Robust inference in two-sample Mendelian randomisation via the zero modal pleiotropy assumption

Fernando Pires Hartwig[1,2]*, George Davey Smith[2,3], Jack Bowden[2,3]

[1]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil.

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.

[3]School of Social and Community Medicine, University of Bristol, United Kingdom.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil) 96020-220. Phone: 55 53 981068670. E-mail: fernandophartwig@gmail.com.

Abstract

Background: Mendelian randomisation (MR) is being increasingly used as a strategy to improve causal inference in observational studies. Availability of summary data of genetic associations for a variety of phenotypes from large genome-wide association studies (GWAS) allows straightforward application of MR in a two-sample design. In addition to the conventional inverse variance weighting (IVW) method, recently developed summary data MR methods, such as the MR-Egger and weighted median approaches, allow a relaxation of the instrumental variable assumptions.

Methods: Here, a new method –the mode estimator – is proposed to obtain a single causal effect estimate from multiple genetic instruments. The mode estimate is consistent when the largest number of similar (identical in infinite samples) individual-instrument causal effect estimates comes from valid instruments, even if the majority of instruments are invalid. We evaluated the performance of the method in simulations designed to mimic the two-sample summary data setting, and demonstrate its use by investigating the causal effect of plasma lipid fractions and urate levels on coronary heart disease risk.

Results: The mode estimates presented less bias and type-I error rates than other methods under the null in many situations. Its power to detect a causal effect was smaller compared to the IVW and weighted median methods, but was larger than that of MR-Egger regression, with sample size requirements typically smaller than those available from GWAS consortia.

Conclusions: The mode estimator relaxes the instrumental variable assumptions, and should be used in combination with other approaches in a sensitivity analysis.

Key words: Causality; Instrumental variables; Genetic variation; Mendelian Randomization; Genetic pleiotropy.

Key Messages

- Two-sample Mendelian randomisation is being increasingly used due to the availability of summary association results from large genome-wide association studies.
- Mendelian randomisation analyses using multiple genetic instruments selected solely based on statistical criteria are prone to bias due to horizontal pleiotropy.
- A mode estimator can be used to obtain causal effects robust against horizontal pleiotropy.
- The mode estimator requires that the most common causal effect estimate is a consistent estimate of the true causal effect, even if the majority of instruments are invalid (i.e., the ZEro Modal Pleiotropy Assumption, or ZEMPA).
- Plotting the smoothed empirical density function is useful to explore the distribution of causal effect estimates.

Introduction

Using germline genetic variants as instrumental variables of modifiable exposure phenotypes can strengthen causal inference in observational studies by applying the principles of Mendelian randomisation (MR).[1,2] This method has already been used to address causality in several exposure-outcome combinations and has become a common feature in the recent epidemiological literature.[3] Causal inference using MR relies on instrumental variable assumptions, which require that the genetic variant is i) associated with the exposure, ii) independent on confounders of the exposure-outcome associations, and iii) independent of the outcome after conditioning on the exposure and exposure-outcome confounders. Germline genetic variants have been proposed as instrumental variables because they are defined at conception (thus preventing reverse causation) and are generally not associated with the confounders that influence conventional observational associations.[1,2,4] MR limitations include bias due to population stratification, linkage disequilibrium, canalization or horizontal pleiotropy (i.e., effects of the genetic instrument on the outcome not fully mediated by the exposure).[1,2]

New applications addressing different limitations of MR feature in the recent literature. Such applications include methods to perform MR with multiple genetic variants, typically single nucleotide polymorphisms (SNPs), using summary data from genome-wide association studies (GWAS).[5] Given the increasing number of publicly available summary statistics from large GWAS consortia, summary data MR methods allow use of these data to interrogate many causal hypotheses without the need to collect individual-level data.[6,7]

Using many instruments increases the probability of including at least one invalid instrument and, therefore, biasing the estimate. For example, the inverse variance weighting [IVW] method requires that either all variants are valid instruments or that there is balanced horizontal pleiotropy (i.e., horizontal pleiotropic effects of individual instruments balance out) and that such pleiotropic effects

are independent of instrument strength across all variants (i.e., the Instrument Strength Independent on Direct Effects – InSIDE – assumption).[5,8] More recently, other summary data MR methods that allow relaxing (but not eliminating) the instrumental variable assumptions regarding horizontal pleiotropy have been proposed.[9,10]

In this paper, we describe a new summary data MR method – the mode estimator. We clarify when this will be a consistent estimate of the causal effect, compare it to the IVW, MR-Egger regression and weighted median methods using simulations, and illustrate its application using a real data example.

Methods

In order to motivate the summary data methods discussed in this paper, we assume the following simple data generating model linking genetic variant G_j ($j = 1, \dots, L$), with all L genetic variants being independent on one another, a continuous exposure X_i and outcome Y_i for subject i.

$$X_i | G_{ij} = \beta_{X0} + \beta_{Xj} G_{ij} + \lambda_{Xij} \quad (1)$$

$$Y_i | G_{ij} = \beta_{Y0} + (\beta\beta_{Xj} + \alpha_j) G_{ij} + \lambda_{Yij} \quad (2)$$

$$= \beta_{Y0} + \beta_{Yj} G_{ij} + \lambda_{Yij}.$$

A more detailed statistical model is given in the Simulations section. Here, β_{Xj} and $\beta_{Yj} = (\beta\beta_{Xj} + \alpha_j)$ represent G_j's true association with the exposure and outcome, respectively. $\beta\beta_{Xj}$ is the effect of G_j on Y through X, where β is the causal effect of X on Y we wish to estimate. The term α_j represents the association between G_j and Y not through the exposure of interest, due to horizontal pleiotropy. The error terms λ_{Xij} and λ_{Yij} will generally be correlated when the collected on the same individuals. However, we will mainly focus on the two-sample setting where the error terms are independent, because independent samples are used to fit models (1) and (2).

Let $\hat{\beta}_{Xj}$ and $\hat{\beta}_{Yj}$ represent the SNP-exposure and SNP-outcome association estimate for variant j, respectively, and let σ^2_{Xj} and σ^2_{Yj} represent the variance of $\hat{\beta}_{Xj}$ and $\hat{\beta}_{Yj}$, respectively. The ratio estimate[11,12] for the causal effect β using variant j alone is equal to

$$\hat{\beta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \quad (3),$$

the standard error of which (σ_{Rj}) can be obtained using the delta method[13] as follows:

$$\sigma_{Rj} = \sqrt{\frac{\sigma^2_{Y_j}}{\hat{\beta}^2_{X_j}} + \frac{\hat{\beta}^2_{Y_j}\sigma^2_{X_j}}{\hat{\beta}^4_{X_j}}} \quad (4).$$

The standard error in (4) can be simplified to $\sigma_{Y_j}/|\beta_{X_j}|$ when the variance of the SNP-exposure association σ^2_{Xj} is small enough to be considered "ignorable", or equivalently that $\hat{\beta}_{Xj} = \beta_{Xj}$. This is called the NO Measurement Error (NOME) assumption.[14]

The ratio estimate $\hat{\beta}_{R_j}$ is a crude measure of causal effect, but has a major advantage over more sophisticated methods in that it can be calculated using summary data estimates for β_{X_j} and β_{Y_j} alone. These estimates can then be used furnish a summary data MR analysis using the framework of a meta-analysis. Furthermore, when multiple genetic variants are available, heterogeneity between the $\hat{\beta}_{R_j}$s can be used to probe for the presence of horizontal pleiotropy.[8,15]

We now briefly review three such summary data methods, and discuss the conditions under which each method returns a consistent causal effect estimate (i.e., estimate converges in probability to the true value as the sample size increases).

Summary data Mendelian randomisation

The IVW estimate for the causal effect is an inverse-variance weighted average of all $\hat{\beta}_{R_j}$s. Under models (1) and (2), variant j is a valid instrument when α_j=0 and invalid when α_j≠0. When variant j is

a valid instrument, $\hat{\beta}_{R_j}$ is a consistent estimate for the true causal effect β (i.e., $\beta_{R_j} = \beta$) and variation between the $\hat{\beta}_{R_j}$s should be due to sampling variation alone. However, when $\alpha_j \neq 0$, $\hat{\beta}_{R_j} = \beta + b_j$, where $b_j = \alpha_j / \beta_{X_j}$ (i.e., a bias term). Therefore, the IVW estimate is only a consistent estimate for β if these bias terms (a) sum to zero and (b) are independent in magnitude of the β_{X_j} parameters (i.e., balanced pleiotropy under the InSIDE assumption[9]). Assessing the heterogeneity between the $\hat{\beta}_{R_j}$s and the adoption of random effects models have therefore been proposed as strategies to detect horizontal pleiotropy.[8,15]

Two new methods for summary data MR that offer increased robustness to directional horizontal pleiotropy (i.e., bias terms do not sum to zero) have been proposed: MR-Egger regression[9] and the weighted median.[10] Briefly, MR-Egger regression fits an inverse-variance weighted linear regression of $\hat{\beta}_{Y_j}$ against $\hat{\beta}_{X_j}$. Its intercept can be interpreted as a test of directional horizontal pleiotropy. It can provide a consistent estimate even when there is directional horizontal pleiotropy, as long as the InSIDE assumption holds. The weighted median estimate is the 50th percentile of the inverse-variance weighted empirical distribution function of all $\hat{\beta}_{R_j}$s. It is a consistent estimate if at least 50% of the weight in the analysis stems from valid instruments. We will implement all three methods using inverse-variance weights calculated under NOME.

The mode estimator

In this paper we propose a new causal effect estimator – the mode estimator – that offers robustness to horizontal pleiotropy in a different manner to that of the IVW, MR-Egger or weighted median methods. Its ability to consistently estimate the true causal effect relies on the following fundamental assumption termed the ZEro Modal Pleiotropy Assumption (ZEMPA):

ZEMPA: *Across all instruments, the most frequent value (i.e., the mode) of b_j is 0.*

In order to formalize this, let $k \in \{1,2,\ldots,L\}$ represent the number of unique values of α_j among the L variants. If all α_j terms are the same then k=1, but if all are unique then $k = L$. Now, let $n_1, n_2, \ldots, n_k$ represent the number of instruments that have the same non-zero value of b_j, where n_1 represents those with the smallest identical value of b_j and n_k represents those with the largest absolute value. Finally, let n_0 represent the number of valid instruments whose b_j terms are identically zero. We then have that $n_0 + n_1 + \cdots + n_k = L$. ZEMPA implies that n_0 is larger than any other n_l for l in $1, 2, \ldots, k$ (i.e., $n_0 > \mathrm{Max}(n_1, \ldots, n_k)$). For a weighted mode estimator, ZEMPA implies that the weights associated with the valid instruments are the largest among all k subsets of instruments (ie., $w_0 > \mathrm{Max}(w_1, \ldots, w_k)$, where w_l is the weight contributed by the lth subset of instruments). If $\mathrm{Max}(n_1, \ldots, n_k)/L$ <50% or $\mathrm{Max}(w_1, \ldots, w_k)/w$ <50%, the breakdown level (i.e., the maximum proportion of information that can come from invalid instruments before the method is inconsistent) of the simple mode estimate is $\{L - [\mathrm{Max}(n_1, \ldots, n_k) + 1]\}/L$, while the breakdown level of the weighted mode estimate is $\{w - [\mathrm{Max}(w_1, \ldots, w_k) + 1]\}/w$, where w denotes the weights from all instruments (i.e., 100% of the weights). Therefore, ZEMPA implies the most frequent individual-instrument causal effect estimate (i.e., the mode) is a consistent estimate for the causal effect.

For example, for the simple mode, ZEMPA would be satisfied if just two variants were valid and the remainder $(L - 2)$ were invalid if the invalid variants have unique values for b_j (i.e., n_0=2 and $n_1 = n_2 = \ldots = n_k = 1$). This is illustrated in Figure 1A, where 6 out of 8 instruments are invalid (so $n_0 = 2$), but all non-zero b_js are unique, implying that $k = L - 1 = 7$ and $n_1 = n_2 = \ldots = n_7 = 1$. In this situation, ZEMPA is satisfied and the simple mode estimate $\hat{\beta}_M$ is a consistent estimate of the causal effect β. However, when the largest number of identical estimates comes from invalid instruments (i.e., $n_0 < n_l$ for some l, ZEMPA violated), then the simple mode estimate will be inconsistent for β (i.e., asymptotically biased). Similarly, the weighted mode estimate will be biased if $w_0 < w_l$ for some l. This is illustrated in Figure 1B, which shows causal effect estimates from 6

invalid and 2 valid variants ($n_0 = 2$). Since three variants have precisely the same horizontal pleiotropic effect in this example ($n_2 = 3$), ZEMPA is violated.

Implementing the mode method

To implement the mode method, we propose using the mode of the smoothed empirical density function of all $\hat{\beta}_{R_j}$s as the causal effect estimate. This strategy is straightforward to implement, easily deals with sampling variation in asymptotically identical $\hat{\beta}_{R_j}$s and allows giving different weights to different instruments. We refer to the mode of the unweighted and inverse-variance weighted empirical density function as the simple and weighted mode methods, respectively. The standardised weights for the weighted mode can be computed as follows:

$$w_j = \sigma_{R_j}^{-2} \Big/ \sum_{j=1}^{L} \sigma_{R_j}^{-2} \quad (5).$$

For the simple mode, $w_1 = w_2 = \cdots = w_L = 1/L$.

Consider the normal kernel density function of the $\hat{\beta}_{R_j}$s

$$f(x) = \frac{1}{h\sqrt{2\pi}} \sum_{j=1}^{L} w_j \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\beta}_{R_j}}{h}\right)^2\right] \quad (6),$$

where h is the smoothing bandwidth parameter.[16] The mode causal effect estimate $\hat{\beta}_M$ is the value of x that maximizes $f(x)$ (i.e., $f\left(\hat{\beta}_M\right) = \max\,[f(x)]$). The h parameter regulates a bias-variance trade-off of the mode estimate, with increasing h leading to higher precision, but also to higher bias. Here, $h = \varphi s$, with φ being a tuning parameter that allows increasing or decreasing the bandwidth, and s being the default bandwidth value according to some criterion. We used the modified Silverman's bandwidth rule proposed by Bickel[17]:

$$s = \frac{0.9\min\left(\mathrm{sd}(\hat{\beta}_{R_J}), 1.4826\mathrm{mad}(\hat{\beta}_{R_J})\right)}{L^{\frac{1}{5}}} \quad (7),$$

where $\mathrm{sd}\left(\hat{\beta}_{R_J}\right)$ and $\mathrm{mad}(\hat{\beta}_{R_J})$ are the standard deviation and median absolute deviation from the median of the L $\hat{\beta}_{R_j}$s, respectively.

An analogy can be drawn between the proposed method and histograms. Consider a hypothetical bimodal variable V, which was simulated by concatenating two normally distributed variables: $V_1 \sim N(0,1)$ and $V_2 \sim N(5,25)$. In this example, the mode of V would be zero, because the distribution with the smallest dispersion is centred at zero. We can use histograms to obtain a rough mode estimate of V using the midpoint of the highest bar.

To make the histogram, it is necessary to define the break points of each bar, which here will be equally-spaced, starting from the minimum value of V and ending at the maximum value of V. As shown in Figure 2A, using six equally-spaced bars does not provide enough resolution to detect bimodality in V. As a result, the mode estimate is biased upwards. In panel B (20 equally-spaced bars), it can be clearly seen that V is bimodal, and the mode estimate is less biased. However, when 100 equally-spaced bars are used (panel C), the data appears to be multimodal, because such spacing criteria was too stringent, so that random fluctuations under the same underlying distribution end up looking like being different distributions in the histogram. While this would not be a problem in infinite samples, it results in numerical instability of the mode estimate in finite samples. So, among panels A-C, the histogram in panel B provides the best bias-variance trade-off, and would be the best one to display the distribution of V.

As panels D-F illustrate, the proposed smoothed empirical density function approach is analogous to the histogram: the smoothed empirical density function is roughly similar to connecting the midpoints of each bar of the histogram (in other words, a "smoothed histogram"); the mode estimate based on the smoothed empirical density function is roughly similar to the mode estimate using the midpoint of the highest bar; and decreasing the tuning parameter φ is analogous to increasing the number of equally-spaced bars in the histogram. Panels D-F also illustrate that the

mode (as estimated by our proposed method) and the mean can be substantially different in case of bimodality, and our mode estimate being close to the centre of the main distribution.

Simulation model

Individual participant data model

The simulations were performed using the following model to generate individual i's exposure X_i, outcome Y_i and confounder U_i based on their underlying genetic data vector $(G_{i1}, \dots, G_{iL})$:

$$U_i = \gamma_U Z_{U_i} + \varepsilon_{U_i} \quad (8)$$

$$X_i = \gamma_X Z_{X_i} + \theta_X U_i + \varepsilon_{X_i} \quad (9)$$

$$Y_i = \gamma_Y Z_{Y_i} + \beta X_i + \theta_Y U_i + \varepsilon_{Y_i} \quad (10),$$

where:

$$Z_{U_i} = \left(\sum_{j=1}^{L} \delta_{U_j} G_{ij}\right) / \sigma_{ZU}\,, Z_{X_i} = \left(\sum_{j=1}^{L} \delta_{X_j} G_{ij}\right) / \sigma_{ZX}\,,\ Z_{Y_i} = \left(\sum_{j=1}^{L} \delta_{Y_j} G_{ij}\right) / \sigma_{ZY}.$$

Z_U, Z_X and Z_Y represent the additive allele scores of L independent SNPs on U, X and Y, modulated by the parameters $\delta_{U_j}, \delta_{X_j}, \delta_{Y_j}$ (j=1,… L). β denotes the true causal effect of X on Y that we wish to estimate. The underlying genetic variables (G_{ij}) were generated independently by sampling from a Binomial (2, p) distribution with p itself drawn from a Uniform(0.1,0.9) distribution, to mimic bi-allelic SNPs in Hardy-Weinberg equilibrium. The resulting allele scores were then divided by their sample standard deviations $(\sigma_{ZU}, \sigma_{ZX}, \sigma_{ZY})$, to set variances to one. The direct effects of U on X and Y are denoted by θ_X and θ_Y, respectively. θ_X and θ_Y are set to positive values in all simulations, so as to always induce positive confounding. Error terms $\varepsilon_{U_i}, \varepsilon_{X_i}, \varepsilon_{Y_i}$ were independently generated from a normal distribution, with mean=0 and variances $\sigma^2_{\varepsilon U}$, $\sigma^2_{\varepsilon X}$ and $\sigma^2_{\varepsilon Y}$ respectively, whose values were chosen to set the variances of U, X and Y to one.

Constraining the variances in this way enables easy interpretation of the parameters in models (8)-(10). For example, β=0.1 implies that the causal effect of X on Y explains 1% of Y variance, and $\gamma_X = \sqrt{0.1}$ implies that the genetic data explains 10% of X variance directly (even though the total amount of X variance explained by Z would be larger in case $\gamma_U \neq 0$). Moreover, β=0.1 implies that one standard deviation increase in X causes a 0.1 standard deviation increase in Y.

Summary data interpretation

Although our underlying model is generated at the individual level, all of the methods we will investigate are applied directly to the set of L SNP-exposure and SNP-outcome association estimates (written as $\hat{\beta}_{Xj}$ and $\hat{\beta}_{Yj}$ in models (1) and (2)). These summary data parameters can be linked to the individual participant data parameters as follows. Firstly, let $Z_{U_{ij}}$, $Z_{X_{ij}}$ and $Z_{Y_{ij}}$ represent the jth component of Z_{U_i}, Z_{X_i} and Z_{Y_i}, respectively. We then have that

$$\beta_{X_j} G_{ij} = \gamma_X Z_{X_{ij}} + \gamma_U \theta_X Z_{U_{ij}} \quad (11)$$

$$\beta_{Y_j} G_{ij} = \gamma_Y Z_{Y_{ij}} + \gamma_U \theta_Y Z_{U_{ij}} + \beta \left(\gamma_X Z_{X_{ij}} + \gamma_U \theta_X Z_{U_{ij}} \right), \quad (12)$$

which yields the ratio estimand

$$\beta_{R_j} = \beta + \frac{\gamma_Y \frac{\delta_{Y_j}}{\sigma_{ZY}} + \gamma_U \theta_Y \frac{\delta_{U_j}}{\sigma_{ZU}}}{\gamma_X \frac{\delta_{X_j}}{\sigma_{ZX}} + \gamma_U \theta_X \frac{\delta_{U_j}}{\sigma_{ZU}}} = \beta + \frac{Total\ direct\ effect\ of\ G_j}{Instrument\ strength\ of\ G_j} = \beta + \frac{\alpha_j}{\beta_{X_j}}. \quad (13)$$

Variant j therefore gives a consistent estimate for β only when the total direct effect of G_j on Y=0. This is equivalent to a_j in model (2). The IVW estimate is only consistent if the total direct effect is zero for all variants and InSIDE holds. The weighted median is consistent if at least 50% of the information in the analysis stems from variants with a zero direct effect. MR-Egger regression can consistently estimate the causal effect if, across all variants, the total direct effects are independent

of the instrument strengths (i.e., InSIDE holds). This is only true in general when horizontal pleiotropy does not affect Y via a confounder (γ_U=0), so that the ratio estimand for variant G_j is:

$$\beta_{R_j} = \beta + \frac{\gamma_Y \frac{\delta_{Y_j}}{\sigma_{ZY}}}{\gamma_X \frac{\delta_{X_j}}{\sigma_{ZX}}}. \quad (14)$$

Finally, the mode estimate is consistent under ZEMPA – i.e., when the most frequent total direct (or pleiotropic) effect is zero (in the case of the simple mode).

Simulation scenarios

Although the consistency property of an estimator provides a formal justification of the approach, it is equally important to understand how well it works in practice for realistically sized data sets in comparison to other methods. Therefore, we evaluated our proposed mode estimator in four different simulation scenarios. In all simulations the number of variants L=30, θ_X=θ_Y=$\sqrt{0.3}$, γ_X=$\sqrt{0.1}$ and γ_U=γ_Y=$\rho\sqrt{0.1}/L$, where $\rho = 0, 3, 6, \ldots, 30$ is the number of invalid instruments.

Simulations 1 and 2 were aimed at evaluating the methods performance under the causal null (β=0) in the two-sample setting. Datasets of 100,000 individuals were simulated and divided in half at random, and each was used to estimate either SNP-exposure or SNP-outcome associations. Simulations 3 and 4 were aimed at evaluating weak instrument bias in the two-sample and single-sample settings; sample sizes used to estimate instrument-exposure (N_X) and instrument-outcome (N_Y) associations were allowed to vary, as described below.

Simulation 1: In this scenario, δ_{U_j} was 0 for all instruments, implying that there is no InSIDE-violating horizontal pleiotropy. InSIDE-respecting horizontal pleiotropic effects δ_{Y_j} were drawn from a Uniform(0.01, 0.2) distribution for the ρ invalid instruments or were set to 0 for valid instruments. Moreover, given that β=0, power can be interpreted as the type-I error rate.

Simulation 2: InSIDE-violating horizontal pleiotropy was induced by setting δ_{Y_j}=0 for all instruments, while δ_{U_j} values were drawn from a Uniform(0.01, 0.2) distribution for the ρ invalid instruments.

Simulation 3: This simulation evaluated the performance of the estimators to detect a positive causal effect of β=0.1 in the two-sample context. δ_{U_j}=δ_{Y_j}=0 for all L instruments, and ρ=0, implying that there is no horizontal pleiotropy. $N_X \in$ {25,000, 50,000, 100,000}, and $N_Y \in$ {25,000, 50,000, 100,000}.

Simulation 4: This simulation evaluated the performance of the estimators under the causal null when SNP-exposure and SNP-outcome associations are estimated in partially (50%) or fully (100%) overlapping samples (the latter being equivalent to the single sample setting). It was implemented as simulation 3, but β=0 and $N_X = N_Y \in$ {1,000, 5,000, 10,000}. We used smaller sample sizes so to purposely increase the bias due to sample overlap, thus facilitating comparisons between methods.

Applied examples: plasma lipid fractions and urate levels and coronary heart disease risk

Do and colleagues[18] performed a two-sample MR analysis to evaluate the causal effect of low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides on coronary heart disease (CHD) risk, using a total of 185 genetics variants. Summary association results were obtained from the Global Lipids Genetics Consortium[19] and the Coronary Artery Disease Genome-Wide Replication and Meta-Analysis Consortium,[20] and were downloaded from Do and colleagues' supplementary material (standard errors were estimated based on the regression coefficients and P-values). Genetic variants were classified as instruments for each lipid fraction using a statistical criterion ($P<1\times10^{-8}$), resulting in 73 instruments for LDL-C, 85 for HDL-C and 31 for triglycerides.

White and colleagues[21] performed a similar analysis, but with plasma urate levels rather than lipid fractions. 31 variants associated with urate levels ($P<5\times10^{-7}$) and the required summary statistics were obtained from the GWAS catalog (https://www.ebi.ac.uk/gwas/).

Statistical analyses

In all simulation scenarios, causal effect estimates were obtained using established MR methods (multiplicative random effects IVW, multiplicative random effects MR-Egger regression and weighted median), as well as the simple mode and the weighted mode methods. Each version of the mode estimator was evaluated without and with making the NOME assumption, thus yielding four mode methods. Each of these four methods was evaluated for two values of the tuning parameter $\varphi \in \{1, 0.5\}$, totalising of eight versions of the mode method. Parametric bootstrap was used to estimate the standard errors of the mode estimates using the median absolute deviation from the median of the bootstrap distribution of causal effect estimates. These were used to derive symmetric confidence intervals, after multiplying by 1.4826 for asymptotically normal consistency.

In each scenario, coverage, power and average causal effect estimates, standard errors, $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ and I^2_{GX} statistics (which quantify the magnitude of violation of the NOME assumption in IVW and MR-Egger regression estimates, respectively[8,14]) were obtained across 10,000 simulated datasets. Power was defined as the proportion of times that 95% confidence intervals excluded zero, and coverage as the proportion of times that 95% confidence intervals included the true causal effect.

MR methods were also applied to estimate the causal effect of plasma lipid fractions and urate levels on CHD risk. The magnitude of regression dilution bias in IVW and MR-Egger regression was assessed by the $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ and I^2_{GX} statistics, respectively. Cochran's Q test was used to test for the presence of horizontal pleiotropy (under the assumption that this is the only source of heterogeneity between $\hat{\beta}_{R_j}$s other than chance).[15] All simulations and analyses were performed using R 3.3.1 (www.r-project.org). R code for implementing the mode estimator is provided in Supplementary Materials.

Results

Performance under the causal null in the two-sample context

The results of simulation 1 – where directional horizontal pleiotropy (if any) occurs only under the InSIDE assumption – are shown in Table 1. When all instruments are valid, all methods were unbiased with type-I error rates ≤5%. As expected, MR-Egger regression (which is consistent if InSIDE holds) was the least biased method in this scenario, especially when many instruments were invalid. The four mode estimates in Table 1 were less biased and less precise than the IVW and the weighted median methods. The simple mode was more biased than the weighted mode (noticeable especially when the proportion of invalid instruments was high). Using weights derived under the NOME assumption increased bias and false rejection rates. Setting φ=0.5 reduced both bias and precision (Supplementary Table 1).

When InSIDE is violated (Table 2), again the mode estimates were less biased than IVW and weighted median methods. In this case, however, they were also less biased than MR-Egger regression estimates, which is highly sensitive to InSIDE violations[8]. The exception was for large proportions (i.e., ≥80%) of invalid instruments, where MR-Egger estimates were the least biased. This is because the degree of InSIDE violation, as quantified by an inverse variance weighted Pearson correlation between instrument strength and horizontal pleiotropic effects[8], is smaller in those situations (Supplementary Table 3). Moreover, in this scenario, the simple mode estimates where generally less biased than their weighted counterparts, and setting φ=0.5 had a smaller effect when compared to simulation 1 (and indeed only clear for the simple mode – Supplementary Table 3). The NOME assumption again increased bias and false rejection rates.

(Table 1 here)

(Table 2 here)

Power to detect a causal effect in the two-sample context

Table 3 displays the results for simulation 3 (no invalid instruments). The IVW method was the most powered to detect a causal effect, followed by the weighted median method, the weighted mode, the simple mode and MR-Egger regression. Assuming NOME reduced the bias towards the null in the weighted mode estimators and improved power. Setting φ=0.5 had no consistent effect on bias, but substantially reduced power (Supplementary Table 4).

(Table 3 here)

Performance under the causal null in overlapping samples

Table 4 displays the performance of the methods under the causal null when the samples used to estimate instrument-exposure and instrument-outcome associations overlap. MR-Egger regression presented the largest bias, followed by the weighted mode assuming NOME, the weighted mode not assuming NOME, weighted median, simple mode and IVW. Setting φ=0.5 slightly increased the bias (Supplementary Table 5). Importantly, the precision of the mode estimates was very low, suggesting that the method may be prohibitively underpowered in small samples, thus being best suited for the two-sample setting using precise summary association results. Gains in precision by making the NOME assumption was more noticeable than in the other simulations with larger sample sizes.

(Table 4 here)

Causal effect of plasma lipid fractions and urate levels on CHD risk

We used these datasets of summary association results to further explore the influence of the φ parameter on the mode methods. First, we visually explored the distribution of ratio estimates (Figure 3). In the case of LDL-C (panel A), most of the distribution was above zero, and increasing the stringency of φ did not reveal substantial multimodality, although there were some pronounced density peaks at the left of the main distribution (which corresponds to the true causal effect under

the ZEMPA assumption), which may result in attenuation of the causal effect estimate. However, setting φ=0.25 resulted in some small peaks in the main distribution which may suggest over-stringency, so we used φ=0.5 in the MR analysis. For HDL-C (panel B), the bulk of the distribution was centred close to zero, and setting φ=0.25 revealed some peaks at the left of the main distribution, suggesting that horizontal pleiotropy could lead to an apparent protective effect. Since setting φ=0.5 was sufficient to substantially reduce the density at the tails, this was used in the MR analysis. Regarding triglycerides (panel C), the main distribution was above zero, and the plot suggested that there may be negative horizontal pleiotropy, leading to an underestimation of the causal effect (φ=0.25 was used in MR analysis). Finally, in the case of urate levels (panel D), by decreasing φ it became increasingly evident that the distribution of was bi-modal, which could only be clearly distinguished by setting φ=0.25 (which was used in MR analysis) because they were similar to one another. Comparing the two distributions, the main one was the closest to zero, suggesting that horizontal pleiotropy is biasing the causal effect estimate upwards.

Results of the MR analysis are shown in Supplementary Table 6. The smallest values of $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ and I^2_{GX} were 0.996 and 0.993, respectively, suggesting that IVW and MR-Egger regression estimates were strong against regression dilution bias. P-values of the Cochran's Q test ranged from 0.0003 (urate) to 1.7×10^{-21} (HDL-C), thus providing strong statistical evidence for heterogeneity between the ratio estimates. Nevertheless, results for LDL-C and triglycerides consistently suggested risk-increasing causal effects. In the case of HDL-C, the IVW method suggested a protective effect, with 1 standard deviation increase in HDL-C being associated with a 0.254 (95% CI: 0.115; 0.393) decrease in CHD ln(odds). However, the other methods did not confirm this result, suggesting that it was due to negative horizontal pleiotropy (as suggested by visually inspecting the distribution of ratio estimates). Finally, the IVW method suggested a 0.163 (95% CI: 0.027; 0.298) increase in CHD ln(odds) per standard deviation increase in urate levels. Other methods did not confirm this finding,

suggesting that it could be a result of positive horizontal pleiotropy (as the empirical density plot suggested).

Discussion

We have proposed a new MR method – the mode estimator – for causal effect estimation using summary data of multiple genetic instruments. Its performance was evaluated in a simulation study and its application illustrated in a real data example. An overview of the summary data MR methods that we evaluated is provided in Table 5.

(Table 5 here)

In our simulations, we evaluated eight different versions of the mode estimator. Decreasing the tuning parameter φ reduced bias (at the cost of reduced precision) when horizontal pleiotropy does not violate the InSIDE assumption. However, when InSIDE is violated, a similar behaviour could only be clearly seen for the simple mode. Choosing the value of the tuning parameter φ is a bias-variance trade-off and depends on how stringent the smoothing bandwidth needs to be and how stringent it can be before being prohibitively imprecise. In our applied example, we identified the stringency required through a graphical examination, and verified that the mode estimates were powered enough to detect a causal effect between HDL-C and triglycerides on CHD risk. Moreover, in the case of urate levels, the weighted mode estimates were similarly precise to that of the IVW and weighted median methods. This suggests that it may be feasible to set φ to very stringent values in practice, especially when instruments are selected based on genome-wide significance. Evaluating a range of φ values through a graphical examination may be useful to investigate how susceptible the mode estimate is susceptible to contamination from invalid instruments.

Assuming NOME increased bias and reduced the coverage of the 95% confidence intervals in the presence of invalid instruments, but reduced regression dilution bias and improved power in the two-sample setting. However, such gains were relatively small and virtually disappeared in

simulations with larger sample sizes. Moreover, the results in the applied example were virtually identical whether or not NOME was assumed. These findings suggest that the NOME assumption is not necessary (and might be even unwarranted) when deriving weights for the mode estimator.

Although the simple mode was less precise than the weighted mode, it was less prone to bias due to violations of the InSIDE assumption. However, it was more prone to bias when InSIDE holds. Indeed, a similar pattern has been previously shown for the simple and weighted median.[10] This suggests that comparing both methods would be a useful sensitivity analysis in practice, although care must be taken since the simple mode may in some cases (as in our real data example with urate levels) be prohibitively imprecise. Importantly, all the recommendations above are general, and we strongly encourage researchers to consider study-specific factors when deciding upon these aspects. One way of doing so is to perform simulations that reflect the study-specific context and compare a range of thresholds and filters in a range of different scenarios, keeping observable parameters (e.g., sample size) constant. Such simulations would also be useful to identify how strong the violations of the assumptions must be in order to obtain the observed results, which may be a useful sensitivity analysis that will either strengthen or weaken causal inference.

In our simulations, the 95% confidence intervals of the mode estimate computed using the normal approximation presented over-coverage (i.e., coverage larger than 95%) when all instruments were valid. This may be due to the mode estimator being less influenced by outlying causal effect estimates (which is indeed the basis of the method), which correspond to the most imprecise ones when all instruments are valid. Therefore, the causal effect estimate fluctuates less around the true causal effect β (i.e., is less influenced by sampling variation). This may also explain the less pronounced over-coverage in the weighted median. We compared the normal approximation with the percentile method (Supplementary Table 7), but over-coverage in the latter was even greater when there were no or few invalid instruments. Moreover, after a certain proportion of invalid instruments (around 50%) coverage of the percentile method reduced markedly, while this occurred

gradually in the normal approximation method. We therefore proposed the latter method to compute confidence intervals of the mode estimate, but there might be better alternatives.

Another aspect of the mode method (and of the weighted median) that requires further research is regression dilution bias in the two-sample setting. Dilution in the IVW estimate and MR-Egger regression estimates can be quantified by the $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$[8] and the I^2_{GX}[14] statistics, respectively. Understanding how regression dilution bias operates in IVW and MR-Egger contributed to develop correction methods,[14] thus reinforcing the importance of research in this area regarding the mode and the weighted median methods.

Although this is the first description of using the mode as a causal effect estimate in MR, other closely related methods have already been published. For example, Guo et al.[22] have recently described a method based on bivariate comparisons of all pairs of instruments, which classify instruments as estimating or not estimating the same causal effect. The largest identified set of concordant instruments can then be used to estimate the causal effect using, for example, the IVW method. Therefore, Guo et al.'s approach also relies on the assumption that the most common causal effect estimate is a consistent estimate of the true causal effect (i.e. ZEMPA). In fact, both our approach and Guo et al.'s can be viewed as methods that fully exploit the power of the *consistency criterion* defined originally by Kang et al,[23] who used it to propose a LASSO-based variable selection procedure to detect and adjust for pleiotropic variants. Guo et al.'s method and the mode estimator (which was developed independently from their work) are very different in their implementation. Ours is designed to be simple to understand and implement, does not require selecting instruments, and is easy to extend to any weighting scheme one desires. Moreover, plotting the empirical density function using different bandwidths may be a useful tool to visually explore the distribution of the $\hat{\beta}_{R_j}$s, and provides an intuitive way to select the optimal bandwidth value. In separate work we conduct a thorough review Guo et al.'s method after translating it to the two-sample context, and suggest some simple modifications to improve its performance.[24]

It is also important to consider that there are strategies to compute the mode of continuous data other than the ones we evaluated. In preliminary simulations, the modified Silverman's rule was both generally more robust against horizontal pleiotropy than the original Silverman's rule[25] and more powered to detect a causal effect. Therefore, we opted by the modified rule. However, many other kernels and bandwidth selection rules could be used, as well as strategies that are not based on the smoothed empirical density function, such as the simple and robust parametric estimators,[17] Grenander's estimators[26] and the half-sample mode method.[16] Further research is required to translate these mode estimators into the summary data MR context and compare their performance under different scenarios.

We propose the simple and weighted mode estimators as additional MR methods that should be used in combination with other approaches in a sensitivity analysis framework. Using several methods that make different assumptions, rather than a single method, is a useful strategy to assess the robustness of the results against violations of the instrumental variable assumption[27,28]. Further developments in this area (including some aspects of the mode estimator itself) will contribute to expanding the arsenal of tools available to applied researchers to interrogate causal hypotheses with observational data.

Funding

The Medical Research Council (MRC) and the University of Bristol fund the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9].

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003; 32(1):1-22.

2. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014; 23(R1):R89-98.

3. Burgess S, Timpson NJ, Ebrahim S, Davey Smith G. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol* 2015; 44(2):379-388.

4. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med* 2007; 4(12):e352.

5. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013; 37(7):658-665.

6. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* 2015; 30(7):543-552.

7. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomisation: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* 2017; In press.

8. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* 2017; In press.

9. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015; 44(2):512-525.

10. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* 2016; 40(4):304-314.

11. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol* 2004; 33(1):21-25.

12. Harbord RM, Didelez V, Palmer TM, Meng S, Sterne JA, Sheehan NA. Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Stat Med* 2013; 32(7):1246-1258.

13. Thomas DC, Lawlor DA, Thompson JR. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Ann Epidemiol* 2007; 17(7):511-513.

14. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol* 2016; doi:10.1093/ije/dyw220.

15. Greco MF, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med* 2015; 34(21):2926-2940.

16. Bickel DR, Frühwirth R. On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis* 2006; 50(12):3500-3530.

17. Bickel DR. Robust and efficient estimation of the mode of continuous data: the mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation* 2002; 73(12):899-912.

18. Do R, Willer CJ, Schmidt EM *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet* 2013; 45(11):1345-1352.

19. Willer CJ, Schmidt EM, Sengupta S *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; 45(11):1274-1283.

20. Deloukas P, Kanoni S, Willenborg C *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; 45(1):25-33.

21. White J, Sofat R, Hemani G *et al.* Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *Lancet Diabetes Endocrinol* 2016; 4(4):327-336.

22. Guo Z, Kang H, Cai TT, Small DS. Confidence Intervals for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding. *arXiv* 2016.

23. Kang H, Zhang A, Cai T, Small D. Instrumental variables estimation with some invalid instruments, and its application to Mendelian randomisation. *JASA* 2016; 111:132–144.

24. Windmeijer F, Hartwig FP, Bowden J, Davey Smith G. Instrumental Variables Estimation of Causal Effects in the Presence of Invalid Instruments. *University of Bristol (Technical Report)* 2017.

25. Silverman BW. Density Estimation for Statistics and Data Analysis. London, UK: Chapman & Hall, 1986.

26. Grenander U. Some direct estimates of the mode. *Annals of Mathematical Statistics* 1965; 36:131-138.

27. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr* 2016; 103:965-978.

28. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* 2017; 28(1):30-42.

Tables

Table 1. Mean estimates from simulation 1: directional horizontal pleiotropy under the InSIDE assumption and zero causal effect (10,000 simulations per scenario).

Estimator	Statistic	Proportion (%) of invalid instruments (Mean $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ [%]; mean I^2_{GX} [%])										
		0 (99.7; 97.4)	10 (99.7; 97.4)	20 (99.7; 97.4)	30 (99.7; 97.4)	40 (99.7; 97.4)	50 (99.7; 97.4)	60 (99.7; 97.4)	70 (99.7; 97.4)	80 (99.7; 97.4)	90 (99.7; 97.4)	100 (99.7; 97.4)
IVW	Beta	0.000	0.081	0.159	0.238	0.315	0.394	0.473	0.550	0.629	0.707	0.784
	SE	0.015	0.058	0.078	0.092	0.102	0.109	0.114	0.117	0.118	0.117	0.115
	Coverage (%)	96.9	86.7	51.1	18.9	4.8	0.6	0.1	0.0	0.0	0.0	0.0
	Power (%)[a]	3.2	13.3	48.9	81.1	95.2	99.4	99.9	100.0	100.0	100.0	100.0
MR-Egger	Beta	0.001	0.003	0.006	0.008	0.004	0.010	0.014	0.011	0.020	0.018	0.018
	SE	0.032	0.127	0.170	0.197	0.215	0.226	0.231	0.230	0.224	0.212	0.191
	Coverage (%)	96.6	95.5	94.7	94.3	94.2	93.9	94.0	93.9	94.0	93.6	93.5
	Power (%)[a]	3.4	4.5	5.3	5.7	5.8	6.1	6.0	6.1	6.1	6.4	6.5
Weighted Median	Beta	0.000	0.008	0.020	0.037	0.076	0.168	0.305	0.433	0.541	0.624	0.692
	SE	0.020	0.021	0.023	0.026	0.031	0.038	0.043	0.044	0.044	0.043	0.043
	Coverage (%)	97.6	95.6	88.1	73.3	48.7	20.4	4.7	0.5	0.0	0.0	0.0
	Power (%)[a]	2.5	4.4	11.9	26.7	51.3	79.6	95.3	99.5	100.0	100.0	100.0
Simple Mode[b]	Beta	0.000	0.000	0.002	0.003	0.013	0.040	0.129	0.294	0.501	0.650	0.742
	SE	0.046	0.054	0.056	0.069	0.084	0.118	0.126	0.148	0.183	0.175	0.183
	Coverage (%)	99.2	98.8	98.5	97.9	96.8	87.0	37.4	9.9	5.6	4.4	4.1
	Power (%)[a]	0.8	1.2	1.5	2.1	3.2	13.0	62.6	90.1	94.4	95.6	95.9
Weighted Mode[b]	Beta	0.000	0.001	0.001	0.003	0.014	0.044	0.125	0.222	0.332	0.430	0.513
	SE	0.040	0.048	0.050	0.063	0.076	0.107	0.103	0.107	0.135	0.132	0.144
	Coverage (%)	98.5	98.0	97.6	96.6	93.8	71.1	19.5	8.2	6.8	5.6	5.1
	Power (%)[a]	1.5	2.0	2.4	3.4	6.2	28.9	80.5	91.8	93.3	94.4	94.9
Simple Mode (Under NOME)[b]	Beta	0.000	0.000	0.002	0.003	0.013	0.040	0.129	0.294	0.501	0.650	0.742
	SE	0.032	0.031	0.031	0.032	0.035	0.043	0.054	0.071	0.076	0.070	0.066
	Coverage (%)	99.1	98.7	98.1	97.4	95.8	84.7	29.8	4.1	1.4	0.6	0.6
	Power (%)[a]	0.9	1.3	1.9	2.6	4.3	15.4	70.2	96.0	98.6	99.4	99.4
Weighted Mode (Under NOME)[b]	Beta	0.000	0.001	0.002	0.004	0.016	0.063	0.193	0.343	0.481	0.577	0.644
	SE	0.026	0.026	0.026	0.026	0.029	0.039	0.045	0.049	0.049	0.045	0.045
	Coverage (%)	98.3	97.6	97.2	95.8	92.3	65.8	12.0	2.3	1.2	0.7	0.8
	Power (%)[a]	1.7	2.4	2.9	4.2	7.7	34.2	88.0	97.8	98.8	99.3	99.2

InSIDE: Instrument Strength Independent on Direct Effect. IVW: Inverse-variance weighting. SE: estimated standard error. NOME: NO Measurement Error.

[a]Given that the true causal effect is zero, power can be interpreted as the type-I error rate.

[b]φ=1.

Table 2. Mean estimates from simulation 2: directional horizontal pleiotropy mediated by a single confounder of the exposure-outcome association (so violating the InSIDE assumption) and zero causal effect (10,000 simulations per scenario).

Estimator	Statistic	Proportion (%) of invalid instruments (Mean $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ [%]; mean I^2_{GX} [%])										
		0 (99.7; 97.4)	10 (99.7; 97.6)	20 (99.7; 97.9)	30 (99.8; 98.0)	40 (99.8; 98.1)	50 (99.8; 98.2)	60 (99.8; 98.2)	70 (99.8; 98.2)	80 (99.8; 98.2)	90 (99.8; 98.1)	100 (99.8; 98.0)
IVW	Beta	0.000	0.066	0.119	0.162	0.199	0.231	0.257	0.281	0.302	0.321	0.337
	SE	0.015	0.031	0.037	0.039	0.040	0.039	0.038	0.037	0.035	0.033	0.030
	Coverage (%)	96.9	44.2	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Power (%)	3.2	55.8	97.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MR-Egger	Beta	0.001	0.111	0.188	0.240	0.274	0.294	0.303	0.302	0.288	0.263	0.223
	SE	0.032	0.067	0.080	0.086	0.090	0.091	0.092	0.092	0.090	0.088	0.084
	Coverage (%)	96.6	61.2	35.3	20.9	14.0	10.7	9.4	9.7	11.7	16.7	26.5
	Power (%)	3.4	38.8	64.7	79.1	86.0	89.3	90.6	90.3	88.3	83.3	73.5
Weighted Median	Beta	0.000	0.019	0.047	0.103	0.176	0.234	0.269	0.292	0.308	0.319	0.328
	SE	0.020	0.022	0.024	0.028	0.029	0.027	0.025	0.023	0.022	0.021	0.020
	Coverage (%)	97.6	88.5	55.8	18.1	2.8	0.2	0.0	0.0	0.0	0.0	0.0
	Power (%)	2.5	11.5	44.2	81.9	97.2	99.8	100.0	100.0	100.0	100.0	100.0
Simple Mode[b]	Beta	0.000	0.001	0.003	0.010	0.024	0.060	0.138	0.231	0.289	0.316	0.326
	SE	0.046	0.045	0.042	0.047	0.050	0.063	0.073	0.066	0.055	0.047	0.043
	Coverage (%)	99.2	99.0	98.7	98.0	95.9	85.9	55.6	21.6	6.0	1.4	0.6
	Power (%)	0.8	1.1	1.3	2.0	4.1	14.1	44.5	78.4	94.0	98.6	99.4
Weighted Mode[b]	Beta	0.000	0.002	0.008	0.035	0.102	0.190	0.248	0.282	0.298	0.307	0.312
	SE	0.040	0.039	0.041	0.051	0.054	0.050	0.043	0.037	0.031	0.029	0.027
	Coverage (%)	98.5	98.1	96.2	88.1	64.9	31.5	10.5	2.9	1.0	0.4	0.2
	Power (%)	1.5	1.9	3.8	11.9	35.1	68.5	89.6	97.2	99.0	99.6	99.8
Simple Mode (Under NOME)[b]	Beta	0.000	0.001	0.003	0.010	0.024	0.060	0.138	0.231	0.289	0.316	0.326
	SE	0.032	0.031	0.032	0.034	0.040	0.056	0.067	0.060	0.051	0.044	0.042
	Coverage (%)	99.1	98.9	98.5	97.9	95.6	85.6	54.7	21.0	5.6	1.2	0.6
	Power (%)	0.9	1.1	1.5	2.1	4.4	14.5	45.3	79.1	94.4	98.8	99.4
Weighted Mode (Under NOME)[b]	Beta	0.000	0.002	0.010	0.048	0.127	0.218	0.271	0.298	0.311	0.318	0.322
	SE	0.026	0.026	0.030	0.040	0.045	0.041	0.035	0.030	0.027	0.026	0.026
	Coverage (%)	98.3	97.8	95.3	84.1	57.2	24.6	7.3	1.6	0.4	0.2	0.1
	Power (%)	1.7	2.2	4.7	15.9	42.8	75.4	92.7	98.4	99.6	99.8	99.9

InSIDE: Instrument Strength Independent on Direct Effect. IVW: Inverse-variance weighting. SE: estimated standard error. NOME: NO Measurement Error.

[a]Given that the true causal effect is zero, power can be interpreted as the type-I error rate.

[b]φ=1.

Table 3. Mean estimates from simulation 3: no horizontal pleiotropy and causal effect β=0.1 (10,000 simulations per scenario). Sample sizes N_X and N_Y are in thousands.

Estimator	Statistic	N	Mean $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ [%]; mean I^2_{GX} [%]								
			99.3; 94.8	99.3; 94.8	99.3; 94.8	99.7; 97.4	99.7; 97.4	99.7; 97.4	99.8; 98.7	99.8; 98.7	99.8; 98.7
		N_X	25	25	25	50	50	50	100	100	100
		N_Y	25	50	100	25	50	100	25	50	100
IVW	Beta		0.099	0.099	0.099	0.099	0.099	0.100	0.100	0.100	0.100
	SE		0.021	0.015	0.011	0.021	0.015	0.011	0.021	0.015	0.010
	Coverage (%)		96.5	96.5	96.4	96.7	96.3	96.7	96.1	96.7	97.0
	Power (%)		99.8	100.0	100.0	99.8	100.0	100.0	99.7	100.0	100.0
MR-Egger	Beta		0.096	0.096	0.096	0.098	0.098	0.098	0.099	0.099	0.099
	SE		0.045	0.032	0.023	0.046	0.032	0.023	0.046	0.033	0.023
	Coverage (%)		96.7	96.2	96.2	96.8	96.5	96.5	96.1	96.8	96.6
	Power (%)		53.7	82.1	97.8	54.2	84.0	98.1	54.9	83.9	98.5
Weighted Median	Beta		0.099	0.098	0.098	0.099	0.099	0.099	0.100	0.099	0.100
	SE		0.029	0.020	0.015	0.029	0.020	0.014	0.029	0.020	0.014
	Coverage (%)		97.3	97.1	97.1	97.1	97.2	97.4	97.0	97.4	97.9
	Power (%)		95.3	100.0	100.0	95.5	100.0	100.0	95.2	100.0	100.0
Simple Mode[a]	Beta		0.099	0.098	0.099	0.099	0.099	0.100	0.100	0.099	0.100
	SE		0.087	0.061	0.047	0.073	0.045	0.035	0.053	0.037	0.027
	Coverage (%)		99.0	99.1	98.9	99.1	99.0	99.1	98.8	98.8	99.2
	Power (%)		59.4	85.7	94.1	61.3	88.6	96.9	64.0	91.0	98.2
Weighted Mode[a]	Beta		0.097	0.097	0.097	0.098	0.098	0.099	0.099	0.098	0.099
	SE		0.079	0.055	0.043	0.065	0.040	0.031	0.044	0.031	0.022
	Coverage (%)		98.4	98.4	98.2	98.3	98.1	98.2	98.0	98.3	98.4
	Power (%)		75.2	90.9	94.7	77.5	94.5	97.1	80.0	96.7	98.7
Simple Mode (Under NOME)[a]	Beta		0.099	0.098	0.099	0.099	0.099	0.100	0.100	0.099	0.100
	SE		0.047	0.033	0.023	0.046	0.032	0.023	0.045	0.033	0.023
	Coverage (%)		98.8	98.9	98.8	99.1	98.9	99.0	98.8	98.9	99.1
	Power (%)		64.1	91.1	98.8	64.2	91.4	99.2	64.9	91.7	99.3
Weighted Mode (Under NOME)[a]	Beta		0.099	0.098	0.098	0.099	0.099	0.099	0.100	0.099	0.100
	SE		0.038	0.027	0.019	0.038	0.026	0.019	0.037	0.026	0.018
	Coverage (%)		98.1	98.0	97.9	98.1	97.9	98.0	97.9	98.3	98.3
	Power (%)		81.5	96.6	99.3	81.0	97.2	99.5	81.5	97.6	99.8

N_X: sample size of the dataset used to estimate instrument-exposure associations. N_Y: sample size of the dataset used to estimate instrument-outcome associations. IVW: Inverse-variance weighting. SE: estimated standard error. NOME: NO Measurement Error.

[a]φ=1.

Table 4. Mean estimates from simulation 4: no horizontal pleiotropy, zero causal effect and partially (50%) of fully (100%) overlapping samples (10,000 simulations per scenario).

Estimator	Statistic	N	Mean $\frac{\bar{F}_{GX}-1}{\bar{F}_{GX}}$ [%]; mean I^2_{GX} [%]					
			89.2; 21.8	97.0; 77.0	98.4; 87.6	89.3; 22.0	97.0; 77.0	98.4; 87.7
			Sample overlap=50%			Sample overlap=100%		
		N_X	1,000	5,000	10,000	1,000	5,000	10,000
		N_Y	1,000	5,000	10,000	1,000	5,000	10,000
IVW	Beta		0.033	0.008	0.004	0.066	0.016	0.008
	SE		0.093	0.046	0.033	0.093	0.046	0.033
	Coverage (%)		94.9	96.3	96.2	90.6	95.2	95.7
	Power (%)		5.1	3.7	3.8	9.4	4.8	4.3
MR-Egger	Beta		0.055	0.020	0.012	0.109	0.041	0.024
	SE		0.176	0.096	0.070	0.176	0.095	0.070
	Coverage (%)		95.6	96.2	96.1	92.4	94.6	95.2
	Power (%)		4.4	3.8	3.9	7.6	5.4	4.8
Weighted Median	Beta		0.039	0.010	0.005	0.078	0.022	0.011
	SE		0.131	0.063	0.045	0.130	0.063	0.045
	Coverage (%)		96.8	97.2	97.3	94.0	96.5	96.7
	Power (%)		3.2	2.8	2.7	6.0	3.5	3.3
Simple Mode[b]	Beta		0.035	0.003	0.003	0.068	0.013	0.006
	SE		2.001	0.403	0.209	1.967	0.362	0.209
	Coverage (%)		99.1	99.1	99.1	98.8	99.1	99.2
	Power (%)[a]		0.9	0.9	0.9	1.2	0.9	0.8
Weighted Mode[b]	Beta		0.041	0.007	0.005	0.076	0.022	0.010
	SE		1.969	0.387	0.197	1.936	0.345	0.197
	Coverage (%)		99.3	98.7	98.6	98.2	98.2	98.3
	Power (%)[a]		0.7	1.4	1.4	1.8	1.8	1.7
Simple Mode (Under NOME)[b]	Beta		0.035	0.003	0.003	0.068	0.013	0.006
	SE		0.213	0.105	0.074	0.212	0.104	0.073
	Coverage (%)		98.4	98.9	99.0	97.8	98.8	99.0
	Power (%)[a]		1.6	1.1	1.0	2.2	1.3	1.0
Weighted Mode (Under NOME)[b]	Beta		0.046	0.007	0.005	0.087	0.023	0.010
	SE		0.168	0.086	0.060	0.167	0.085	0.060
	Coverage (%)		96.8	97.7	97.9	93.6	96.9	97.6
	Power (%)[a]		3.2	2.3	2.1	6.4	3.1	2.4

N_X: sample size of the dataset used to estimate instrument-exposure associations. N_Y: sample size of the dataset used to estimate instrument-outcome associations. IVW: Inverse-variance weighting. SE: estimated standard error. NOME: NO Measurement Error.

[a]Given that the true causal effect is zero, power can be interpreted as the type-I error rate.

[b]φ=1.

Table 5. Breakdown level and assumptions regarding horizontal pleiotropy of the inverse variance weighted (IVW), MR-Egger regression, weighted median and weighted mode estimator methods.

Method	Breakdown level	Assumptions regarding horizontal pleiotropy
IVW	0%	Consistent if the sum of horizontal pleiotropic effects of all instruments is zero.
MR-Egger regression	100%	Consistent even if all instruments are invalid if InSIDE holds.
Weighted median	50%	Consistent if less than 50% of the weight is contributed by valid instruments, regardless of the type of horizontal pleiotropy.
Simple mode[a]	$\frac{L - [\mathrm{Max}(n_1, \dots, n_k) + 1]}{L}$	Consistent if the most common horizontal pleiotropy value is zero (i.e., ZEMPA), regardless of the type of horizontal pleiotropy.
Weighted mode[b]	$\frac{w - [\mathrm{Max}(w_1, \dots, w_k) + 1]}{w}$	Consistent if the largest weights among the k subsets are contributed by valid instruments (i.e., ZEMPA), regardless of the type of horizontal pleiotropy.

InSIDE: Instrument Strength Independent on Direct Effect. ZEMPA: ZEro Mode Pleiotropy Assumption.
[a]The breakdown level formula requires that $\mathrm{Max}(n_1, \dots, n_k)/L$ <50%.
[b]The breakdown level formula requires that $\mathrm{Max}(w_1, \dots, w_k)/w$ <50%.

Figure legends

Figure 1. Illustration of the ZEro Modal Pleiotropy Assumption (ZEMPA) in the simple mode estimate.

$\hat{\beta}_M$ is the simple mode causal effect estimate and β is the true causal effect. n_l denotes the number of variants with a given horizontal pleiotropic effect (n_0 denotes the number of valid instruments). Panel A: ZEMPA is satisfied. Panel B: ZEMPA is violated.

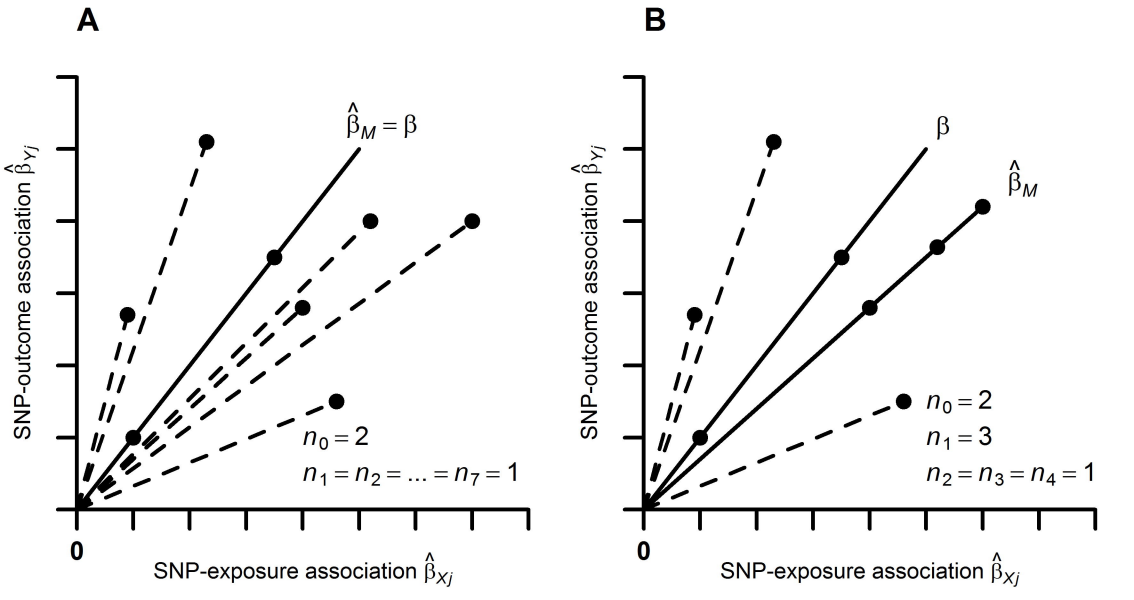
Figure 2. Histograms and smoothed empirical density plots of a hypothetical bimodal variable V (true mode is zero).

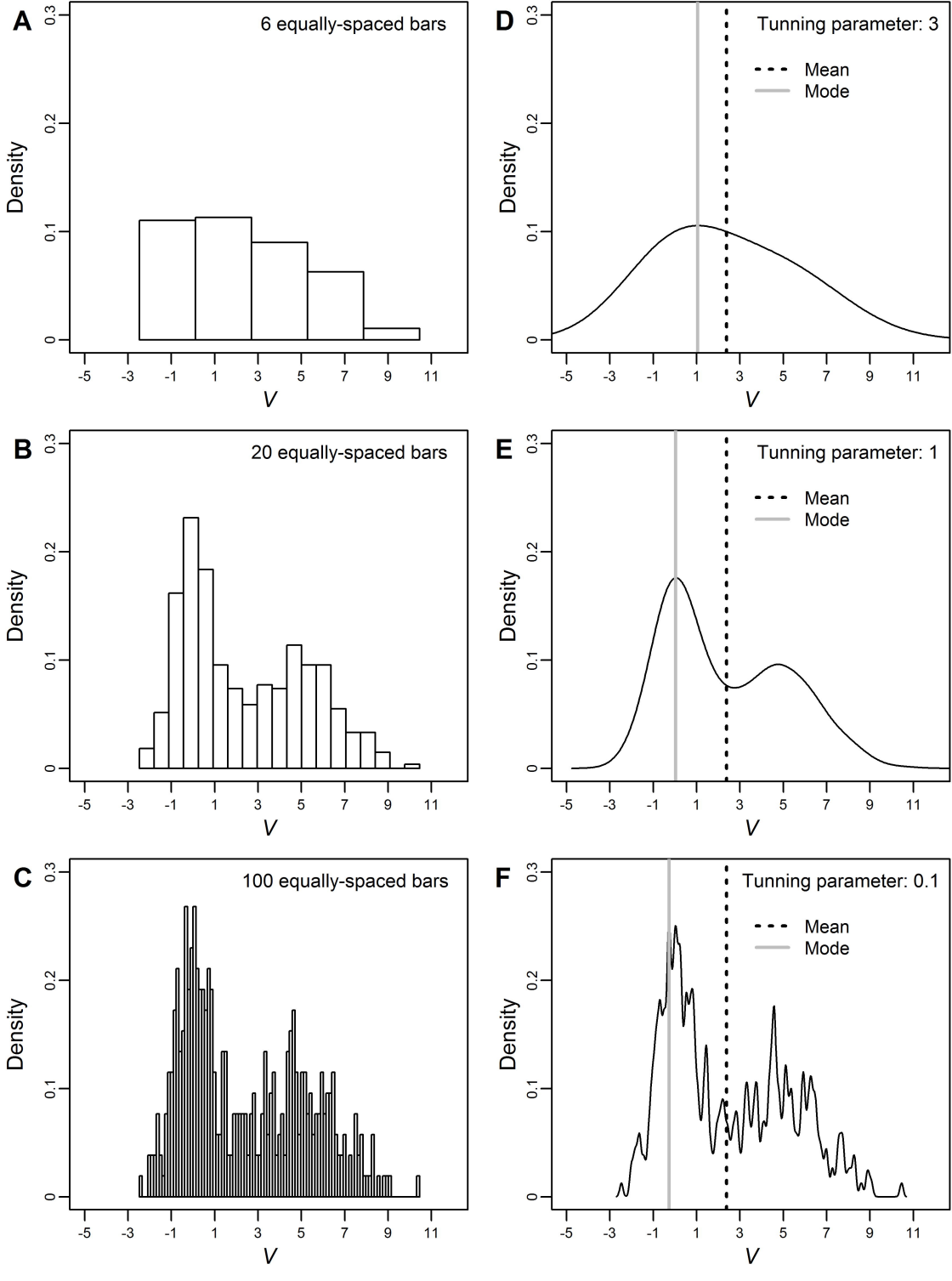
Figure 3. Weighted[a] empirical density function of all individual-instrument ratio causal effect estimates ($\widehat{\beta}_{R_j}$) of plasma LDL-C (panel A), HDL-C (panel B), triglycerides (panel C) and urate (panel D) levels on coronary heart disease risk for different values of the tuning parameter φ.

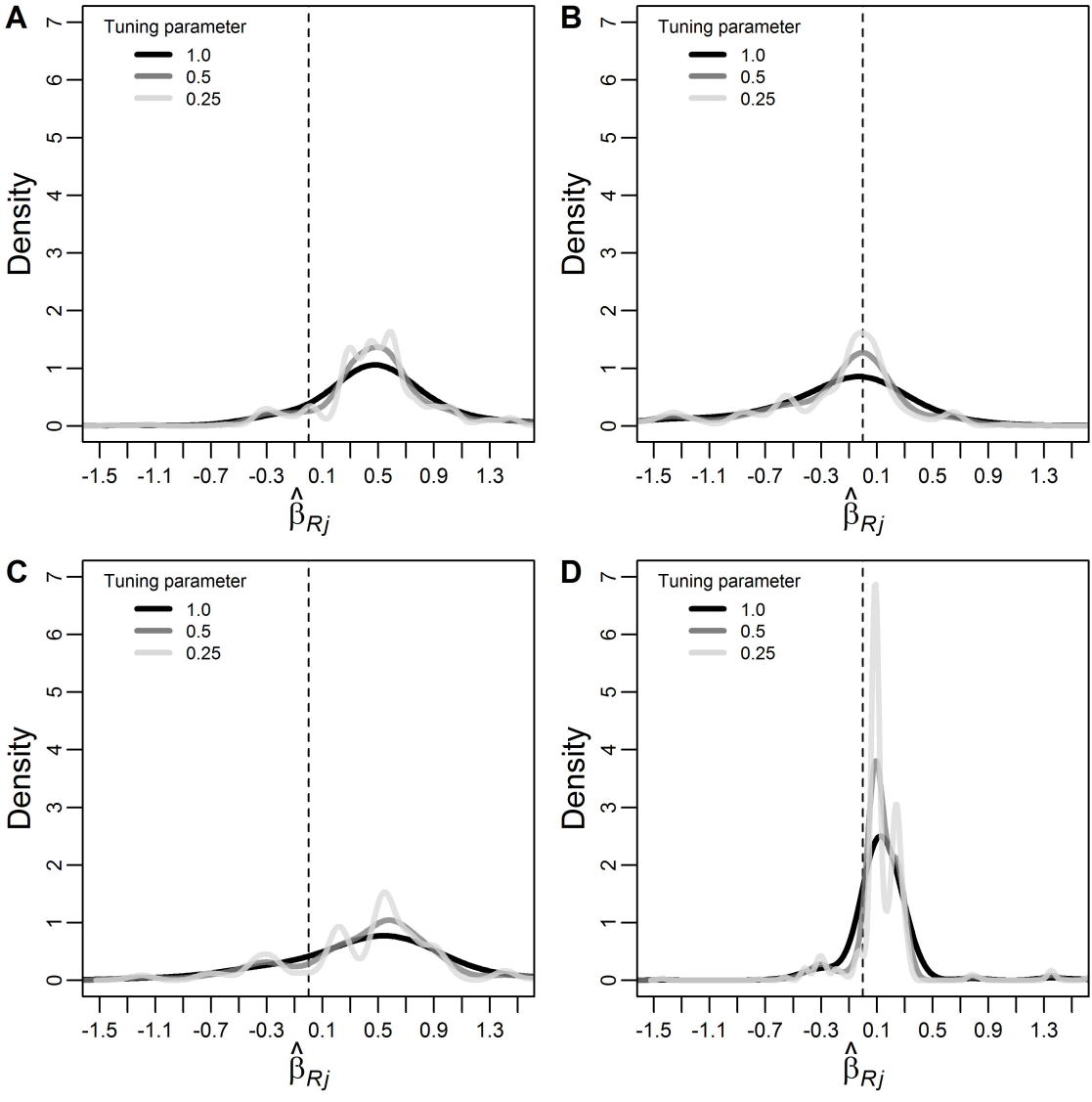
LDL-C: low-density lipoprotein cholesterol. HDL-C: high-density lipoprotein cholesterol.

The dashed line indicates the zero value.

[a]Weights were calculated without making the NOME assumption.


A
$\hat{\beta}_M = \beta$
$n_0 = 2$
$n_1 = n_2 = \ldots = n_7 = 1$
SNP-outcome association $\hat{\beta}_{Yj}$
SNP-exposure association $\hat{\beta}_{Xj}$
0
B
β
$\hat{\beta}_M$
$n_0 = 2$
$n_1 = 3$
$n_2 = n_3 = n_4 = 1$
SNP-outcome association $\hat{\beta}_{Yj}$
SNP-exposure association $\hat{\beta}_{Xj}$
0


A
6 equally-spaced bars
B
20 equally-spaced bars
C
100 equally-spaced bars
D
Tunning parameter: 3
E
Tunning parameter: 1
F
Tunning parameter: 0.1
Mean
Mode
Density
V
0
0.1
0.2
0.3
-5
-3
-1
1
3
5
7
9
11


A
B
C
D
Tuning parameter
1.0
0.5
0.25
Density
0
1
2
3
4
5
6
7
-1.5
-1.1
-0.7
-0.3
0.1
0.5
0.9
1.3
$\hat{\beta}_{Rj}$