International Journal of Epidemiology, 2017, 1–14
doi: 10.1093/ije/dyx102
Original article

Original article

Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption

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

[1]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil, [2]MRC Integrative Epidemiology Unit and [3]School of Social and Community Medicine, University of Bristol, Bristol, UK

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

Editorial decision 16 May 2017; Accepted 22 May 2017

Abstract

Background: Mendelian randomization (MR) is being increasingly used to strengthen 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 using summary data methods, typically 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-based estimate (MBE) - is proposed to obtain a single causal effect estimate from multiple genetic instruments. The MBE 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 evaluate 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 MBE presented less bias and lower type-I error rates than other methods under the null in many situations. Its power to detect a causal effect was smaller compared with 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 MBE relaxes the instrumental variable assumptions, and should be used in combination with other approaches in sensitivity analyses.

Key words: Causality, instrumental variables, genetic variation, Mendelian randomization, genetic pleiotropy

Key Messages

- Summary data Mendelian randomization, typically in a two-sample setting, is being increasingly used due to the availability of summary association results from large genome-wide association studies.
- Mendelian randomization analyses using multiple genetic instruments are prone to bias due to horizontal pleiotropy, especially when genetic instruments are selected based solely on statistical criteria.
- A causal effect estimate robust to horizontal pleiotropy can be obtained using the mode-based estimate (MBE).
- The MBE 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, and to understand how the MBE is determined.

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 randomization (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 the instrumental variable assumptions, which require that the genetic variant is: (i) associated with the exposure; (ii) independent of confounders of the exposure-outcome association; and (iii) independent of the outcome after conditioning on the exposure and all exposure-outcome confounders.

Recent MR methods allow performing MR with multiple genetic instruments, typically single nucleotide polymorphisms (SNPs), using summary data estimates from genome-wide association studies (GWAS).[4] Given the increasing number of publicly available summary statistics from large GWAS consortia, summary data MR methods enable many causal hypotheses to be rapidly interrogated without the administrative burden and cooperation required to perform equivalent individual-level data analyses.[5,6]

However, using many instruments in an MR analysis increases the probability of including at least one invalid instrument, which could easily bias 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 sum to zero) and that such pleiotropic effects are independent of instrument strength across all variants (i.e. the Instrument Strength Independent of Direct Effects – InSIDE – assumption).[4,7] More recently, other summary data MR methods that allow relaxion (but not elimination) of the instrumental variable assumptions regarding horizontal pleiotropy have been proposed.[8,9]

In this paper, we describe a new summary data MR method – the mode-based estimate (MBE). We clarify when this will be a consistent estimate of the causal effect, compare it with established summary data MR methods using simulations and illustrate its application using real data examples.

Methods

In order to motivate the summary data methods discussed in this paper, we assume the following data-generating model linking genetic variant G_j ($j = 1, \ldots, L$), a continuous exposure X and outcome Y for subject i:

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

$$\begin{aligned} Y_i|G_{ij} &= \beta_{Y0} + \left(\beta\beta_{Xj} + \alpha_j\right)G_{ij} + \lambda_{Yij} \\ &= \beta_{Y0} + \beta_{Yj}G_{ij} + \lambda_{Yij}. \end{aligned} \quad (2)$$

Here, β_{Xj} and $\beta_{Yj} = \left(\beta\beta_{Xj} + \alpha_j\right)$ 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 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). For simplicity, we will also assume that all L genetic variants are mutually independent of one another.

Let $\hat{\beta}_{Xj}$ and $\hat{\beta}_{Yj}$ represent the SNP-exposure and SNP-outcome association estimates 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[10,11] for the causal effect β using variant j alone is equal to:

$$\hat{\beta}_{Rj} = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}} \tag{3}$$

the standard error of which (σ_{Rj}) can be obtained using the delta method[12] as follows:

$$\sigma_{Rj} = \sqrt{\frac{\sigma^2_{Yj}}{\hat{\beta}^2_{Xj}} + \frac{\hat{\beta}^2_{Yj}\sigma^2_{Xj}}{\hat{\beta}^4_{Xj}}} \tag{4}$$

The standard error in (4) can be simplified to $\sigma_{Yj}/|\hat{\beta}_{Xj}|$ 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 referred to as the NO Measurement Error (NOME) assumption.[13]

The ratio estimate $\hat{\beta}_{Rj}$ 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_{Xj} and β_{Yj} alone. These estimates can then be used to furnish a summary data MR analysis using the framework of a meta-analysis.

Under models (1) and (2), variant j is a valid instrument when $\alpha_j = 0$ and invalid when $\alpha_j \neq 0$. When $\alpha_j \neq 0$, then $\beta_{Rj} = \beta + b_j$, where $b_j = \alpha_j/\beta_{Xj}$ (i.e. a bias term). In the Supplementary Methods (available as Supplementary data at *IJE* online), we briefly review three such summary data methods – IVW,[4] MR-Egger regression[8] and weighted median[9] – 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).

The MBE

In this paper we propose a new causal effect estimator – the MBE – 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): 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 b_j among the L variants. If all b_j terms are identical 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 non-zero identical value of b_j and n_k represents those with the largest non-zero identical value. Finally, let n_0 represent the number of valid instruments whose b_j terms are identically zero. We then have that $n_0 + n_1 + \ldots + n_k = L$. ZEMPA implies that n_0 is larger than any other n_l for l in $1, 2, \ldots, k$ (i.e., $n_0 > \max(n_1, \ldots, n_k)$). For a weighted version of the MBE, that is an MBE derived by allowing the weight given to each ratio estimate to vary, ZEMPA implies that the weights associated with the valid instruments are the largest among all k subsets of instruments (ie. $w_0 > \max(w_1, \ldots, w_k)$, where w_l is the weight contributed by the lth subset of instruments using our previous subset definition based on b_j.

The breakdown level (i.e. the maximum proportion of information that can come from invalid instruments before the method is inconsistent) of the simple (i.e. unweighted) MBE ranges from $100\left(\frac{L/2+1}{L}\right)\%$ to $100\left(\frac{L-2}{L}\right)\%$. The lower limit corresponds to the situation where there are some valid instruments, but all invalid instruments estimate the same (biased) causal effect parameter (i.e. $k = 2$) implying that ZEMPA is satisfied (i.e. $n_0 > \max(n_1, \ldots, n_k)$) if up to, but not including, half of the instruments are invalid. The upper limit corresponds to the situation where all invalid instruments estimate different causal effect parameters (i.e. $n_1 = n_2 = \ldots = n_k = 1$), implying that ZEMPA would be satisfied if just two variants were valid ($n_0 = 2$) and the remainder ($L - 2$) were invalid. Given that $\max(n_1, \ldots, n_k)$ is often unknown and is likely to vary depending on the set of genetic instruments and the outcome variable, the true breakdown level of the MBE in any given applied investigation is difficult to determine.

For example, in Figure 1A, six out of eight 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 MBE 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 MBE will be inconsistent for β (i.e. asymptotically biased). This is illustrated in Figure 1B, which shows causal effect estimates from six invalid and two valid variants ($n_0 = 2$). Since three variants have precisely the same horizontal pleiotropic effect in this example ($n_2 = 3$), ZEMPA is violated.

The breakdown level of the weighted MBE can be similarly defined as ranging from 50% (exclusive) to 100% (exclusive). In other words, the weighted MBE is biased if $w_0 < w_l$ for some l. Of note, the limits are open intervals because the weights are real numbers, unlike number of instruments (in the case of the simple MBE), which is a natural number. However, as L increases, then the lower and upper limits of the breakdown level of the simple MBE also tend to 50% and 100%, respectively.

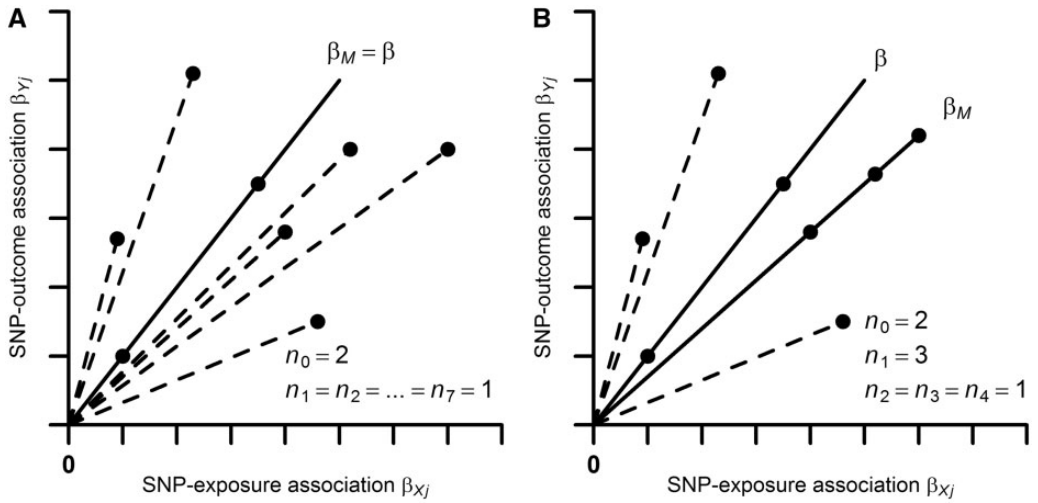

Figure 1. Illustration of the ZEro Modal Pleiotropy Assumption (ZEMPA) in the simple (i.e. unweighted) mode-based estimate (MBE). β_M is the simple MBE causal effect and β is the true causal effect; n_l denotes the number of variants with a given horizontal pleiotropic effect (n_0 denotes the number of valid instruments). Panel A: ZEMPA is satisfied. Panel B: ZEMPA is violated. SNP, single nucleotide polymorphism.

Implementing the MBE

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

$$w_j = \sigma_{R_j}^{-2} / \sum_{j=1}^{L} \sigma_{R_j}^{-2} \tag{5}$$

For the simple MBE, $w_1 = w_2 = \ldots = w_L = 1/L$.

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

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

where h is the smoothing bandwidth parameter.[14] The causal effect estimate obtained using the MBE method $\hat{\beta}_M$ is the value of x that maximizes $f(x)$ (i.e. $f(\hat{\beta}_M) = \max[f(x)]$). The h parameter regulates a bias-variance trade-off of the MBE, 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 chosen according to some criterion. We used the modified Silverman's bandwidth rule proposed by Bickel[15]:

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

where $\mathrm{sd}(\hat{\beta}_{RJ})$ and $\mathrm{mad}\left(\hat{\beta}_{RJ}\right)$ are the standard deviation and median absolute deviation from the median of the L $\hat{\beta}_{Rj}$s, respectively. An intuitive explanation of the MBE based on an analogy with histograms is provided in the Supplementary Methods (available as Supplementary data at *IJE* online).

Simulation 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 $(G_{i1}, \ldots, G_{iL})$:

$$U_i = \gamma_U Z_{Ui} + \varepsilon_{Ui} \tag{8}$$
$$X_i = \gamma_X Z_{Xi} + \theta_X U_i + \varepsilon_{Xi} \tag{9}$$
$$Y_i = \gamma_Y Z_{Yi} + \beta X_i + \theta_Y U_i + \varepsilon_{Yi} \tag{10}$$

where:

$$Z_{Ui} = \left(\sum_{j=1}^{L} \delta_{Uj} G_{ij}\right)/\sigma_{ZU},\ Z_{Xi} = \left(\sum_{j=1}^{L} \delta_{Xj} G_{ij}\right)/\sigma_{ZX},$$

$$Z_{Yi} = \left(\sum_{j=1}^{L} \delta_{Yj} G_{ij}\right)/\sigma_{ZY}.$$

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 δ_{Uj}, δ_{Xj}, δ_{Yj} ($j = 1, \ldots L$). 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 $\varepsilon_{Ui},\ \varepsilon_{Xi},\ \varepsilon_{Yi}$ 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, $\beta = 0.1$ implies that one standard deviation increment in X causes a 0.1 standard deviation increment in Y, and that the causal effect of X on Y explains $0.1^2 = 1\%$ of Y variance. A summary data interpretation of our simulation model is provided in the Supplementary Methods (available as Supplementary data at *IJE* online).

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 datasets in comparison with other methods. Therefore, we evaluated our proposed estimator in four different simulation scenarios. In all simulations, the number of variants $L = 30$, $\theta_X = \theta_Y = \sqrt{0.3}$, $\gamma_X = \sqrt{0.1}$ and $\gamma_U = \gamma_Y = \rho\sqrt{0.1}/L$, where $\rho = 0,\ 3,\ 6, \ldots, 30$ is the number of invalid instruments.

Simulations 1 and 2 were aimed at evaluating the performance of the MBE under the causal null ($\beta = 0$) 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_{Uj} was 0 for all instruments, implying that there is no InSIDE-violating horizontal pleiotropy. InSIDE-respecting horizontal pleiotropic effects δ_{Yj} were drawn from a Uniform(0.01, 0.2) distribution for the ρ invalid instruments or were set to 0 for valid instruments. Given that $\beta = 0$, power can be interpreted as the type-I error rate.

Simulation 2. InSIDE-violating horizontal pleiotropy was induced by setting $\delta_{Yj} = 0$ for all instruments, whereas δ_{Uj} 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 $\beta = 0.1$ in the two-sample context. $\rho = 0$, implying that there is no horizontal pleiotropy, and $N_X \in$ {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 for simulation 3, except $\beta = 0$ and $N_X = N_Y \in$ {1 000, 5 000, 10 000}. We used smaller sample sizes 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[16] 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[17] and the Coronary Artery Disease Genome-Wide Replication and Meta-Analysis Consortium,[18] 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 ($P < 1 \times 10^{-8}$), resulting in 73 instruments for LDL-C, 85 for HDL-C and 31 for triglycerides.

White and colleagues[19] performed a similar analysis, but with plasma urate levels rather than lipid fractions. 31 variants associated with urate levels ($P < 5 \times 10^{-7}$) were used as genetic instruments, and the required summary statistics were obtained from the GWAS catalogue [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,[7] multiplicative random effects MR-Egger regression[7] and weighted median, all implemented using inverse-variance weights calculated under NOME), as well as the simple and the weighted MBEs. Each version of the MBE was evaluated using weights calculated with and without making the NOME assumption, thus yielding four MBEs. Each of these four methods was evaluated for two values of the tuning parameter $\varphi \in \{1, 0.5\}$, totalling eight versions of the MBE method. Parametric bootstrap was used to estimate the standard errors of the MBE using the median absolute deviation from the median (multiplied by 1.4826 for asymptotically normal consistency) of the bootstrap distribution of causal effect estimates. These were used to derive symmetrical confidence intervals.

In each scenario, coverage, power and average causal effect estimates, standard errors, $\frac{\overline{F}_{GX}-1}{\overline{F}_{GX}}$ and I^2_{GX} statistics (which quantify the magnitude of violation of the NOME assumption in IVW and MR-Egger regression estimates, respectively[7,13]) 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 $\frac{\overline{F}_{GX}-1}{\overline{F}_{GX}}$ 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 $\hat{\beta}_{Rj}$s other than chance).[20] All simulations and analyses were performed using R 3.3.1 [www.r-project.org]. R code for implementing the MBE is provided in Supplementary Methods (available as Supplementary data at *IJE* online).

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 were valid, all methods were unbiased with type-I error rates $\leq$ 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 MBEs in Table 1 were less biased and less precise than the IVW and the weighted median methods. The simple MBE was more biased than the weighted MBE (noticeable especially when the proportion of invalid instruments was high). Using weights derived under the NOME assumption increased bias and false rejection rates. Setting $\varphi = 0.5$ (i.e. setting the bandwidth to half of the default value) reduced both bias and precision (Supplementary Table 1, available as Supplementary data at *IJE* online).

When InSIDE is violated (Table 2), again the MBEs 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 known to be highly sensitive to InSIDE violation.[8] The exception was for large proportions (i.e. $\geq$ 80%) of invalid instruments, where MR-Egger estimates were the least biased. This is because the degree of InSIDE violation, as quantified by the inverse-variance weighted Pearson correlation between instrument strength and horizontal pleiotropic effects,[8] is smaller in those situations (Supplementary Table 2, available as Supplementary data at *IJE* online). Moreover, in this scenario, the simple MBE was generally less biased than the weighted counterparts, and setting $\varphi = 0.5$ had a smaller effect when compared with simulation 1 (and indeed only clear for the simple MBE–Supplementary Table 3). The NOME assumption again increased bias and false rejection rates.

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 MBE, the simple MBE and MR-Egger regression. Assuming NOME reduced the bias towards the null in the weighted MBEs and improved power. Setting $\varphi = 0.5$ had no consistent effect on bias, but substantially reduced power (Supplementary Table 4, available as Supplementary data at *IJE* online).

Performance under the causal null in overlapping samples

Supplementary Table 5 (available as Supplementary data at *IJE* online) 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 MBE assuming NOME, the weighted MBE not assuming NOME, weighted median, simple MBE and IVW. Setting $\varphi = 0.5$ slightly increased the bias (Supplementary Table 6, available as Supplementary data at *IJE* online). Importantly, the

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 $\frac{\overline{F}_{GX}-1}{\overline{F}_{GX}}$ [%]; 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	Beta	0.000	0.008	0.020	0.037	0.076	0.168	0.305	0.433	0.541	0.624	0.692
Median	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	Beta	0.000	0.000	0.002	0.003	0.013	0.040	0.129	0.294	0.501	0.650	0.742
MBE[b]	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	Beta	0.000	0.001	0.001	0.003	0.014	0.044	0.125	0.222	0.332	0.430	0.513
MBE[b]	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	Beta	0.000	0.000	0.002	0.003	0.013	0.040	0.129	0.294	0.501	0.650	0.742
MBE	SE	0.032	0.031	0.031	0.032	0.035	0.043	0.054	0.071	0.076	0.070	0.066
(Under	Coverage (%)	99.1	98.7	98.1	97.4	95.8	84.7	29.8	4.1	1.4	0.6	0.6
NOME)[b]	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	Beta	0.000	0.001	0.002	0.004	0.016	0.063	0.193	0.343	0.481	0.577	0.644
MBE	SE	0.026	0.026	0.026	0.026	0.029	0.039	0.045	0.049	0.049	0.045	0.045
(Under	Coverage (%)	98.3	97.6	97.2	95.8	92.3	65.8	12.0	2.3	1.2	0.7	0.8
NOME)[b]	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 of Direct Effect; IVW, inverse-variance weighting; SE, estimated standard error; NOME, NO Measurement Error; MBE, mode-based estimate.

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

[b]$\varphi = 1$.

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 $\frac{\overline{F}_{GX}-1}{\overline{F}_{GX}}$ [%]; 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 (%)[a]	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 (%)[a]	3.4	38.8	64.7	79.1	86.0	89.3	90.6	90.3	88.3	83.3	73.5
Weighted	Beta	0.000	0.019	0.047	0.103	0.176	0.234	0.269	0.292	0.308	0.319	0.328
Median	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 (%)[a]	2.5	11.5	44.2	81.9	97.2	99.8	100.0	100.0	100.0	100.0	100.0
Simple	Beta	0.000	0.001	0.003	0.010	0.024	0.060	0.138	0.231	0.289	0.316	0.326
MBE[b]	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 (%)[a]	0.8	1.1	1.3	2.0	4.1	14.1	44.5	78.4	94.0	98.6	99.4
Weighted	Beta	0.000	0.002	0.008	0.035	0.102	0.190	0.248	0.282	0.298	0.307	0.312
MBE[b]	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 (%)[a]	1.5	1.9	3.8	11.9	35.1	68.5	89.6	97.2	99.0	99.6	99.8
Simple	Beta	0.000	0.001	0.003	0.010	0.024	0.060	0.138	0.231	0.289	0.316	0.326
MBE	SE	0.032	0.031	0.032	0.034	0.040	0.056	0.067	0.060	0.051	0.044	0.042
(under	Coverage (%)	99.1	98.9	98.5	97.9	95.6	85.6	54.7	21.0	5.6	1.2	0.6
NOME)[b]	Power (%)[a]	0.9	1.1	1.5	2.1	4.4	14.5	45.3	79.1	94.4	98.8	99.4
Weighted	Beta	0.000	0.002	0.010	0.048	0.127	0.218	0.271	0.298	0.311	0.318	0.322
MBE	SE	0.026	0.026	0.030	0.040	0.045	0.041	0.035	0.030	0.027	0.026	0.026
(under	Coverage (%)	98.3	97.8	95.3	84.1	57.2	24.6	7.3	1.6	0.4	0.2	0.1
NOME)[b]	Power (%)[a]	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 of Direct Effect; IVW, inverse-variance weighting; SE, estimated standard error; NOME, NO Measurement Error; MBE, mode-based estimate.

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

[b]$\varphi = 1$.

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

Estimator	Statistic	N	Mean $\frac{\overline{F}_{GX}-1}{\overline{F}_{GX}}$ [%]; 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	Beta		0.099	0.098	0.098	0.099	0.099	0.099	0.100	0.099	0.100
Median	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	Beta		0.099	0.098	0.099	0.099	0.099	0.100	0.100	0.099	0.100
MBE[a]	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	Beta		0.097	0.097	0.097	0.098	0.098	0.099	0.099	0.098	0.099
MBE[a]	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	Beta		0.099	0.098	0.099	0.099	0.099	0.100	0.100	0.099	0.100
MBE	SE		0.047	0.033	0.023	0.046	0.032	0.023	0.045	0.033	0.023
(under	Coverage (%)		98.8	98.9	98.8	99.1	98.9	99.0	98.8	98.9	99.1
NOME)[a]	Power (%)		64.1	91.1	98.8	64.2	91.4	99.2	64.9	91.7	99.3
Weighted	Beta		0.099	0.098	0.098	0.099	0.099	0.099	0.100	0.099	0.100
MBE	SE		0.038	0.027	0.019	0.038	0.026	0.019	0.037	0.026	0.018
(under	Coverage (%)		98.1	98.0	97.9	98.1	97.9	98.0	97.9	98.3	98.3
NOME)[a]	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; MBE, mode-based estimate.

[a] $\varphi = 1$.

precision of the MBE 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 were more noticeable than in the other simulations with larger sample sizes.

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

We used real datasets of summary association results to further explore the influence of the φ parameter on the MBE. First, we visually explored the distribution of ratio estimates (Figure 2). 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 $\varphi = 0.25$ resulted in some small peaks in the main distribution which may suggest over-stringency, so we used $\varphi = 0.5$ in the MR analysis. For HDL-C (panel B), the bulk of the distribution was centred close to zero, and setting $\varphi = 0.25$ revealed some peaks at the left of the main distribution, suggesting that horizontal pleiotropy could lead to an apparent protective effect. Since setting $\varphi = 0.5$ 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

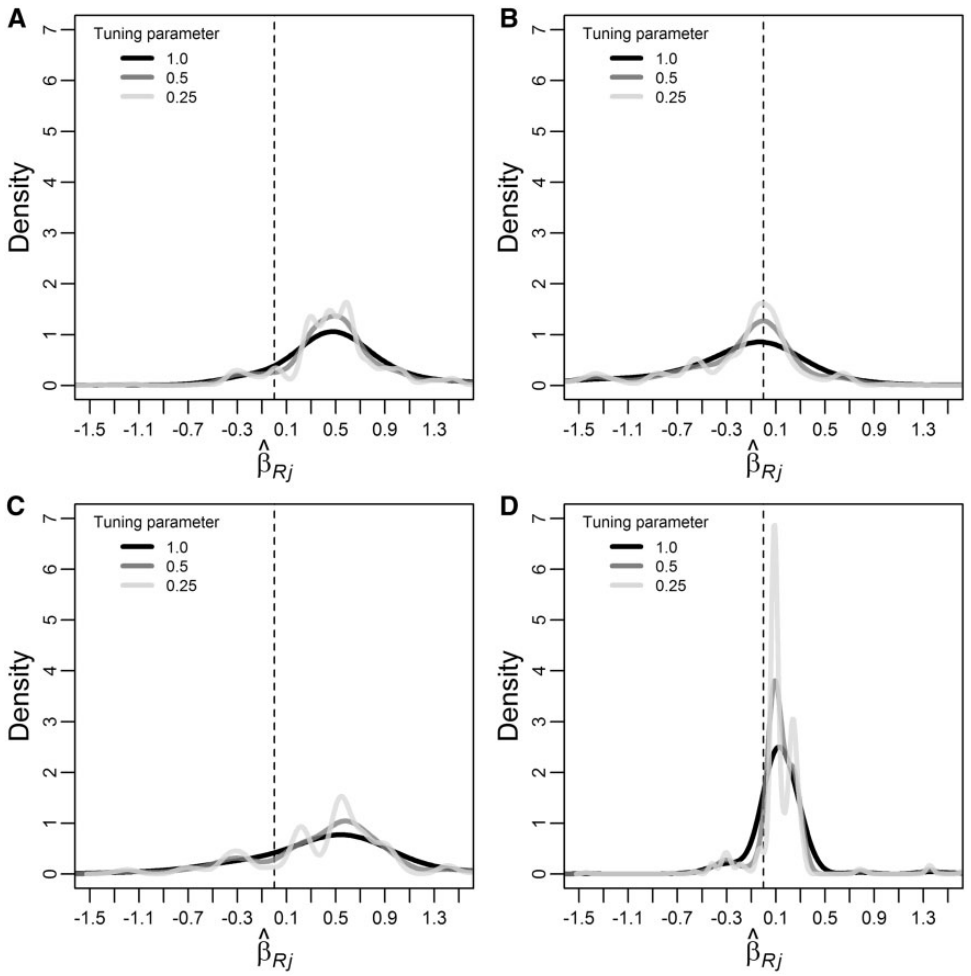

Figure 2. Weighted[a] empirical density function of all individual-instrument ratio causal effect estimates ($\hat{\beta}_{Rj}$) of plasma LDL-C (panel A), HDL-C (panel B), triglycerides (panel C) and urate (panel D) levels on ln(odds ratio) of coronary heart disease for different values of the tuning 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.

and the plot suggested that there may be negative horizontal pleiotropy, leading to an underestimation of the causal effect ($\varphi = 0.25$ was used in MR analysis). Finally, in the case of urate levels (panel D), by decreasing φ it became increasingly evident that the distribution was bi-modal, which could only be clearly distinguished by setting $\varphi = 0.25$ (which was used in MR analysis) because the main peaks 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 Table 4. 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 not materially affected by 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 one 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 MBE – for causal effect estimation using summary data of multiple

Table 4. Mendelian randomization estimates of the causal effect of urate plasma levels (in standard deviation units) on CHD risk [in ln(odds)] using 31 genetic instruments

Exposure	Estimator	Beta	SE	95% CI	*P*-value
LDL-C	IVW	0.476	0.060	0.357; 0.595	1.8×10^{-11}
	MR-Egger β_0	−0.009	0.005	−0.020; 0.001	0.083
	MR-Egger β_1	0.624	0.103	0.419; 0.828	5.3×10^{-8}
	Weighted median	0.457	0.064	0.331; 0.583	7.4×10^{-10}
	Simple MBE[a]	0.422	0.187	0.056; 0.788	0.027
	Weighted MBE[a,b]	0.491	0.109	0.276; 0.705	2.7×10^{-5}
HDL-C	IVW	−0.254	0.070	−0.393; -0.115	4.9×10^{-4}
	MR-Egger β_0	−0.014	0.005	−0.025; -0.003	0.011
	MR-Egger - β_1	−0.013	0.115	−0.241; 0.215	0.913
	Weighted median	−0.069	0.068	−0.202; 0.065	0.314
	Simple MBE[a]	−0.174	0.171	−0.509; 0.161	0.311
	Weighted MBE[a,b]	−0.003	0.088	−0.175; 0.170	0.974
Triglycerides	IVW	0.416	0.081	0.252; 0.580	6.0×10^{-6}
	MR-Egger - β_0	0.000	0.007	−0.015; 0.015	0.962
	MR-Egger - β_1	0.422	0.140	0.140; 0.704	0.004
	Weighted median	0.516	0.083	0.352; 0.679	1.5×10^{-7}
	Simple MBE[c]	0.875	0.259	0.367; 1.383	0.002
	Weighted MBE[c,b]	0.547	0.134	0.284; 0.810	1.8×10^{-4}
Urate levels	IVW	0.163	0.066	0.027; 0.298	0.020
	MR-Egger - β_0	0.008	0.005	−0.002; 0.018	0.118
	MR-Egger - β_1	0.048	0.096	−0.148; 0.245	0.614
	Weighted median	0.119	0.061	−0.001; 0.239	0.061
	Simple MBE[c]	0.188	0.163	−0.132; 0.507	0.259
	Weighted MBE[c,b]	0.092	0.066	−0.038; 0.221	0.175

LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; IVW, inverse-variance weighting; SE, standard error; CI, confidence interval; MBE, mode-based estimate.

[a] $\varphi = 0.5$.

[b] Not under the NO Measurement Error (NOME) assumption.

[c] $\varphi = 0.25$.

genetic instruments. Its performance was evaluated in a simulation study and its application illustrated in real data examples. An overview of the summary data MR methods that we evaluated (as well as the simple median) is provided in Table 5.

Consistent causal effect estimation using the MBE requires that ZEMPA holds. ZEMPA is an assumption that relates to the underlying bias parameters (the b_j) that contribute to the ratio estimand $\beta_j = \beta + b_j$ identified by the jth genetic instrument. If ZEMPA is satisfied, then the MBE yields a consistent estimate for the causal effect. However, due to imprecision in the $\hat{\beta}_j$'s in finite samples, in practice the MBE may be contaminated by some invalid invariants even if ZEMPA holds. This can be seen in our simulations, where ZEMPA is only violated when all instruments are invalid, but nevertheless there is bias in the MBE when some of the instruments are valid. In practice, the MBE also depends on the magnitude of the bias, with invalid genetic instruments identifying causal effect parameters that are close to the true causal effect being more likely to contaminate the MBE estimate. However, this also means that genetic instruments that would introduce strong bias are less likely to contaminate the MBE.

In our simulations, we evaluated eight different versions of the MBE. Decreasing the tuning parameter φ reduced bias (at the cost of reduced precision) when horizontal pleiotropy did not violate the InSIDE assumption. However, when InSIDE was violated, a similar behaviour could only be clearly seen for the simple MBE. 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 MBEs 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 MBE was similarly precise to the IVW and weighted median methods. This suggests that it may be feasible to set φ to stringent values in practice, especially when there are multiple instruments selected based on genome-wide significance. Evaluating a range of φ values through a

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

Method	Breakdown level	Assumptions regarding horizontal pleiotropy
IVW	0%	Consistent if the sum of horizontal pleiotropic effects of all instruments is zero and InSIDE holds
MR-Egger regression	100%	Consistent even if all instruments are invalid if InSIDE holds
Simple median	$100\left(\frac{L/2+1}{L}\right)\%$	Consistent if less than 50% of instruments are invalid, regardless of the type of horizontal pleiotropy
Weighted median	50% (exclusive)	Consistent if less than 50% of the weight is contributed by invalid instruments, regardless of the type of horizontal pleiotropy
Simple MBE	Ranges from $100\left(\frac{L/2+1}{L}\right)\%$ to $100\left(\frac{L-2}{L}\right)\%$	Consistent if the most common horizontal pleiotropy value is zero (i.e. ZEMPA), regardless of the type of horizontal pleiotropy
Weighted MBE	Ranges from 50% (exclusive) to 100% (exclusive)	Consistent if the largest weights among the k subsets are contributed by valid instruments (i.e. ZEMPA), regardless of the type of horizontal pleiotropy

IVW, inverse-variance weighting; InSIDE, Instrument Strength Independent of Direct Effect; ZEMPA, ZEro Mode Pleiotropy Assumption; MBE, mode-based estimate.

graphical examination may be useful to investigate how susceptible the MBE is 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 MBE.

Although the simple MBE was less precise than the weighted MBE, it was less prone to bias due to violations of the InSIDE assumption. However, it was more prone to bias when InSIDE held. Indeed, a similar pattern has been previously shown for the simple and weighted median.[9] This suggests that comparing both methods would be a useful sensitivity analysis in practice, although care must be taken since the simple MBE 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 different 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 MBE computed using the normal approximation presented over-coverage (i.e. coverage larger than 95%). This may be due to the MBE being less influenced by outlying instruments (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, available as Supplementary data at *IJE* online), 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, whereas this occurred gradually in the normal approximation method. We therefore proposed the latter method to compute confidence intervals, but there might be better alternatives.

Another aspect of the MBE method (and of the weighted median) that requires further research is regression dilution bias in the two-sample setting. Understanding how regression dilution bias operates in IVW and MR-Egger contributed to developing correction methods,[13] thus reinforcing the importance of research in this area regarding the MBE and the weighted median.

Although this is the first description of using the MBE as a causal effect estimate in MR, other closely related

methods have already been published. For example, Guo *et al.*[21] 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*,[22] who used it to propose a LASSO-based variable selection procedure to detects and adjusts for horizontally pleiotropic variants. However, Guo *et al.*'s method and the MBE (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 $\hat{\beta}_{Rj}$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.[23]

It is also important to consider that there are other strategies to compute the mode of continuous data. In preliminary simulations, the modified Silverman's rule was both generally more robust against horizontal pleiotropy than the original Silverman's rule[24] and more powered to detect a causal effect. Therefore, we opted for 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,[15] Grenander's estimators[25] and the half-sample mode method.[14] 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 MBE as an additional MR method 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 assumptions.[26,27] Further developments in this area (including some aspects of the MBE itself) will contribute to expanding the arsenal of tools available to applied researchers to interrogate causal hypotheses with observational data.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

The Medical Research Council (MRC) and the University of Bristol support the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9]. J.B. is additionally supported by an MRC Methodology Research Fellowship (grant MR/N501906/1).

Conflict of interest: None declared.

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32:**1–22.
2. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23:**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:**379–88.
4. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013;**37:**658–65.
5. 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:**543–52.
6. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* 2016;**45:**1717–26.
7. Bowden J, Del Greco MF, 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;**36:**1783–807.
8. 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:**512–25.
9. 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:**304–14.
10. Thomas DC, Conti DV. Commentary: The concept of 'Mendelian randomization'. *Int J Epidemiol* 2004;**33:**21–25.
11. 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:**1246–58.
12. 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:**511–13.
13. 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 I^2 statistic. *Int J Epidemiol* 2016;**45:**1961–74.

14. Bickel DR, Frühwirth R. On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Comput Stat Data Ana* 2006;**50:**3500–30.
15. Bickel DR. Robust and efficient estimation of the mode of continuous data: the mode as a viable measure of central tendency. *J Stat Comput Simul* 2002;**73:**899–912.
16. Do R, Willer CJ, Schmidt EM *et al*. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet* 2013;**45:**1345–52.
17. Willer CJ, Schmidt EM, Sengupta S *et al*. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; **45:**1274–83.
18. Deloukas P, Kanoni S, Willenborg C *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013;**45:**25–33.
19. White J, Sofat R, Hemani G *et al*. Plasma urate concentration and risk of coronary heart disease: a Mendelian randomization analysis. *Lancet Diabetes Endocrinol* 2016;**4:**327–36.
20. Greco MF, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomization studies with summary data and a continuous outcome. *Stat Med* 2015;**34:**2926–40.
21. Guo Z, Kang H, Cai TT, Small DS. Confidence intervals for causal effects with invalid instruments using two-stage hard thresholding. *arXiv* 2016:1603.05224 [math.ST].
22. Kang H, Zhang A, Cai T, Small D. Instrumental variables estimation with some invalid instruments, and its application to Mendelian randomization. *JASA* 2016;**111:**132–44.
23. Windmeijer F, Hartwig FP, Bowden J, Davey Smith G. Instrumental variables estimation of causal effects in the presence of invalid instruments. Technical Report. University of Bristol, 2017.
24. Silverman BW. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.
25. Grenander U. Some direct estimates of the mode. *Ann Math Stat* 1965;**36:**131–38.
26. 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:**965–78.
27. 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:**30–42.