy the following assumptions: (1) it is associated with S; (2) it does not affect Y except through its effect on S; and (3) it is not associated with any confounders of the exposure-outcome relationship.[1] A causal diagram is given in Figure 1. Let $\boldsymbol{Z}$ be a vector of measured covariates, such as age, gender, and principal components for ancestry. In addition, let the first component of $\boldsymbol{Z}$ be the constant 1, and let $\boldsymbol{X} = (\boldsymbol{G}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$.

We consider the following pair of linear models:

$$S = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \epsilon_S, \qquad \text{(Equation 1)}$$

and

$$Y = \gamma S + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z} + \epsilon_Y, \qquad \text{(Equation 2)}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression parameters, γ represents the causal effect of the exposure variable on the outcome, and $(\epsilon_S, \epsilon_Y)^{\mathrm{T}}$ is a zero-mean bivariate normal random vector, with $\mathrm{Var}(\epsilon_S) = \sigma_S^2$, $\mathrm{Var}(\epsilon_Y) = \sigma_Y^2$, and $\mathrm{Corr}(\epsilon_S, \epsilon_Y) = \rho$. The joint density function of (Y, S) given $(\boldsymbol{X}, \boldsymbol{Z})$ is

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

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \gamma, \sigma_S^2, \sigma_Y^2, \rho)^{\mathrm{T}}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of $(\epsilon_S, \epsilon_Y)^{\mathrm{T}}$. When the normality assumption is in doubt, we perform the inverse-normal transformation on the outcome and the measured values of the exposure.

We use R to indicate, by the values one versus zero, whether the measurement of S is exact or incomplete, respectively. When $R = 0$, the exposure variable S is only known to belong to an interval C, where $C = (-\infty, L)$ if S is below the lower detection limit L, $C = (U, \infty)$ if S is above the upper detection limit U, and $C = (-\infty, \infty)$ if S is not measured. Although measurements of the exposure variable are usually non-negative, we allow negative values to accommodate situations in which standardization, log-transformation, or other transformations are performed. The detection limits are allowed to be subject-specific. For a sample of size n, the observed data consist of $\{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, R_i, R_iS_i + (1 - R_i)C_i\}$ for $i = 1, \ldots, n$.

We assume that R_i, L_i, and U_i are random variables whose joint distribution does not contain information about $\boldsymbol{\theta}$. We impose the missing-at-random assumption on R_i. We further assume that data on the outcome, genotypes, and measure covariates are complete. The observed-data likelihood function for $\boldsymbol{\theta}$ is proportional to

$$\prod_{i=1}^{n}\left\{f(Y_i, S_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})^{R_i}\left[\int_{s \in C_i} f(Y_i, s|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})ds\right]^{1-R_i}\right\},$$

where the integration is taken over all possible values of the missing S_i. To maximize this likelihood function, we adopt the EM algorithm with S treated as potentially missing data. The complete-data log likelihood function is

$$\ell(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \log f(Y_i, S_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}).$$

In the E-step, we evaluate the conditional expectation of $\ell(\boldsymbol{\theta})$ given the observed data at the current parameter estimates:

$$\begin{aligned}
\widehat{\ell}(\boldsymbol{\theta}) = & -n\log(2\pi) - \frac{n}{2}\log(\sigma_S^2) - \frac{n}{2}\log(\sigma_Y^2) - \frac{n}{2}\log(1-\rho^2) \\
& - \frac{1}{2(1-\rho^2)\sigma_S^2}\sum_{i=1}^{n}\left[\widehat{E}(S_i^2) + (\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)^2 - 2\widehat{E}(S_i)(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\right] \\
& - \frac{1}{2(1-\rho^2)\sigma_Y^2}\sum_{i=1}^{n}\left[\gamma^2\widehat{E}(S_i^2) + (Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)^2\right. \\
& \left. - 2\gamma\widehat{E}(S_i)(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)\right] \\
& + \frac{\rho}{(1-\rho^2)\sigma_S\sigma_Y}\sum_{i=1}^{n}\left[-\gamma\widehat{E}(S_i^2) - (\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)\right. \\
& \left. + \widehat{E}(S_i)(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i + \gamma\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\right], \qquad \text{(Equation 3)}
\end{aligned}$$

where $\widehat{E}$ denotes the conditional expectation given the observed data, evaluated at the current parameter estimates. In the M-step, we update the parameters with the maximizer of the expected complete-data log likelihood function. We iterate between the E-step and the M-step until the Euclidean distance between the parameter values at two successive iterations is less than a pre-specified small positive constant (e.g., 10^{-5}). The resulting estimator of $\boldsymbol{\theta}$ is denoted by $\widehat{\boldsymbol{\theta}}$. Details of the E-step and the M-step are given in Appendix A.

The estimated covariance matrix of $\widehat{\boldsymbol{\theta}}$ is derived using the Louis formula.[4] We first compute the complete-data information matrix $\mathbf{I}_{\mathrm{c}}$, which is the negative of the Hessian matrix of $\widehat{\ell}(\boldsymbol{\theta})$ evaluated at $\widehat{\boldsymbol{\theta}}$. In addition, we calculate the gradient of $\log f(Y_i, S_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, denoted by $\boldsymbol{U}_i$, and evaluate the conditional expectations of $\boldsymbol{U}_i$ and $\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}$ given the observed data at $\widehat{\boldsymbol{\theta}}$. We finally obtain the observed-data information matrix

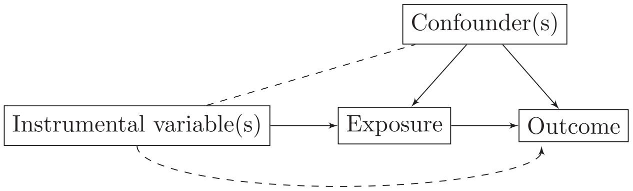

Figure 1. The causal diagram for MR analysis, where a set of genetic variants are used as IVs to estimate the causal effect of an exposure on an outcome with observational data

A dashed arrow from the IV(s) to the outcome means that there should not be any direct effect(s) of the IV(s) on the outcome. Similarly, a dashed line between the IV(s) and the confounder(s) means that the IV(s) should not be associated with any confounder(s) of the exposure-outcome relationship.

$$\mathbf{I}_{\text{obs}} = \mathbf{I}_{\text{c}} - \sum_{i=1}^{n} I_{(R_i=0)}\left[\widehat{E}\left(\boldsymbol{U}_i\boldsymbol{U}_i^{\text{T}}\right) - \widehat{E}(\boldsymbol{U}_i)\widehat{E}(\boldsymbol{U}_i)^{\text{T}}\right], \qquad \text{(Equation 4)}$$

where the second term on the right-hand side is the conditional covariance matrix of the gradient of $\ell(\boldsymbol{\theta})$ given the observed data. Then, the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $\mathbf{I}_{\text{obs}}^{-1}$; see Appendix A for details.

Simulation studies

We conducted extensive simulation studies to compare the proposed method with existing ones. We set $Z_1 = 1$ and independently generated Z_2 from the standard normal distribution, Z_3 from the Bernoulli distribution with 0.5 success probability, and Z_4 from the standard uniform distribution; Z_2, Z_3, and Z_4 represent the first principal component for ancestry, gender, and (normalized) age, respectively. We generated the IV (i.e., G) from the $\text{Binomial}(2, p)$ distribution, which represents the genotype of a genetic variant with minor allele frequency p under the Hardy-Weinberg equilibrium; we set $p = e^{0.5Z_2}/(1+e^{0.5Z_2})$ to create population stratification. Then, we generated the exposure variable S from model (1), where $\boldsymbol{X} = (G, \boldsymbol{Z}^{\text{T}})^{\text{T}}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_4)^{\text{T}}$, and generated the outcome variable Y from model (2).

For the parameter values, the genetic effect on the exposure was set to 0.25; the intercepts in both models (1) and (2) were set to 1; the coefficients of Z_2, Z_3, and Z_4 in both models (1) and (2) were set to 0.5; γ was set to 0 or 0.25; σ_S^2 and σ_Y^2 were both set to 1; and ρ was set to vary from 0 to 0.5 with a 0.1 increment. The choice of $\gamma = 0.25$ represents a moderate effect of the exposure on the outcome. We set the sample size n to be 9,000 and let 70% of the exposure values be unmeasured completely at random. We altered the lower detection limit for the exposure values from -1 to 1 with an increment of 0.1; we assumed that all the individuals had the same lower detection limit and that there was no upper detection limit.

The value of $\boldsymbol{\alpha}$ was selected such that when there was no lower detection limit, the mean of the 2.5% and the 97.5% quantiles of the measured exposure variable were -0.78 and 4.28, respectively. As the lower detection limit increased from -1 to 1, the proportion of subjects (within the subsample consisting of individuals with measured exposure) with undetectable exposure values increased from 0% to 30.3%, which covered the situations in the HCHS/SOL data. To measure the strength of the IV, we evaluated the (partial) F-test statistic for the coefficient of G in model (1) based on individuals with available S. As the lower detection limit increased, the number of complete cases became smaller, resulting in a decrease in the mean F-statistic from 76.6 to 35.3; however, the selection of the genetic effect size ensured that the IV remained sufficiently strong since the F-statistics were greater than 10.[16]

We considered four existing methods: complete-case analysis, "imputation at limit," "imputation at mid-point," and multiple imputation. For complete-case analysis, we removed all subjects with unmeasured or undetectable exposure values and estimated the parameters using standard MLE. For the single imputation methods, we imputed the measurements below the detection limit L by L for the "imputation at limit" method and by $L - \log 2$ for the "imputation at mid-point" method; in addition, we removed subjects with unmeasured exposure values. We then performed standard MLE on the resulting data. Here, we assumed S to be the log-transformation of the original measurement. Thus, for the imputation at mid-point method, the imputed value is the log of the mid-point between e^L (the lower detection limit on the original scale) and 0 (the lower bound of the exposure values). For multiple imputation, we imputed each missing value 20 times from the distribution in Equation 5. We calculated the MLE on each complete dataset and obtained the combined estimate, standard error estimate, and p value.[4] Estimates from the imputation at mid-point method were used as parameter values in Equation 5 for multiple imputation and as initial values for the EM algorithm in the proposed method. For each method, we performed the Wald test on the null hypothesis of $\gamma = 0$ at the nominal significance level of 0.001. We simulated 10,000 and 10 million replicates for $\gamma = 0.25$ and $\gamma = 0$, respectively.

Figures 2, S1, and S2 show the simulation results for the scenario of $\gamma = 0.25$. The proposed causal effect estimator is nearly unbiased, the proposed standard error estimator is accurate, and the 95% confidence interval has adequate empirical coverage probability. The complete-case analysis yields a nearly unbiased causal effect estimator when ρ is zero (i.e., when there are no unmeasured confounders), but the estimator has a much larger standard error than the proposed estimator, so the complete-case analysis has substantially lower power for testing γ than the proposed method. When ρ is nonzero, the complete-case analysis yields a negatively biased estimator; as the lower detection limit increases, the bias becomes more severe, the standard error increases, and the power for the test on γ decreases. The bias of the causal effect estimator from the imputation at limit method is away from the null and becomes more severe as the lower detection limit increases. The causal effect estimator from the imputation at mid-point method is also biased, although the magnitude is smaller than that of

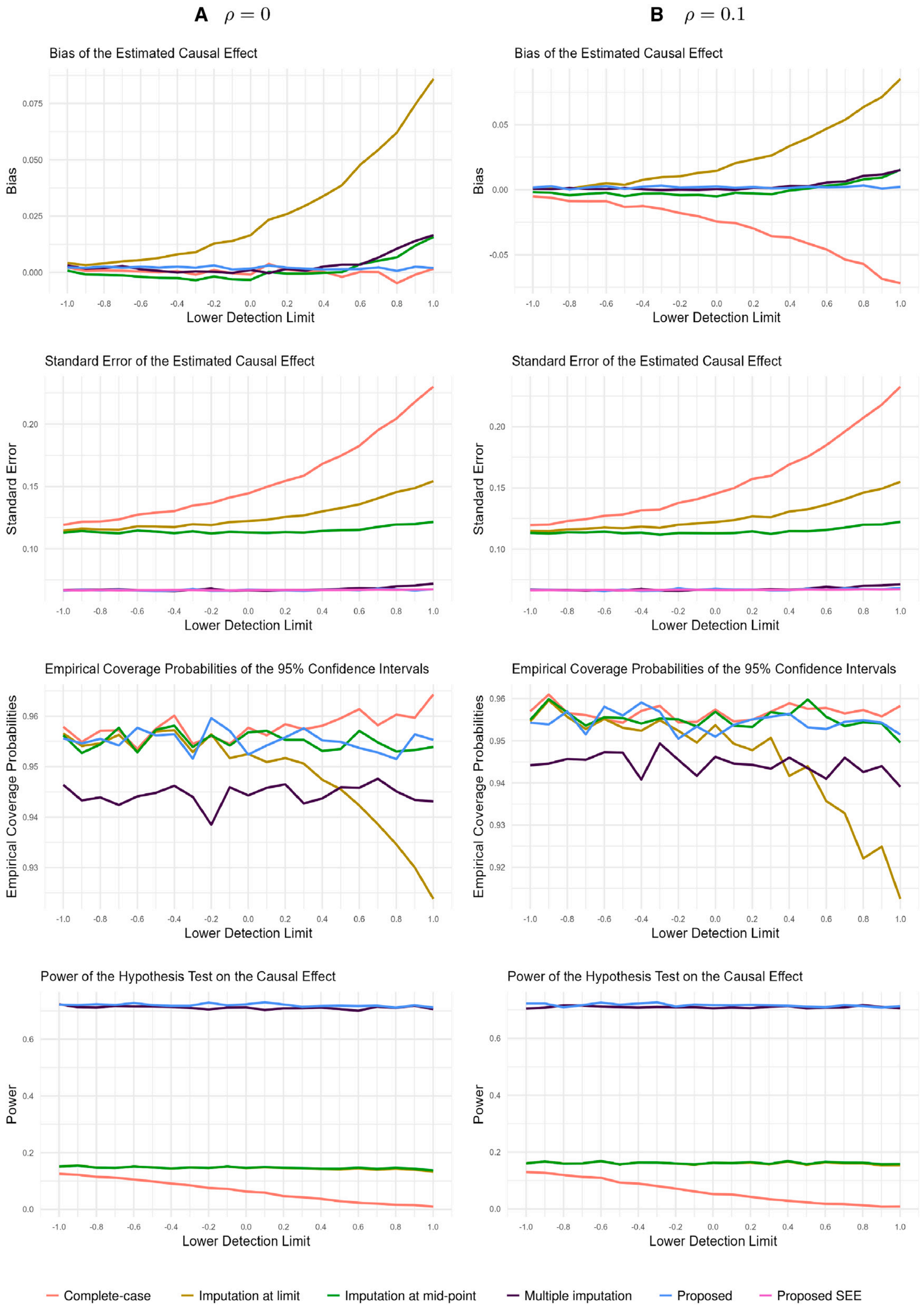

Figure 2. Simulation results for the scenario where the true value of the causal effect γ is 0.25

Panels A and B correspond to the cases where the correlation is 0 and 0.1, respectively. Bias and standard error of the causal effect estimators, the empirical coverage probabilities of the 95% confidence intervals, and the empirical power of the significance test on γ are plotted against the lower detection limit of the exposure variable. The red, brown, green, purple, and blue curves correspond to the complete-case analysis, the imputation at limit method, the imputation at mid-point method, multiple imputation, and the proposed method, respectively. The pink curve represents the mean of the standard error estimator (SEE) given by the proposed method.

the imputation at limit estimator. The two single imputation methods yield larger standard errors and thus are less powerful in testing γ than the proposed method. Multiple imputation performs similarly to the proposed method when the lower detection limit is small; however, when the lower detection limit is greater than 0.5, under which the proportion of undetectable values reaches about 17.8%, the effect estimator becomes positively biased, and its standard error becomes larger than that of the proposed estimator.

Figures 3, S3, and S4 show the simulation results for the scenario of $\gamma = 0$. For the proposed method, the conclusions regarding the bias of the causal effect estimator, standard error estimation, and coverage of confidence intervals are the same as those in the case of $\gamma = 0.25$, but the empirical type I error tends to be below the nominal significance level. The complete-case analysis yields a virtually unbiased causal effect estimator when ρ is zero. However, when ρ is nonzero, the complete-case estimator is biased, and the bias becomes more severe as ρ or the detection limit increases. Compared with the proposed estimator, the complete-case estimator also has a higher standard error, which increases drastically as the lower detection limit increases. The two single imputation approaches yield nearly unbiased estimators of γ, but the estimators have higher standard errors than the proposed estimator. The empirical type I errors yielded by the two single imputation methods are similar but larger than the type I error of the proposed method, exceeding the nominal level for $\rho \geq 0.4$. Finally, multiple imputation provides a nearly unbiased causal effect estimator but yields an inflated type I error when $\rho = 0.5$.

We performed additional simulation studies to evaluate the robustness of each method to the non-normality of $(\epsilon_S, \epsilon_Y)^{\mathrm{T}}$. In the first scenario, we set $\epsilon_S = (T_S + T_0)/\sqrt{3}$ and $\epsilon_Y = (T_Y + T_0)/\sqrt{3}$, where T_S, T_Y, and T_0 followed independent *t*-distributions with 6 degrees of freedom. In the second scenario, we let $\epsilon_S = (\Gamma_S + \Gamma_0 - 16)/4\sqrt{2}$ and $\epsilon_Y = (\Gamma_Y + \Gamma_0 - 16)/4\sqrt{2}$, where Γ_S, Γ_Y, and Γ_0 independently followed the Gamma distribution with a shape parameter of 4 and a scale parameter of 2. The constants (i.e., $\sqrt{3}$, 16, and $4\sqrt{2}$) were incorporated so that ϵ_S and ϵ_Y had means of zero and variances of one. The other simulation parameters remained unchanged. When the lower detection limit increased from -1 to 1, the proportion of individuals with undetectable exposure increased from 0% to 31.1%. We performed the inverse-normal transformation on the outcome and the measured values of the exposure. The corresponding parameter should be interpreted as the effect of the transformed exposure on the transformed outcome rather than the effect on the original scale, and the two effect sizes are generally different when the true effect on the original scale is nonzero.

The results are shown in Figures S5 and S6. When the true causal effect is nonzero, the proposed method remains more powerful than complete-case analysis and single imputation; multiple imputation has similar power to the proposed method when the proportion of undetectable values is small but is less powerful when more than 14.3% of the exposures are undetectable. However, the estimators no longer reflect the effect size on the original scale due to the inverse-normal transformation. When the true causal effect is zero, the type I errors of the imputation methods are inflated, whereas the type I errors yielded by the proposed method and complete-case analysis remain below the nominal level; in addition, the proposed estimator is almost unbiased.

The simulation results show that removing the undetectable values of the exposure (which is an endogenous variable in our models) or imputing the undetectable values with fixed values generally leads to biased causal effect estimators. In addition, since the complete-case analysis and the two single imputation methods remove a large proportion of subjects, they yield much larger standard errors and are substantially less powerful than the proposed method in detecting a causal effect. Multiple imputation also performs worse than the proposed method when the proportion of the undetectable values is large or when the normality assumption does not hold, and it is computationally less efficient than the proposed method.

Under the normality assumption, the proposed method yields a nearly unbiased causal effect estimator, with accurate standard error estimation, but the empirical type I error under the null hypothesis tends to be below the nominal level. By inspecting the empirical distributions of the *z*-values over the replicates, where the *z*-value is defined as the ratio of the causal effect estimate to the standard error estimate, we found that the deflation of type I error is attributed to the thin-tailed distributions of the *z*-values. For scenarios where $\rho \geq 0.2$, the complete-case analysis yields a severely biased causal effect estimator when the lower detection limit is large, but the corresponding empirical type I error remains deflated. This is because the distributions of the *z*-values are highly leptokurtic, and less than 0.1% of the values exceed the range between the 0.05% and the 99.95% quantiles of the standard normal distribution. When $\rho \geq 0.4$, the two single imputation approaches yield inflated type I errors since the distributions of the *z*-values are positively skewed and have heavy right tails; similar phenomena were observed for the multiple imputation method when $\rho = 0.5$.

Application to the HCHS/SOL

In this section, we use the proposed method to evaluate the causal effects of the metabolites of interest on estimated glomerular filtration rate (eGFR) and some lipoprotein outcomes using data from the HCHS/SOL. Feofanova et al.[17] conducted a genome-wide association study and discovered some previously unreported single nucleotide polymorphism (SNP)-metabolite associations in the HCHS/SOL; the authors also reported some metabolite-phenotype pairs with an absolute Pearson's correlation greater than 0.1. We used the significant SNPs identified

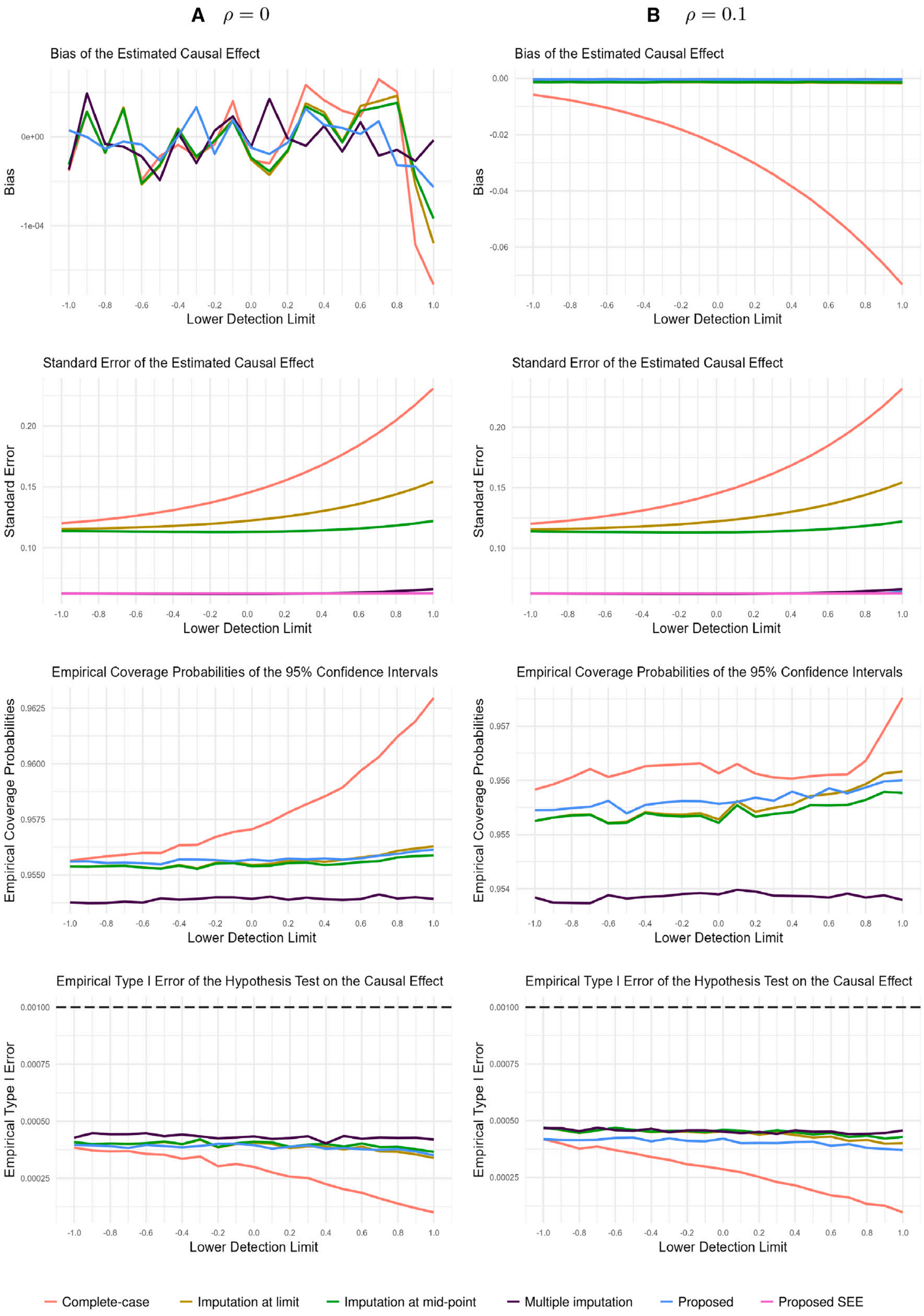

Figure 3. Simulation results for the scenario where the true value of the causal effect γ is 0

Panels A and B correspond to the cases where the correlation is 0 and 0.1, respectively. Bias and standard error of the causal effect estimators, the empirical coverage probabilities of the 95% confidence intervals, and the empirical type I error of the significance test on γ are plotted against the lower detection limit of the exposure variable. The red, brown, green, purple, and blue curves correspond to the complete-case analysis, the imputation at limit method, the imputation at mid-point method, multiple imputation, and the proposed method, respectively. The pink curve represents the mean of the standard error estimator (SEE) given by the proposed method. The black dashed line indicates the nominal significance level of 0.001.

Table 1. **SNP-metabolite-phenotype combinations of interest**

SNP	Metabolite	Phenotype
rs1047891	isovalerylglycine	eGFR
rs1047891	isobutyrylglycine	eGFR
rs10889335	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	TC
		TG
		LDL-C
rs174554	1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	TC
rs247617	1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	HDL-C
		LDL-C
		TC
		TG

eGFR, estimated glomerular filtration rate; TC, total cholesterol; TG, triglycerides; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol.

by the authors as the IVs and performed MR analyses on these metabolite-phenotype pairs.

HCHS/SOL

The HCHS/SOL is a multicenter community-based prospective cohort study of US Hispanic/Latino individuals. A stratified two-stage area probability sampling design was adopted in the recruitment process, so that the selected individuals would represent the target population.[15] In total, 16,415 individuals between 18 and 74 years old who self-identified as Hispanic or Latino were recruited from communities in the Bronx, Chicago, Miami, and San Diego.[14] The study included participants from the following backgrounds: Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American.

Participants' fasting blood specimens were collected at each recruitment center with a standardized protocol, and daily fresh and frozen specimens were sent to a central laboratory for further measurements.[14] The serum was separated from the cells within 2 h of collection and stored at $-\ 70^\circ$C until assayed. Serum and plasma metabolite measurements were available only for a randomly selected subset of the study participants, and only data from the participants' baseline visit were used in our analysis. The HCHS/SOL study was approved by institutional review boards at participating institutions. Written informed consent was obtained from all participants. More details on the laboratory procedures and measurements are described in the HCHS/SOL Manual 7 (Addendum), available at https://sites.cscc.unc.edu/hchs/manuals-forms.

Selection of outcomes, exposures, and IVs

We focused on 10 of the 15 metabolite-phenotype pairs that were reported to have an absolute Pearson's correlation greater than 0.1: we excluded three pairs with a binary phenotype; we excluded two more pairs involving urine creatinine because the measurements can fluctuate depending on diet, hydration, and other factors[18] and because a single observation is usually insufficient to reflect the actual level of urine creatinine. In Feofanova et al.,[17] one SNP was reported to be significantly associated with each of the 10 metabolites used in our analysis. The metabolite-phenotype pairs of interest and the corresponding SNP for each metabolite are shown in Table 1, and details on the SNP-metabolite associations reported by Feofanova et al.[17] are provided in Table S1. Throughout our analysis, participants with chronic kidney disease were excluded.[19]

Data collection and processing

Total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and creatinine were measured in serum on a Roche Modular P Chemistry Analyzer (Roche Diagnostics Corporation), and eGFR was computed based on serum creatinine level, age, and sex, using the method described by Inker et al.[20] We removed individuals with missing outcome measurements. In the analyses of the lipid outcomes, we also excluded non-fasting participants and corrected the outcome values for individuals using statins (which are lipid-lowering medications) based on the findings of Wu et al.[21]

Stored serum samples were used for metabolomic profiling, which was conducted in 2017 at Metabolon Inc. (Durham, NC) with the Metabolon Discovery HD4 platform, and serum metabolites were quantified with an untargeted, liquid chromatography-mass spectrometry (LC-MS)-based method.[22,23] For each of the metabolites included in our analysis, there are values below the lower detection limit. The minimum measurement among the individuals whose metabolites were measured and detectable was considered to be the (intrinsic) lower detection limit, denoted by L_0.

Outliers exist in the positively skewed metabolite data, and we set them as being beyond certain detection limits before data analysis. Specifically, we first computed the sample mean and sample standard deviation of the log-transformed measurements among the measured and detectable values, denoted by m and s, respectively. Then, we introduced an upper detection limit of $\exp(m+3s)$ and redefined the lower detection limit as $\max\{\exp(m\ -\ 3s), L_0\}$. Thus, the log-scale measurements that were three sample standard deviations above and below the sample mean were labeled as being above the upper detection limit and below the lower detection limit, respectively.

The HCHS/SOL participants were a subset of the individuals in the Population Architecture using Genomics and Epidemiology (PAGE) study; in the course of that study, the participants were genotyped on the MultiEthnic Genotyping Array at the Center for Inherited Disease Research.[24] Four SNPs (rs1047891, rs10889335, rs174554, and rs247617) were used in our analysis, and the imputation was performed with the TOPMed imputation server using the TOPMed freeze 8 imputation reference panel.[25–27] We obtained genotypic data on 11,604 HCHS/SOL participants for each SNP and excluded individuals with unavailable genetic data. To account for the population structure, we also derived the principal components for ancestry and incorporated them into our model as covariates.

Complete data on age and gender were collected during individuals' baseline visits. The Hispanic/Latino background (Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American) of the individuals was derived using the method of Conomos et al.[28] We included the Hispanic/Latino background variable in the analysis since it is associated with lifestyle factors, such as diet, that affect metabolite values.

Statistical analysis

We performed the inverse-normal transformation on the outcome variables and the measured values of the exposures, such that the measured and detectable metabolite values approximately followed a truncated normal distribution. Then, we used the proposed method to assess the causal effects of the transformed metabolite levels on the transformed outcomes, using the corresponding SNP in Table 1 as an IV and using age, gender, the center of recruitment, Hispanic/Latino background, and the first five principal components for ancestry as the measured covariates. We also employed multiple imputation and the imputation at mid-point method for comparisons. For imputation at mid-point, individuals with unmeasured metabolites were removed, and values above the upper detection limit or below the lower detection limit were imputed with the upper detection limit or half of the lower detection limit on the original scale, respectively.

We performed a preliminary assessment of the IV assumptions. Similar to the simulation studies, we assessed the IV strength by calculating the (partial) *F*-test statistic for the coefficient of the genetic variant in model (1) using individuals with available metabolite data. The validity of the IVs is more difficult to verify. There has been little literature on the reliable assessment of horizontal pleiotropy when a single SNP is used and a large proportion of the measurements on the exposure variable are unavailable; we discuss this issue in Section discussion.

Endogamous mating exists in the Hispanic/Latino community and induces genetic relatedness in the HCHS/SOL participants. As a sensitivity analysis, we used Pedigree Reconstruction and Identification of a Maximum Unrelated Set (PRIMUS) to identify the maximum unrelated subset of individuals such that $\widehat{\pi} \leq 0.2$ for any individual pairs, where $\widehat{\pi}$ is the estimated proportion of alleles shared identical by descent based on identity by state and allele frequency.[29] We then repeated the statistical analysis using only the unrelated individuals.

Results

Descriptive statistics of the phenotypes, metabolites, and covariates for the 11,604 individuals with available genetic data are shown in Table S2. In addition, for each metabolite-phenotype pair of interest, the number of individuals with unmeasured or undetectable exposure values is shown in Table S3. In summary, the metabolite is unmeasured in about two-thirds of the individuals, and the proportion of undetectable metabolite values varies from 0.13% to 5.72%.

The proposed method detected a significant ($p < 0.05$) causal relationship for all 10 metabolite-phenotype pairs (Table 2). Both isovalerylglycine and isobutyrylglycine were found to have a negative causal effect on eGFR. The metabolite 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) had strong and positive causal effects on TC, TG, and LDL-C. A positive causal effect of 1-palmitoyl-2-stearoyl-GPC (16:0/18:0) on TC was also identified. The effects of 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) on the lipid phenotypes were found to be in opposite directions, with positive causal effects on HDL-C and TC and negative causal effects on LDL-C and TG. After applying the Bonferroni correction, which produced an adjusted significance level of $0.05/10 = 0.005$, most of the causal effects detected by the proposed method remained significant, except for the effects of 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) on TC and LDL-C.

The results from multiple imputation were close to those of the proposed method. This was consistent with the simulation results that the two methods performed similarly when the proportion of undetectable values was small. The effect estimates from the imputation at mid-point method were similar to the proposed estimates, and the estimated causal effects from the two methods had the same directions. However, the imputation at mid-point method generated wider confidence intervals and larger p values, reflecting lower statistical efficiency than the proposed method. At the significance level of 0.05, the effect of 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4) on LDL-C and the effects of 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) on TC and LDL-C were all reported to be insignificant by the imputation at mid-point method, while significant causal relationships were identified by the proposed method for these three pairs. Results given by the complete-case analysis and the imputation at limit method are shown in Table S4.

In the assessment of IV strength, the *F*-statistics in the 10 sets of analyses ranged from 23.7 to 80.4 and were greater than the rule-of-thumb value of 10,[16] so the IVs were strong. For IV validity, we recognized that potential pleiotropic effects might lead to bias in the causal effect estimation.

In the sensitivity analysis, the number of excluded individuals ranged from 1,300 to 1,400. The effect directions yielded by different methods remained the same as those in the main analyses. However, with a smaller sample size, the confidence intervals were wider. As a result, the proposed method failed to identify a significant causal relationship between 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) and TC, and the causal effect of isobutyrylglycine on eGFR given by the imputation at mid-point method became insignificant. The causal effects for the other

Table 2. Estimated causal effect of the metabolite on the phenotype for each pair of interest

			Proposed method			Multiple imputation			Imputation at mid-point		
Metabolite	Phenotype	n	Est	95% CI	p value	Est	95% CI	p value	Est	95% CI	p value
isovalerylglycine	eGFR	9540	−0.171	(−0.276, −0.066)	0.001	−0.170	(−0.280, −0.060)	0.002	−0.207	(−0.385, −0.030)	0.022
isobutyrylglycine	eGFR	9540	−0.225	(−0.371, −0.080)	0.002	−0.229	(−0.378, −0.080)	0.003	−0.270	(−0.507, −0.033)	0.025
1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	TC	9382	0.543	(0.361, 0.724)	<0.001	0.552	(0.368, 0.736)	<0.001	0.460	(0.203, 0.717)	<0.001
	TG	9382	0.662	(0.463, 0.860)	<0.001	0.661	(0.463, 0.859)	<0.001	0.610	(0.355, 0.865)	<0.001
	LDL-C	9236	0.330	(0.147, 0.512)	<0.001	0.334	(0.150, 0.518)	<0.001	0.169	(−0.139, 0.477)	0.283
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	TC	9382	0.292	(0.159, 0.424)	<0.001	0.291	(0.156, 0.426)	<0.001	0.430	(0.239, 0.622)	<0.001
1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	HDL-C	9381	1.676	(1.132, 2.221)	<0.001	1.777	(1.232, 2.322)	<0.001	1.661	(1.103, 2.219)	<0.001
	LDL-C	9236	−0.275	(−0.504, −0.046)	0.019	−0.270	(−0.499, −0.041)	0.021	−0.368	(−0.769, 0.034)	0.073
	TC	9382	0.195	(0.001, 0.389)	0.049	0.199	(0.005, 0.393)	0.044	0.121	(−0.212, 0.454)	0.476
	TG	9382	−0.317	(−0.531, −0.103)	0.004	−0.325	(−0.541, −0.109)	0.003	−0.361	(−0.693, −0.030)	0.033

n, sample size; Est, causal effect estimate; CI, confidence interval; eGFR, estimated glomerular filtration rate; TC, total cholesterol; TG, triglycerides; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol.

metabolite-phenotype pairs remained significant. Finally, the IVs remained strong in the sensitivity analysis.

Discussion

In this article, we consider MR analysis with a continuous exposure variable that may be unobserved or beyond detection limits. We propose an MLE approach that incorporates both the outcome model and the exposure model into the likelihood function, so the estimator is valid and statistically efficient. Extensive simulation studies demonstrate that under the normality assumption, the proposed method provides reliable inference on the causal effect of interest even when exact exposure values are only available for a small subset of individuals. The proposed method also performs better than existing methods when the normality assumption does not hold. In addition to incomplete measurements, the proposed method can deal with outliers in the exposure data, as in the analysis of the HCHS/SOL data: values above or below certain thresholds can be treated as being above the upper or below the lower detection limit before the proposed method is applied. The common practice of excluding outliers from analysis or winsorizing may introduce bias and reduce power.

The proposed method involves the use of the EM algorithm, which requires initial parameter estimates. The EM algorithm may be sensitive to the choice of initial values, and it may fail to converge if the initial values are far away from the optimizer of the target function. Using the estimates from the other existing methods is one possible way to select the initial parameter values. Multiple imputation performs well in the simulation studies, but it is time-consuming. Among the other methods, we have shown that the causal effect estimate from the imputation at mid-point method is closer to the true value than that from the complete-case analysis and imputation at limit. Therefore, we recommend using the imputation at mid-point method to compute the initial values for the EM algorithm. With this choice of initial values, the proportion of non-convergence was less than 0.1% in the simulation studies. The solution from the EM algorithm may be a local rather than the global maximizer of the observed-data likelihood function. We suggest running the EM algorithm with different initial values and choosing the solution with the highest likelihood value.

The HCHS/SOL data example involved only a single IV for each exposure variable, and thus we merely conducted simulation studies under the single-IV scenario. However, the proposed framework accommodates a multivariate $\boldsymbol{G}$, so the proposed method is still applicable when there are multiple IVs. When the number of IVs increases, it takes longer for the proposed EM algorithm to converge. Thus, we recommend combining multiple IVs into genetic risk scores before applying the proposed method, which not only leads to higher time efficiency but also increases the first-stage (partial) F-statistic and thus reduces the weak instrument bias.[30]

The HCHS/SOL participants were selected through a stratified multistage cluster sampling design.[15] Incorporating sampling weight into analysis to account for the complex survey sampling design can resolve potential over-representation or under-representation issues so that the effect estimates can be generalized to the target population. In addition, taking into account the complex correlation structure induced by household sharing and endogamous mating within the Hispanic/Latino community can lead to more reliable inference on the effect of interest.[31] Thus, it would be valuable to extend our current method to accommodate a complicated sampling scheme with unequal selection probabilities.

Evaluating the IV assumptions is essential for an MR analysis. Pleiotropy-robust methods and the assessment of IV strength have been frequently discussed with regard to MR studies with individual-level data.[32–38] However, the approaches in these studies require that data on the exposure variable are complete. Although a preliminary assessment was performed in our analysis of the HCHS/SOL data, we have yet to develop rigorous methods for assessing the IV assumptions when exposure data are incomplete. To fill this gap, we are exploring whether the proposed method can be incorporated into existing approaches that accommodate more complex relationships between variables, such as network MR,[39] and we aim to develop reliable guidelines for assessing IV strength with incomplete exposure data.

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 the proposed method is available at https://github.com/OSylli/MRIE.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2023.100245.

Acknowledgments

The authors are grateful to the support from the National Heart, Lung, and Blood Institute (NHLBI) under R01HL143885 to develop this methodology. The authors are also grateful to 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 Hispanic Community Health Study/Study of Latinos (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 (HHSN268201300003I/N01-HC-65236 Northwestern University), and San Diego State University (HHSN268201300005I/N01-HC-65237). The following institutes/centers/offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research (NIDCR), National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, and National Institutes of Health Office of Dietary Supplements. 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 in Precision Medicine (TOPMed) program imputation panel (version TOPMed-r2), which is supported by the NHLBI; see www.nhlbiwgs.org. TOPMed study investigators contributed data to the reference panel, which can be accessed through the Michigan Imputation Server; see https://imputationserver.sph.umich.edu. 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.

Declaration of interests

The authors declare no competing interests.

Received: May 2, 2023
Accepted: October 2, 2023

Web resources

The laboratory procedures and measurements are described in the HCHS/SOL Manual 7 (Addendum), available at https://sites.cscc.unc.edu/hchs/manuals-forms.

Appendix A: Details of the proposed method

In this appendix, we present details of the EM algorithm and derive the estimated covariance matrix of the estimators. The complete-data log likelihood function is

$$\ell(\boldsymbol{\theta}) = -n\log(2\pi) - \frac{n}{2}\log\left(\sigma_S^2\right) - \frac{n}{2}\log\left(\sigma_Y^2\right) - \frac{n}{2}\log\left(1-\rho^2\right) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}\left[\frac{\left(S_i-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i\right)^2}{\sigma_S^2} + \frac{\left(Y_i-\gamma S_i-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i\right)^2}{\sigma_Y^2} - 2\rho\frac{\left(S_i-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i\right)\left(Y_i-\gamma S_i-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i\right)}{\sigma_S\sigma_Y}\right].$$

We denote the initial parameter vector by $\boldsymbol{\theta}_{(0)}$.

E-step

At the jth iteration ($j = 1, 2, \ldots$), we compute the conditional expectations of S_i and S_i^2 given the observed data at the current parameter values $\boldsymbol{\theta}_{(j-1)} \equiv \left(\boldsymbol{\alpha}_{(j-1)}^{\mathrm{T}}, \boldsymbol{\beta}_{(j-1)}^{\mathrm{T}}, \gamma_{(j-1)}, \sigma_{S(j-1)}^2, \sigma_{Y(j-1)}^2, \rho_{(j-1)}\right)^{\mathrm{T}}$.

If $R_i = 1$, then S_i is observed, $\widehat{E}(S_i) = S_i$, and $\widehat{E}(S_i^2) = S_i^2$. If $R_i = 0$, then the conditional density function of S_i given the observed data Y_i, $\boldsymbol{Z}_i$, and $\boldsymbol{X}_i$ at the current values of parameters $\boldsymbol{\theta}_{(j-1)}$, denoted by $f(s|Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, C_i; \boldsymbol{\theta}_{(j-1)})$, satisfies

$$f\left(s\middle|Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, C_i; \boldsymbol{\theta}_{(j-1)}\right) \propto \frac{1}{\sqrt{2\pi a}} \exp\left[-\frac{(s - b_i)^2}{2a}\right] I_{(s \in C_i)}, \qquad \text{(Equation 5)}$$

where

$$a = \left(1 - \rho_{(j-1)}^2\right)\left[\frac{\gamma_{(j-1)}^2}{\sigma_{Y(j-1)}^2} + \frac{1}{\sigma_{S(j-1)}^2} + \frac{2\rho_{(j-1)}\gamma_{(j-1)}}{\sigma_{S(j-1)}\sigma_{Y(j-1)}}\right]^{-1},$$

and

$$b_i = \frac{a}{1 - \rho_{(j-1)}^2}\left[\frac{\gamma_{(j-1)}\left(Y_i - \boldsymbol{\beta}_{(j-1)}^{\mathrm{T}}\boldsymbol{Z}_i\right)}{\sigma_{Y(j-1)}^2} + \frac{\boldsymbol{\alpha}_{(j-1)}^{\mathrm{T}}\boldsymbol{X}_i}{\sigma_{S(j-1)}^2} + \frac{\rho_{(j-1)}\left(Y_i - \boldsymbol{\beta}_{(j-1)}^{\mathrm{T}}\boldsymbol{Z}_i + \gamma_{(j-1)}\boldsymbol{\alpha}_{(j-1)}^{\mathrm{T}}\boldsymbol{X}_i\right)}{\sigma_{S(j-1)}\sigma_{Y(j-1)}}\right].$$

Therefore, if S_i is unmeasured, then $C_i = (-\infty, \infty)$, $\widehat{E}(S_i) = b_i$, and $\widehat{E}(S_i^2) = b_i^2 + a$. If S_i is below the lower detection limit, then $C_i = (-\infty, L_i)$, $\widehat{E}(S_i) = b_i - \sqrt{a}\varphi(L_{0i})/\Phi(L_{0i})$, and $\widehat{E}(S_i^2) = b_i^2 + a - (2b_i\sqrt{a} + aL_{0i})\varphi(L_{0i})/\Phi(L_{0i})$, where $L_{0i} = (L_i - b_i)/\sqrt{a}$, and φ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. If S_i is above the upper detection limit, then $C_i = (U_i, \infty)$, $\widehat{E}(S_i) = b_i + \sqrt{a}\varphi(U_{0i})/\Phi(U_{0i})$, and $\widehat{E}(S_i^2) = b_i^2 + a + (2b_i\sqrt{a} - aU_{0i})\varphi(U_{0i})/\Phi(U_{0i})$, where $U_{0i} = (-U_i + b_i)/\sqrt{a}$. Then we can obtain the expression of $\widehat{\ell}(\boldsymbol{\theta})$ in Equation 3.

M-step

By computing the first-order partial derivatives of $\widehat{\ell}(\boldsymbol{\theta})$ and setting them to zero, we find that for any fixed $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ, $\widehat{\ell}(\boldsymbol{\theta})$ is maximized at

$$\widehat{\sigma}_S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \frac{1}{n}\left[\boldsymbol{\alpha}^{\mathrm{T}}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)\boldsymbol{\alpha} - 2\sum_{i=1}^{n}\widehat{E}(S_i)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\alpha} + \sum_{i=1}^{n}\widehat{E}(S_i^2)\right], \qquad \text{(Equation 6)}$$

$$\widehat{\sigma}_Y^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \frac{1}{n}\left[\gamma^2\sum_{i=1}^{n}\widehat{E}(S_i^2) - 2\gamma\sum_{i=1}^{n}\widehat{E}(S_i)(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i) + \sum_{i=1}^{n}(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i)^2\right], \qquad \text{(Equation 7)}$$

and

$$\widehat{\rho}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = -\frac{1}{n\widehat{\sigma}_S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\widehat{\sigma}_Y(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)}\left[\gamma\sum_{i=1}^{n}\widehat{E}(S_i^2) + \sum_{i=1}^{n}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i) - \sum_{i=1}^{n}\widehat{E}(S_i)(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i + \gamma\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\right]. \qquad \text{(Equation 8)}$$

Plugging in the maximizers into $\widehat{\ell}(\boldsymbol{\theta})$, we have

$$\widehat{\ell}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \widehat{\sigma}_S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma), \widehat{\sigma}_Y^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma), \widehat{\rho}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\right) = -n\log(2\pi) - n - \frac{n}{2}\log\left\{\widehat{\sigma}_S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\widehat{\sigma}_Y^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\left[1 - \widehat{\rho}^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\right]\right\},$$

so it suffices to minimize $\log\{\widehat{\sigma}_S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)\widehat{\sigma}_Y^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)[1 - \widehat{\rho}^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)]\}$. At the jth iteration ($j = 1, 2, \ldots$), we find the optimizers $\boldsymbol{\alpha}_{(j)}$, $\boldsymbol{\beta}_{(j)}$, and $\gamma_{(j)}$ by applying the Newton-Raphson algorithm and then get the updated parameters $\sigma_{S(j)}^2$, $\sigma_{Y(j)}^2$, and $\rho_{(j)}$ based on Equations 6, 7, and 8; this completes the M-step.

We iterate between the E-step and the M-step until the Euclidean distance between $\boldsymbol{\theta}_{(j-1)}$ and $\boldsymbol{\theta}_{(j)}$ is smaller than 10^{-5}.

Estimated covariance matrix

Define $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^{\mathrm{T}}$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^{\mathrm{T}}$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{E} = (\widehat{E}(S_1), \ldots, \widehat{E}(S_n))^{\mathrm{T}}$, $\boldsymbol{e}_S = \boldsymbol{E} - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}$, $\boldsymbol{e}_Y = \boldsymbol{Y} - \widehat{\gamma}\boldsymbol{E} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$, and $A = \sum_{i=1}^{n}\widehat{E}(S_i^2)$. It is straightforward to show that the complete-data information matrix $\mathbf{I}_{\mathrm{c}}$ is a symmetric matrix with the following lower triangular elements:

$$\frac{1}{1-\widehat{\rho}^2}\begin{bmatrix} \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{\widehat{\sigma}_S^2} & & & & & \\ \frac{-\widehat{\rho}\mathbf{Z}^{\mathrm{T}}\mathbf{X}}{\widehat{\sigma}_S\widehat{\sigma}_Y} & \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}{\widehat{\sigma}_Y^2} & & & & \\ \frac{-\widehat{\rho}\boldsymbol{E}^{\mathrm{T}}\mathbf{X}}{\widehat{\sigma}_S\widehat{\sigma}_Y} & \frac{\boldsymbol{E}^{\mathrm{T}}\mathbf{Z}}{\widehat{\sigma}_Y^2} & \frac{A}{\widehat{\sigma}_Y^2} & & & \\ \frac{\boldsymbol{e}_S^{\mathrm{T}}\mathbf{X}}{\widehat{\sigma}_S^4}-\frac{\widehat{\rho}\boldsymbol{e}_Y^{\mathrm{T}}\mathbf{X}}{2\widehat{\sigma}_S^3\widehat{\sigma}_Y} & \frac{-\widehat{\rho}\boldsymbol{e}_S^{\mathrm{T}}\mathbf{Z}}{2\widehat{\sigma}_S^3\widehat{\sigma}_Y} & \frac{-\widehat{\rho}\left(A-\boldsymbol{E}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\alpha}}\right)}{2\widehat{\sigma}_S^3\widehat{\sigma}_Y} & \frac{n(2-\widehat{\rho}^2)}{4\widehat{\sigma}_S^4} & & \\ \frac{-\widehat{\rho}\boldsymbol{e}_Y^{\mathrm{T}}\mathbf{X}}{2\widehat{\sigma}_S\widehat{\sigma}_Y^3} & \frac{\boldsymbol{e}_Y^{\mathrm{T}}\mathbf{Z}}{\widehat{\sigma}_Y^4}-\frac{\widehat{\rho}\boldsymbol{e}_S^{\mathrm{T}}\mathbf{Z}}{2\widehat{\sigma}_S\widehat{\sigma}_Y^3} & \frac{\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{Y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})-\widehat{\gamma}A}{\widehat{\sigma}_Y^4}-\frac{\widehat{\rho}\left(A-\boldsymbol{E}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\alpha}}\right)}{2\widehat{\sigma}_S\widehat{\sigma}_Y^3} & \frac{-n\widehat{\rho}^2}{4\widehat{\sigma}_S^2\widehat{\sigma}_Y^2} & \frac{n(2-\widehat{\rho}^2)}{4\widehat{\sigma}_Y^4} & \\ \frac{-2\widehat{\rho}\boldsymbol{e}_S^{\mathrm{T}}\mathbf{X}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S^2}+\frac{(1+\widehat{\rho}^2)\boldsymbol{e}_Y^{\mathrm{T}}\mathbf{X}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y} & \frac{-2\widehat{\rho}\boldsymbol{e}_Y^{\mathrm{T}}\mathbf{Z}}{(1-\widehat{\rho}^2)\widehat{\sigma}_Y^2}+\frac{(1+\widehat{\rho}^2)\boldsymbol{e}_S^{\mathrm{T}}\mathbf{Z}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y} & \frac{-2\widehat{\rho}\left[\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{Y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})-\widehat{\gamma}A\right]}{(1-\widehat{\rho}^2)\widehat{\sigma}_Y^2}+\frac{(1+\widehat{\rho}^2)\left(A-\boldsymbol{E}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\alpha}}\right)}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y} & \frac{-n\widehat{\rho}}{2\widehat{\sigma}_S^2} & \frac{-n\widehat{\rho}}{2\widehat{\sigma}_Y^2} & \frac{n(1+\widehat{\rho}^2)}{(1-\widehat{\rho}^2)} \end{bmatrix}.$$

In addition, we compute $\boldsymbol{U}_i$, the gradient of $\log f(Y_i, S_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\widehat{\boldsymbol{\theta}}$, and the conditional expectations $\widehat{E}(\boldsymbol{U}_i)$ and $\widehat{E}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})$. It can be verified that $\boldsymbol{U}_i = \boldsymbol{V}_{i1} + S_i\boldsymbol{V}_{i2} + S_i^2\boldsymbol{V}_{i3}$, where

$$\boldsymbol{V}_{i1} = \begin{bmatrix} -\frac{1}{(1-\widehat{\rho}^2)\widehat{\sigma}_S^2}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i)\boldsymbol{X}_i - \frac{\widehat{\rho}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y}(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i)\boldsymbol{X}_i \\ \frac{1}{(1-\widehat{\rho}^2)\widehat{\sigma}_Y^2}(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i)\boldsymbol{Z}_i + \frac{\widehat{\rho}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i)\boldsymbol{Z}_i \\ 0 \\ -\frac{1}{2\widehat{\sigma}_S^2} + \frac{1}{2(1-\widehat{\rho}^2)\widehat{\sigma}_S^4}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i\right)^2 + \frac{\widehat{\rho}}{2(1-\widehat{\rho}^2)\widehat{\sigma}_S^3\widehat{\sigma}_Y}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right)\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) \\ -\frac{1}{2\widehat{\sigma}_Y^2} + \frac{1}{2(1-\widehat{\rho}^2)\widehat{\sigma}_Y^4}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right)^2 + \frac{\widehat{\rho}}{2(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y^3}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right)\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) \\ \frac{\widehat{\rho}}{1-\widehat{\rho}^2} - \frac{\widehat{\rho}}{(1-\widehat{\rho}^2)^2\widehat{\sigma}_S^2}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right)^2 - \frac{\widehat{\rho}}{(1-\widehat{\rho}^2)^2\widehat{\sigma}_Y^2}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right)^2 - \frac{1+\widehat{\rho}^2}{(1-\widehat{\rho}^2)^2\widehat{\sigma}_S\widehat{\sigma}_Y}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right)\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) \end{bmatrix},$$

$$\boldsymbol{V}_{i2} = \frac{1}{1-\widehat{\rho}^2}\begin{bmatrix} \frac{1}{\widehat{\sigma}_S^2}\boldsymbol{X}_i + \frac{\widehat{\rho}\widehat{\gamma}}{\widehat{\sigma}_S\widehat{\sigma}_Y}\boldsymbol{X}_i \\ -\frac{\widehat{\gamma}}{\widehat{\sigma}_Y^2}\boldsymbol{Z}_i - \frac{\widehat{\rho}}{\widehat{\sigma}_S\widehat{\sigma}_Y}\boldsymbol{Z}_i \\ \frac{1}{\widehat{\sigma}_Y^2}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) + \frac{\widehat{\rho}}{\widehat{\sigma}_S\widehat{\sigma}_Y}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i \\ -\frac{1}{\widehat{\sigma}_S^4}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i - \frac{\widehat{\rho}}{2\widehat{\sigma}_S^3\widehat{\sigma}_Y}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i + \widehat{\gamma}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right) \\ -\frac{\widehat{\gamma}}{\widehat{\sigma}_Y^4}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) - \frac{\widehat{\rho}}{2\widehat{\sigma}_S\widehat{\sigma}_Y^3}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i + \widehat{\gamma}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right) \\ \frac{2\widehat{\rho}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S^2}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i + \frac{2\widehat{\rho}\widehat{\gamma}}{(1-\widehat{\rho}^2)\widehat{\sigma}_Y^2}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i\right) + \frac{1+\widehat{\rho}^2}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y}\left(Y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i + \widehat{\gamma}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_i\right) \end{bmatrix},$$

and

$$\boldsymbol{V}_{i3} = \frac{1}{1-\widehat{\rho}^2}\begin{bmatrix} 0 \\ 0 \\ -\dfrac{\widehat{\gamma}}{\widehat{\sigma}_Y^2} - \dfrac{\widehat{\rho}}{\widehat{\sigma}_S\widehat{\sigma}_Y} \\ \dfrac{1}{2\widehat{\sigma}_S^4} + \dfrac{\widehat{\rho}\widehat{\gamma}}{2\widehat{\sigma}_S^3\widehat{\sigma}_Y} \\ \dfrac{\widehat{\gamma}^2}{2\widehat{\sigma}_Y^4} + \dfrac{\widehat{\rho}\widehat{\gamma}}{2\widehat{\sigma}_S\widehat{\sigma}_Y^3} \\ -\dfrac{\widehat{\rho}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S^2} - \dfrac{\widehat{\rho}\widehat{\gamma}^2}{(1-\widehat{\rho}^2)\widehat{\sigma}_Y^2} - \dfrac{(1+\widehat{\rho}^2)\widehat{\gamma}}{(1-\widehat{\rho}^2)\widehat{\sigma}_S\widehat{\sigma}_Y} \end{bmatrix}.$$

Thus, $\widehat{E}(\boldsymbol{U}_i) = \boldsymbol{V}_{i1} + \widehat{E}(S_i)\boldsymbol{V}_{i2} + \widehat{E}(S_i^2)\boldsymbol{V}_{i3}$, and

$$\begin{aligned}\widehat{E}\left(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}\right) = {} & \boldsymbol{V}_{i1}\boldsymbol{V}_{i1}^{\mathrm{T}} + \left(\boldsymbol{V}_{i1}\boldsymbol{V}_{i2}^{\mathrm{T}} + \boldsymbol{V}_{i2}\boldsymbol{V}_{i1}^{\mathrm{T}}\right)\widehat{E}(S_i) \\ & + \left(\boldsymbol{V}_{i1}\boldsymbol{V}_{i3}^{\mathrm{T}} + \boldsymbol{V}_{i2}\boldsymbol{V}_{i2}^{\mathrm{T}} + \boldsymbol{V}_{i3}\boldsymbol{V}_{i1}^{\mathrm{T}}\right)\widehat{E}\left(S_i^2\right) \\ & + \left(\boldsymbol{V}_{i2}\boldsymbol{V}_{i3}^{\mathrm{T}} + \boldsymbol{V}_{i3}\boldsymbol{V}_{i2}^{\mathrm{T}}\right)\widehat{E}\left(S_i^3\right) + \boldsymbol{V}_{i3}\boldsymbol{V}_{i3}^{\mathrm{T}}\widehat{E}\left(S_i^4\right).\end{aligned}$$

To evaluate the conditional expectation $\widehat{E}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}})$, we need to derive $\widehat{E}(S_i^3)$ and $\widehat{E}(S_i^4)$. If S_i is observed, then $\widehat{E}(S_i^3) = S_i^3$ and $\widehat{E}(S_i^4) = S_i^4$. If S_i is unmeasured, then $\widehat{E}(S_i^3) = b_i^3 + 3b_ia$, and $\widehat{E}(S_i^4) = b_i^4 + 6b_i^2a + 3a^2$. If S_i is below the lower detection limit, then

$$\widehat{E}\left(S_i^3\right) = b_i^3 + 3b_ia - \frac{\varphi(L_{0i})}{\Phi(L_{0i})}\left(3b_i^2\sqrt{a} + 2a^{\frac{3}{2}} + 3b_iaL_{0i} + a^{\frac{3}{2}}L_{0i}^2\right),$$

and

$$\begin{aligned}\widehat{E}\left(S_i^4\right) = {} & b_i^4 + 6b_i^2a + 3a^2 - \frac{\varphi(L_{0i})}{\Phi(L_{0i})}\left[4b_i^3\sqrt{a} + 8b_ia^{\frac{3}{2}}\right. \\ & \left. + \left(6b_i^2a + 3a^2\right)L_{0i} + 4b_ia^{\frac{3}{2}}L_{0i}^2 + a^2L_{0i}^3\right].\end{aligned}$$

If S_i is above the upper detection limit, then

$$\widehat{E}\left(S_i^3\right) = b_i^3 + 3b_ia + \frac{\varphi(U_{0i})}{\Phi(U_{0i})}\left(3b_i^2\sqrt{a} + 2a^{\frac{3}{2}} - 3b_iaU_{0i} + a^{\frac{3}{2}}U_{0i}^2\right),$$

and

$$\begin{aligned}\widehat{E}\left(S_i^4\right) = {} & b_i^4 + 6b_i^2a + 3a^2 + \frac{\varphi(U_{0i})}{\Phi(U_{0i})}\left[4b_i^3\sqrt{a} + 8b_ia^{\frac{3}{2}}\right. \\ & \left. - \left(6b_i^2a + 3a^2\right)U_{0i} + 4b_ia^{\frac{3}{2}}U_{0i}^2 - a^2U_{0i}^3\right].\end{aligned}$$

Then we can obtain the observed-data information matrix in Equation 4.

References

1. Burgess, S., and Thompson, S.G. (2015). Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation (CRC Press).
2. Wehby, G.L., Fletcher, J.M., Lehrer, S.F., Moreno, L.M., Murray, J.C., Wilcox, A., and Lie, R.T. (2011). A genetic instrumental variables analysis of the effects of prenatal smoking on birth weight: Evidence from two samples. Biodemogr. Soc. Biol. *57*, 3–32.
3. Davies, N.M., von Hinke Kessler Scholder, S., Farbmacher, H., Burgess, S., Windmeijer, F., and Smith, G.D. (2015). The many weak instruments problem and Mendelian randomization. Stat. Med. *34*, 454–468.
4. Little, R.J., and Rubin, D.B. (2020). Statistical Analysis with Missing Data, 3rd Edition (John Wiley & Sons).
5. Pierce, B.L., and Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am. J. Epidemiol. *178*, 1177–1184.
6. 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.
7. 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.
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. Lin, D.Y., Zeng, D., and Couper, D. (2020). A general framework for integrative analysis of incomplete multiomics data. Genet. Epidemiol. *44*, 646–664.
10. Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. Environ. Health Perspect. *112*, 1691–1696.
11. Labaki, W.W., Gu, T., Murray, S., Curtis, J.L., Yeomans, L., Bowler, R.P., Barr, R.G., Comellas, A.P., Hansel, N.N., Cooper, C.B., et al. (2019). Serum amino acid concentrations and clinical outcomes in smokers: SPIROMICS metabolomics study. Sci. Rep. *9*, 11367.
12. Surendran, P., Stewart, I.D., Au Yeung, V.P.W., Pietzner, M., Raffler, J., Wörheide, M.A., Li, C., Smith, R.F., Wittemans, L.B.L., Bomba, L., et al. (2022). Rare and common genetic determinants of metabolic individuality and their effects on human health. Nat. Med. *28*, 2321–2332.
13. Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., and Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Br. Med. J. *338*, b2393.
14. 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.
15. 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.
16. 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).
17. 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.

18. Cánovas, R., Cuartero, M., and Crespo, G.A. (2019). Modern creatinine (bio) sensing: Challenges of point-of-care platforms. Biosens. Bioelectron. *130*, 110–124.
19. Bulbul, M.C., Dagel, T., Afsar, B., Ulusu, N.N., Kuwabara, M., Covic, A., and Kanbay, M. (2018). Disorders of lipid metabolism in chronic kidney disease. Blood Purif. *46*, 144–152.
20. Inker, L.A., Eneanya, N.D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., Crews, D.C., Doria, A., Estrella, M.M., Froissart, M., et al. (2021). New creatinine- and cystatin C–based equations to estimate GFR without race. N. Engl. J. Med. *385*, 1737–1749.
21. 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.
22. Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M., and Milgram, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal. Chem. *81*, 6656–6667.
23. Ohta, T., Masutomi, N., Tsutsui, N., Sakairi, T., Mitchell, M., Milburn, M.V., Ryals, J.A., Beebe, K.D., and Guo, L. (2009). Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. Toxicol. Pathol. *37*, 521–535.
24. 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.
25. 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.
26. 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.
27. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.
28. 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.
29. 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.
30. 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.
31. Lin, D.Y., Tao, R., Kalsbeek, W.D., Zeng, D., Gonzalez, F., II, 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.
32. Kang, H., Zhang, A., Cai, T.T., and Small, D.S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. J. Am. Stat. Assoc. *111*, 132–144.
33. Jiang, L., Oualkacha, K., Didelez, V., Ciampi, A., Rosa-Neto, P., Benedet, A.L., Mathotaarachchi, S., Richards, J.B., Greenwood, C.M.T.; and Alzheimer's Disease Neuroimaging Initiative (2019). Constrained instruments and their application to Mendelian randomization with pleiotropy. Genet. Epidemiol. *43*, 373–401.
34. Tchetgen, E.T., Sun, B., and Walter, S. (2021). The GENIUS approach to robust Mendelian randomization inference. Stat. Sci. *36*, 443–464.
35. Van Kippersluis, H., and Rietveld, C.A. (2018). Pleiotropy-robust Mendelian randomization. Int. J. Epidemiol. *47*, 1279–1288.
36. Spiller, W., Slichter, D., Bowden, J., and Davey Smith, G. (2019). Detecting and correcting for bias in Mendelian randomization analyses using gene-by-environment interactions. Int. J. Epidemiol. *48*, 702–712.
37. Guo, Z., Kang, H., Tony Cai, T., and Small, D.S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. J. Roy. Stat. Soc. B *80*, 793–815.
38. Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. J. Am. Stat. Assoc. *114*, 1339–1350.
39. Burgess, S., Daniel, R.M., Butterworth, A.S., Thompson, S.G.; and EPIC-InterAct Consortium (2015). Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. Int. J. Epidemiol. *44*, 484–495.