International Journal of Epidemiology, 2018, 1242–1254
doi: 10.1093/ije/dyy080
Advance Access Publication Date: 26 May 2018
Original article

Methods

Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid

Stephen Burgess,[1,2]* Verena Zuber,[1] Apostolos Gkatzionis[1] and Christopher N Foley[1]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK and [2]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

*Corresponding author. MRC Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge, CB2 0SR, UK. E-mail: sb452@medschl.cam.ac.uk

Editorial decision 13 April 2018; Accepted 19 April 2018

Abstract

Background: A robust method for Mendelian randomization does not require all genetic variants to be valid instruments to give consistent estimates of a causal parameter. Several such methods have been developed, including a mode-based estimation method giving consistent estimates if a plurality of genetic variants are valid instruments; i.e. there is no larger subset of invalid instruments estimating the same causal parameter than the subset of valid instruments.

Methods: We here develop a model-averaging method that gives consistent estimates under the same 'plurality of valid instruments' assumption. The method considers a mixture distribution of estimates derived from each subset of genetic variants. The estimates are weighted such that subsets with more genetic variants receive more weight, unless variants in the subset have heterogeneous causal estimates, in which case that subset is severely down-weighted. The mode of this mixture distribution is the causal estimate. This heterogeneity-penalized model-averaging method has several technical advantages over the previously proposed mode-based estimation method.

Results: The heterogeneity-penalized model-averaging method outperformed the mode-based estimation in terms of efficiency and outperformed other robust methods in terms of Type 1 error rate in an extensive simulation analysis. The proposed method suggests two distinct mechanisms by which inflammation affects coronary heart disease risk, with subsets of variants suggesting both positive and negative causal effects.

Conclusions: The heterogeneity-penalized model-averaging method is an additional robust method for Mendelian randomization with excellent theoretical and practical

properties, and can reveal features in the data such as the presence of multiple causal mechanisms.

Key words: Mendelian randomization, instrumental variables, robust methods, invalid instruments, model averaging

Key Messages

- We propose a heterogeneity-penalized model-averaging method that gives consistent causal estimates if a weighted plurality of the genetic variants are valid instruments.
- The method calculates causal estimates based on all subsets of genetic variants, and up-weights subsets containing several genetic variants with similar causal estimates.
- The method is asymptotically efficient and does not rely on bootstrapping to obtain a confidence interval, nor is the confidence interval constrained to be symmetric.
- In particular, the confidence interval can include multiple disjoint intervals, suggesting the presence of multiple causal mechanisms by which the risk factor influences the outcome.
- The method can incorporate biological knowledge to up-weight the contribution of genetic variants with stronger plausibility of being valid instruments.

Introduction

Mendelian randomization is an epidemiological approach for making causal inferences from observational data by using genetic variants as instrumental variables.[1,2] If a genetic variant is a valid instrument for the risk factor, then any association of the variant with the outcome is indicative of a causal effect of the risk factor on the outcome.[3] To be a valid instrumental variable, a genetic variant must be:

- IV1: associated with the risk factor (relevance);
- IV2: independent of any confounder of the risk factor–outcome association (exchangeable);
- IV3: independent of the outcome conditional on the risk factor and confounders (exclusion restriction).

Violation of any of these assumptions means that an instrumental variable is not valid.

When there are multiple genetic variants that are all valid instrumental variables, and under certain parametric assumptions (most notably that all relationships between variables are linear and there is no effect modification), an efficient test of the causal null hypothesis as the sample size increases can be obtained using the two-stage least-squares method (based on individual-level data)[4] or equivalently the inverse-variance weighted (IVW) method (based on summarized data).[5] With uncorrelated instruments, the IVW estimate [equal to the two-stage least-squares (2SLS) estimate] is a weighted mean of the Wald (or ratio) estimates obtained separately from each individual instrumental variable.

Whereas the 2SLS/IVW estimator is asymptotically efficient, it is not robust to violations of the instrumental variable assumptions. Specifically, if a genetic variant is a valid instrument, then the ratio estimate based on that variant is a consistent estimate of the causal effect. Hence the weighted mean of these ratio estimates is a consistent estimate of the causal effect if all genetic variants are valid instruments, but not in general if at least one variant is not a valid instrument.[6] This has motivated the development of robust methods for instrumental variable analysis based on only a subset of the genetic variants being valid instruments. For example, Kang *et al.*[7] developed a method using L1-penalization that gives consistent estimates if at least 50% of the instrumental variables are valid. Bowden *et al.*[8] considered simple and weighted median methods that again are consistent if at least 50% of the candidate instrumental variables are valid; the simple median method is a median of the variant-specific ratio estimates. Most recently, Guo *et al.*[9] introduced a method that provides a consistent estimate if a plurality of the candidate instruments are valid, meaning that the largest subset of genetic variants with the same ratio estimate (in a large sample size) comprises the valid instruments. Invalid instruments may have different ratio estimates asymptotically, but the assumption is that there is no larger subset of invalid instruments with the

same ratio estimate than the subset of valid instruments. Intuitively, this means that the true causal estimate can be identified asymptotically as the mode of the variant-specific ratio estimates. In parallel, Hartwig *et al.*[10] have developed a modal-based estimation method that can be implemented using summarized data and provides a consistent estimate under this plurality assumption, which they term the 'zero modal pleiotropy assumption' (ZEMPA).

The idea of a modal-based estimate is an attractive one due to the high breakdown point of the mode as an estimator and its insensitivity to extreme values. However, there are several issues with the implementations of Guo *et al.* and Hartwig *et al.*'s methods that could be improved upon. In particular, Hartwig *et al.*'s implementation of this approach fits a kernel-density-smoothed function to the variant-specific ratio estimates, and calculates confidence intervals based on the median absolute deviation of a bootstrapped distribution. Varying the bandwidth of the kernel density can result in substantial changes to the estimate and its confidence interval, as demonstrated later in this paper. Guo *et al.*'s individual-level data method is implemented by pairwise comparison of estimates from different candidate instruments. When two genetic variants have similar estimates, they 'vote' for each other. The overall estimate is based on the set of genetic variants with the greatest number of these votes. However, as these binary votes are determined by a fixed threshold, estimates from the Guo *et al.* method (called 'two-stage hard thresholding') will be sensitive to small changes in the data when the comparison measures are close to the threshold.

In this paper, we propose an alternative way of constructing a density function for the causal effect estimate as a heterogeneity-penalized weighted mixture distribution. This approach up-weights estimates that are supported by multiple genetic variants, but severely down-weights heterogeneity. We show that the mode of this distribution will be an asymptotically consistent estimator of the causal effect if a weighted plurality of the genetic variants are valid instruments. We first introduce this method, and then we demonstrate its performance in a simulation study compared with other robust methods. We consider its behaviour in two applied examples. Finally, we discuss the results of this paper and their relevance to applied research. In particular, we consider how to incorporate biological knowledge into the weighting procedure. Software code for implementing the proposed method is provided in Supplementary Material A.1, available as Supplementary data at *IJE* online.

Methods

In this section, we first introduce the data requirements and parametric assumptions necessary for summarized data Mendelian randomization. We then recall the IVW method, and subsequently introduce the model-averaging procedure proposed in this paper.

Data requirements and assumptions

For practical reasons, many modern Mendelian randomization investigations are conducted using summarized data on genetic associations with the risk factor (X) and outcome (Y) taken from univariable regression models of the risk factor (or outcome) regressed on the genetic variants in turn.[11] We assume, as is common in applied practice, that the genetic variants are all uncorrelated (not in linkage disequilibrium). For each genetic variant G_j ($j = 1, 2, \ldots, J$), we assume that we have an estimate $\hat{\beta}_{Xj}$ of the association of the genetic variant with the risk factor obtained from linear regression. Similar association estimates are assumed to be available for the outcome ($\hat{\beta}_{Yj}$). The standard error of the association estimate with the outcome is $\text{se}(\hat{\beta}_{Yj})$. If any of the variables is binary, then these summarized association estimates may be replaced with association estimates from logistic regression; as has been shown previously, the interpretation of the causal estimate in this case is not clear due to non-collapsibility, but estimates still represent valid tests of the causal null hypothesis.[12,13] See Bowden *et al.*[14] for a more detailed exposition of the parametric assumptions typically made in summarized data Mendelian randomization investigations that are also made here.

IVW method

The ratio estimate based on genetic variant j is $\hat{\theta}_j = \hat{\beta}_{Yj}/\hat{\beta}_{Xj}$, with standard error taken as $\text{se}(\hat{\theta}_j) = \text{se}(\hat{\beta}_{Yj})/\hat{\beta}_{Xj}$ (the leading order term from the delta expansion for the standard error of the ratio of two variables). The IVW estimate is a weighted mean of the ratio estimates:

$$\hat{\theta}_{IVW} = \frac{\sum_j \hat{\theta}_j \text{se}(\hat{\theta}_j)^{-2}}{\sum_j \text{se}(\hat{\theta}_j)^{-2}} = \frac{\sum_j \hat{\beta}_{Yj}\hat{\beta}_{Xj}\text{se}(\hat{\beta}_{Yj})^{-2}}{\sum_j \hat{\beta}_{Xj}^2 \text{se}(\hat{\beta}_{Yj})^{-2}}. \quad (1)$$

The same estimate can be obtained from the weighted regression:

$$\hat{\beta}_{Yj} = \theta_{IVW}\hat{\beta}_{Xj} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, \text{se}(\hat{\beta}_{Yj})^2). \quad (2)$$

For uncorrelated variants, this estimate is also equivalent to the estimate obtained from two-stage least-squares—a method typically used for instrumental variable analysis with individual-level data.[5] These estimates do not take into account uncertainty in the genetic associations

with the risk factor; however, these associations are typically more precisely estimated than those with the outcome, and ignoring this uncertainty does not lead to inflated Type 1 error rates in realistic scenarios.[15] This is because genetic associations with the risk factor are typically estimated in larger sample sizes (as they are estimated in cross-sectional datasets, whereas associations with disease outcomes are estimated in case–control studies), because risk factors are continuous (outcomes are often binary) and because genetic variants are chosen as those having strong associations with the risk factor. If these conditions are not met, then alternative approaches are possible.[16] Additionally, we assume that the standard errors of genetic associations are known without error; as associations are typically estimated in large sample sizes, this is usually a reasonable assumption.

The standard error of the IVW estimate based on a fixed-effect meta-analysis model is:

$$\mathrm{se}(\hat{\theta}_{IVW}) = \frac{1}{\sqrt{\sum_j \mathrm{se}(\hat{\theta}_j)^{-2}}} = \frac{1}{\sqrt{\sum_j \hat{\beta}_{Xj}^2 \mathrm{se}(\hat{\beta}_{Yj})^{-2}}}. \quad (3)$$

We also consider a multiplicative random-effects model based on the weighted linear regression above:

$$\hat{\beta}_{Yj} = \theta_{IVW}\hat{\beta}_{Xj} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, \psi^2 \mathrm{se}(\hat{\beta}_{Yj})^2), \quad (4)$$

where ψ is the residual standard error. Most statistical software packages estimate this additional parameter by default in a weighted linear regression model. A fixed-effect analysis can be performed by fixing the value of ψ to 1.[17] To ensure that the standard error of the IVW estimate is never more precise than that from a fixed-effect analysis, we allow ψ to take values above 1 (corresponding to over-dispersion of the genetic association estimates), but not values below 1 (corresponding to under-dispersion). If all genetic variants estimate the same causal parameter, then ψ should tend to 1 asymptotically.

Heterogeneity-penalized model-averaging method

We seek to define a function with the property that the mode (the maximum value) of the function will tend to the true causal effect when a plurality of the genetic variants are valid instruments. For making statistical inferences, it is convenient if this function is a likelihood for the causal effect parameter. We present the method in a somewhat informal way; a more technical explanation is provided in Supplementary Material A.2, available as Supplementary data at *IJE* online. We consider a model-averaging procedure with $2^J - J - 1$ candidate models, where J is the total number of genetic variants. Each model corresponds to one of the $2^J - J - 1$ subsets of genetic variants (subsets including 0 or 1 genetic variants are ignored throughout). Our likelihood function is a mixture of $2^J - J - 1$ normal distributions, where the kth normal distribution has mean and standard deviation corresponding to the IVW estimate and standard error based on all the variants in the kth subset:

$$\hat{\theta}_{IVW,k} = \frac{\sum_{j\in\sigma_k} \hat{\theta}_j \mathrm{se}(\hat{\theta}_j)^{-2}}{\sum_{j\in\sigma_k} \mathrm{se}(\hat{\theta}_j)^{-2}} \quad (5)$$

$$\mathrm{se}(\hat{\theta}_{IVWr,k}) = \frac{\hat{\psi}_k}{\sqrt{\sum_{j\in\sigma_k} \mathrm{se}(\hat{\theta}_j)^{-2}}}, \quad (6)$$

where $\sigma_k = (\sigma_{k1}, \sigma_{k2}, \ldots, \sigma_{kJ}) : \sigma_{kj} \in \{0, 1\}$ represents a subset of the genetic variants, $j \in \sigma_k$ when $\sigma_{kj} = 1$ (this means that $\hat{\theta}_{IVW,k}$ is the IVW estimate based on all the variants in subset k) and

$$\hat{\psi}_k = \max\left(1, \quad \frac{1}{K-1}\sum_{j\in\sigma_k} \mathrm{se}(\hat{\beta}_{Yj})^{-2} (\hat{\beta}_{Yj} - \hat{\theta}_{IVW,k}\hat{\beta}_{Xj})^2\right), \quad (7)$$

where K is the number of variants included in subset k. The random-effects versions of the standard errors $\mathrm{se}(\hat{\theta}_{IVWr,k})$ are used in this mixture distribution to appropriately allow for heterogeneity between the variant-specific ratio estimates in the overall causal estimate (hence the additional subscripted r).

The weight given to each of these normal distributions is calculated as:

$$w_k = \prod_{j\in\sigma_k} \mathrm{se}(\hat{\theta}_j)^{-1} \exp\left[-\frac{(\hat{\theta}_j - \hat{\theta}_{IVW,k})^2}{2\mathrm{se}(\hat{\theta}_j)^2}\right]. \quad (8)$$

Aside from the constant term, this is a distance measure that will be greater when more variants are included in the subset k due to the $\mathrm{se}(\hat{\theta}_j)^{-1}$ terms, but they will reduce sharply if there is more heterogeneity between the variant-specific ratio estimates for variants in the subset than would be expected due to statistical uncertainty alone if all variants estimated the same causal parameter. If the variant-specific ratio estimates for variants in a particular subset substantially differ, then the weight for that subset will be low. Note that the reason for excluding subsets with one variant is that heterogeneity cannot be estimated for these subsets. We then normalize the weights so that they sum to 1:

$$w'_k = \frac{w_k}{\sum_k w_k}. \quad (9)$$

The causal estimate is the mode of the likelihood of the mixture of normal distributions using these weights:

$$\hat{\theta}_{MODE} = \arg\max_{\theta} \sum_{k} w'_k \mathrm{se}(\hat{\theta}_{IVWr,k})^{-1} \exp\left[-\frac{(\theta - \hat{\theta}_{IVW,k})^2}{2\mathrm{se}(\hat{\theta}_{IVWr,k})^2}\right]. \quad (10)$$

We use this likelihood for making inferences about the causal effect θ.

Consistency and efficiency

In the asymptotic limit for a fixed number of genetic variants but as the sample size tends to infinity (and hence the standard errors of the ratio estimates decrease to 0), the weighted mixture distribution (i.e. the likelihood for θ) tends to a series of spikes about the IVW estimates based on each subset of variants. The height of each spike depends on the total weight of variants that have that causal estimate, and the tallest spike is the estimate with the greatest weight of evidence. The modal estimate will be the IVW estimate corresponding to the subset k of variants all having the same ratio estimate which has the greatest product of the inverse standard errors of the ratio estimates $\prod_{j \in \sigma_k} \mathrm{se}(\hat{\theta}_j)^{-1}$. Therefore, a consistent estimate is obtained under a Hartwig's weighted ZEMPA assumption.[10] The intuition of this assumption is that a weighted plurality of the genetic variants is required to be valid instruments (as opposed to median-based methods that require a majority or weighted majority of variants to be valid instruments). The term 'plurality' is taken from the terminology of elections; a political party winning more votes than any other is said to have a plurality of the votes. We note the similarity between this procedure and maximum likelihood estimation, which gives the mode of a likelihood as its point estimate.

Under this assumption, the heterogeneity-penalized model-averaging method is asymptotically efficient, as the weight of the IVW estimate based on all the valid instruments will increase to 1 as the sample size tends to infinity. This can be seen as the weight for any subset containing variants with different ratio estimates will decrease to 0 rapidly. The weight of the largest subset of variants with the same ratio estimates will be the greatest of all subsets by the ZEMPA assumption, and the ratio of this weight to all other weights will increase to infinity as the sample size increases. However, asymptotic efficiency is not necessarily an important property in practice, as infinite sample sizes are rarely encountered in applied investigations. The model-averaging estimate should be efficient for finite sample sizes when several variants have similar ratio estimates.

Inferences on the weighted model-averaged distribution

We perform causal inferences based on the model-averaged distribution using a generalized likelihood ratio test to construct a confidence interval. We take twice the log-likelihood function, and construct a confidence interval consisting of all points for which twice their log-likelihood is within a given vertical distance from the modal estimate. For a 95% confidence interval, this distance is 3.841 (95th percentile of a chi-squared distribution with one degree of freedom). This is based on the result that twice the difference in the log-likelihood at the estimate and at the true value of the parameter has a chi-squared distribution (here with one degree of freedom as the parameter is one-dimensional). This results in inference without requiring resampling techniques (such as bootstrapping). The confidence interval is not guaranteed to be symmetrical or to be a single range of values (see later for an example of a bimodal mixture distribution resulting in a composite confidence interval).

Practically, the modal estimate and confidence interval were obtained using a grid search approach. The likelihood was evaluated at a series of points (in the simulation study, from –1 to +1 at intervals of 0.001—so estimates and confidence intervals were estimated to three decimal places). The modal estimate was taken as the point with the greatest value of the likelihood function, and the 95% confidence interval was taken as the set of points for which twice the log-likelihood was within 3.841 of the twice the log-likelihood at the modal estimate. If the log-likelihood function is multimodal, this may result in a composite confidence interval that consists of more than one range of values.

Simulation study

To consider the expected performance of this proposed method in realistic situations as well as in comparison to alternative robust methods, we perform a simulation study. We consider four scenarios:

1. no pleiotropy—all genetic variants are valid instruments;
2. balanced pleiotropy (violation of assumption IV3)—some genetic variants have direct (pleiotropic) effects on the outcome, and these pleiotropic effects are equally likely to be positive as negative;
3. directional pleiotropy (violation of IV3)—some genetic variants have direct (pleiotropic) effects on the outcome, and these pleiotropic effects are simulated to be positive;
4. pleiotropy via a confounder (violation of IV2)—some genetic variants have pleiotropic effects on the outcome

via a confounder. These pleiotropic effects are correlated with the instrument strength.

In the first three scenarios, the Instrument Strength Independent of Direct Effect (InSIDE) assumption[6] is satisfied; in Scenario 4, it is violated. This is the assumption required for the MR-Egger method to provide consistent estimates. This choice of scenarios enables us to explore cases where the consistency assumptions for the different methods are satisfied and violated to provide a fair comparison between different methods.

We simulate data for a risk factor X, outcome Y, confounder U (assumed unmeasured) and J genetic variants $G_j, j = 1, \ldots, J$. Individuals are indexed by i. The data-generating model for the simulation study is as follows:

$$U_i = \sum_{j=1}^{J} \zeta_j G_{ij} + \epsilon_{Ui} \tag{11}$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_{Xi}$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \theta X_i + U_i + \epsilon_{Yi}$$

$$G_{ij} \sim \text{Binomial}(2, 0.3) \text{ independently for all } j = 1, \ldots, J$$

$$\epsilon_{Ui}, \epsilon_{Xi}, \epsilon_{Yi} \sim \mathcal{N}(0, 1) \text{ independently}$$

$$\gamma_j \sim \text{Uniform}(0.03, 0.1) \text{ independently for all } j = 1, \ldots, J.$$

The risk factor and outcome are positively correlated due to confounding even when the causal effect θ is 0 through the unmeasured confounder U. The genetic variants are modelled as single-nucleotide polymorphisms (SNPs) with a minor allele frequency of 30%. A total of $J = 10$ genetic variants are used in each analysis. As the proposed model-averaging method calculates weights for all $2^J - J - 1$ possible models, the model scales exponentially with the number of variants, and so including more variants was not computationally feasible in a simulation setting. For each of Scenarios 2 to 4, we considered cases with two, three and five invalid instruments. For valid instruments, the α_j and ζ_j parameters were set to 0. For invalid instruments, the α_j parameters were either drawn from a uniform distribution on the interval from –0.1 to 0.1 (Scenario 2) or from 0 to 0.1 (Scenario 3) or set to 0 (Scenario 4). The ζ_j parameters were either set to 0 (Scenarios 2 and 3) or drawn from a uniform distribution on the interval from –0.1 to 0.1 (Scenario 4). The causal effect θ was either set to 0 (no causal effect) or 0.2 (positive causal effect). The γ_j parameters were drawn from a uniform distribution on 0.03 to 0.1, meaning that the average value of the R^2 statistic for the 10 variants across simulated datasets was 1.0% (from 1.1 to 1.4% in Scenario 4) corresponding to an average F statistic of 20.4 (from 23.4 to 27.5 in Scenario 4).

In total, 10 000 datasets were generated in each scenario. We considered a two-sample setting in which genetic associations with the risk factor and outcome were estimated on non-overlapping groups of 20 000 individuals. We compared estimates from the proposed heterogeneity-penalized model-averaging method with those from a variety of methods: the standard IVW method, MR-Egger[6] (both using random-effects), the weighted and simple median methods[8] and the mode-based estimate (MBE) of Hartwig *et al.*[10] Each of the methods was implemented using summarized data only.

Results

Results for all of the methods are provided in Tables 1 (Scenario 1) and 2 (Scenarios 2 to 4). We provide the mean estimate, the standard deviation of estimates, the mean standard error (Table 1 only) and the empirical power of the 95% confidence interval (the proportion of 95% confidence intervals excluding the null; this is the Type 1 error rate with a null causal effect). Results for the MBE method are only provided for 1000 simulated datasets per scenario. This is for computational reasons—the MBE method took around 20 times longer to run than all the other methods put together. Results for the MBE method correspond to simple (unweighted) and weighted versions of the method not assuming NOME (no measurement error) with the recommended bandwidth parameter from the modified Silverman rule ($\phi = 1$)[18]; in total, 12 different versions of the MBE method are proposed by Hartwig *et al.*

Table 1 shows the efficiency of the model-averaging method when all genetic variants are valid instruments. The method is considerably more efficient than the MR-Egger and MBE methods, with less variable estimates and greater power to detect a causal effect, and similar in efficiency to the median-based methods. Coverage under the null is conservative for all methods, but particularly for the MBE and model-averaging methods.

Table 2 shows the robustness of the model-averaging method in a range of invalid instrument scenarios. Type 1 error rates are well controlled (less than 7.5%) in all scenarios when 2 or 3 out of the 10 variants are invalid, and generally below those of other methods even when 5 variants are invalid. Compared with the model-averaging method, Type 1 error rates with five invalid instruments

Table 1. Mean, standard deviation (SD), mean standard error (mean SE) of estimates and empirical power (%) for Scenario 1 (all variants valid instruments)

Method	Scenario 1: all instruments valid			
	Mean	SD	Mean SE	Power
Null causal effect: $\theta = 0$				
Inverse-variance weighted	0.001	0.072	0.077	3.9
MR-Egger	0.003	0.223	0.236	3.6
Simple median	0.001	0.092	0.105	2.1
Weighted median	0.002	0.086	0.096	2.8
Simple mode-based estimate (Hartwig)	0.003	0.113	0.149	0.3
Weighted mode-based estimate (Hartwig)	0.002	0.098	0.128	1.2
Heterogeneity-penalized model averaging	0.001	0.080	–	1.4
Positive causal effect: $\theta = +0.2$				
Inverse-variance weighted	0.191	0.080	0.086	61.9
MR-Egger	0.130	0.250	0.263	7.0
Simple median	0.201	0.104	0.119	39.0
Weighted median	0.185	0.096	0.109	39.9
Simple mode-based estimate (Hartwig)	0.195	0.136	0.167	18.5
Weighted mode-based estimate (Hartwig)	0.172	0.115	0.142	22.4
Heterogeneity-penalized model averaging	0.188	0.090	–	38.8

for the MR-Egger method are lower in Scenario 3; however, they are far higher in Scenario 4, and the power of the MR-Egger method to detect a positive causal effect was low throughout. Equally, Type 1 error rates are slightly lower for the simple median method in Scenario 4, but higher in Scenario 3. The empirical power of the model-averaging method to detect a causal effect was generally lower than that for other methods. However, when a method suffers from Type 1 error inflation, this comparison is not a fair one. The power of the model-averaging method to detect a positive causal effect was not dominated by any method that had well-controlled Type 1 error rates. Indeed, in Scenario 2, the power of the model-averaging method even exceeded that of the IVW method with three and five invalid variants. This is because models including the invalid variants are down-weighted in the model-averaging method, whereas these variants inflate the standard error in the IVW method. Similar patterns were observed in the bias of estimates, with the model-averaging method generally having low bias. Although some methods were less biased in particular scenarios, no method was less biased across all scenarios.

In comparison to the MBE method of Hartwig *et al.*, Type 1 error rates for the model-averaging method were slightly higher than those for the simple MBE method, but lower than those for the weighted MBE method, particularly in Scenario 4, where the Type 1 error rate for the weighted MBE method was not well controlled even with only two invalid instruments. Power to detect a positive causal effect was greater for the model averaging than for the simple MBE method in all cases by at least 10%, and greater than for the weighted MBE method in all cases except in Scenario 4, where the weighted MBE method had inflated Type 1 error rates.

In an additional simulation, we considered the performance of the model-averaging method with six invalid instruments using the same sample size and a sample size of 100 000 (five times the original sample size) for each of the gene–risk factor and gene–outcome associations (Supplementary Table A1, available as Supplementary data at *IJE* online). Although all methods performed poorly with the original sample size, in comparison with the IVW and weighted median methods, for which bias was almost identical for the two sample sizes, bias for the model-averaging method reduced sharply as the sample size increased. In comparison with the MBE method, the model-averaging method performed similarly well with the original sample size, but the improvement in bias and Type 1 error rate with the increased sample size was much better for the model-averaging method, with little improvement in Type 1 error rates for the MBE method. In a further simulation, we considered the performance of the model-averaging method with four invalid instruments, but in which all the invalid instruments were simulated to have the same pleiotropic effect on the outcome (Supplementary Table A2, available as Supplementary data at *IJE* online). This resulted in a confidence interval that was not a single range of values for around 18% of simulated datasets with the majority of variants having a null causal effect. Despite this, the median estimate from the model-averaging

Table 2. Mean, standard deviation (SD) of estimates and empirical power (%) for Scenarios 2, 3 and 4. MBE, mode-based estimate of Hartwig *et al.*[10]

Method	Two invalid variants			Three invalid variants			Five invalid variants		
	Mean	SD	Power	Mean	SD	Power	Mean	SD	Power
Null causal effect: $\theta = 0$									
Scenario 2: Balanced pleiotropy, InSIDE satisfied									
Inverse-variance weighted	–0.001	0.140	6.3	0.002	0.163	7.5	0.000	0.202	7.8
MR-Egger	0.001	0.436	7.7	0.004	0.509	8.2	0.007	0.629	9.3
Simple median	0.000	0.113	3.8	0.002	0.129	5.5	0.000	0.175	10.2
Weighted median	0.001	0.109	5.2	0.001	0.125	7.5	0.000	0.178	15.0
Simple MBE	0.000	0.126	1.0	0.008	0.131	1.8	0.006	0.196	4.0
Weighted MBE	0.004	0.105	2.4	0.000	0.113	3.1	0.005	0.172	8.3
Model averaging	0.000	0.100	2.4	0.000	0.115	3.2	–0.001	0.187	6.0
Scenario 3: Directional pleiotropy, InSIDE satisfied									
Inverse-variance weighted	0.136	0.101	10.8	0.206	0.113	20.9	0.342	0.131	52.2
MR-Egger	0.004	0.421	7.8	0.002	0.479	8.2	0.011	0.539	8.5
Simple median	0.065	0.104	5.2	0.113	0.118	11.1	0.273	0.172	44.5
Weighted median	0.054	0.104	6.9	0.096	0.123	13.1	0.225	0.182	40.9
Simple MBE	0.020	0.122	1.7	0.044	0.138	2.3	0.146	0.220	9.4
Weighted MBE	0.013	0.102	2.9	0.041	0.123	5.1	0.114	0.177	12.8
Model averaging	0.021	0.098	2.6	0.043	0.121	3.9	0.133	0.214	11.8
Scenario 4: Pleiotropy via confounder, InSIDE violated									
Inverse-variance weighted	0.104	0.125	19.4	0.150	0.135	26.2	0.232	0.140	38.3
MR-Egger	0.240	0.433	35.9	0.304	0.440	39.0	0.401	0.411	40.7
Simple median	0.023	0.111	4.1	0.044	0.125	6.5	0.095	0.164	16.9
Weighted median	0.090	0.144	20.8	0.143	0.164	34.1	0.247	0.178	60.5
Simple MBE	0.018	0.133	2.6	0.043	0.155	4.5	0.091	0.194	12.5
Weighted MBE	0.072	0.171	16.4	0.128	0.197	28.2	0.216	0.204	47.6
Model averaging	0.023	0.118	4.3	0.050	0.146	7.4	0.139	0.206	22.1
Positive causal effect: $\theta = +0.2$									
Scenario 2: Balanced pleiotropy, InSIDE satisfied									
Inverse-variance weighted	0.193	0.143	33.3	0.188	0.168	26.5	0.195	0.206	19.5
MR-Egger	0.129	0.452	9.4	0.137	0.526	9.6	0.135	0.644	8.9
Simple median	0.204	0.127	34.6	0.200	0.143	33.2	0.206	0.191	33.0
Weighted median	0.186	0.122	36.4	0.186	0.140	36.2	0.190	0.188	37.0
Simple MBE	0.198	0.139	17.2	0.193	0.156	19.5	0.202	0.205	18.1
Weighted MBE	0.173	0.118	21.1	0.166	0.132	22.7	0.154	0.166	21.9
Model averaging	0.189	0.115	31.8	0.189	0.135	29.5	0.193	0.207	25.6
Scenario 3: Directional pleiotropy, InSIDE satisfied									
Inverse-variance weighted	0.329	0.110	72.7	0.397	0.121	79.8	0.532	0.140	92.1
MR-Egger	0.138	0.432	9.5	0.140	0.486	9.8	0.136	0.552	9.4
Simple median	0.274	0.120	55.0	0.328	0.136	65.7	0.489	0.186	87.2
Weighted median	0.247	0.117	55.3	0.292	0.137	65.0	0.419	0.189	82.6
Simple MBE	0.216	0.141	20.8	0.254	0.154	26.1	0.356	0.226	39.3
Weighted MBE	0.187	0.117	24.8	0.211	0.122	31.0	0.283	0.165	48.0
Model averaging	0.218	0.116	41.8	0.243	0.136	43.9	0.339	0.218	52.6
Scenario 4: Pleiotropy via confounder, InSIDE violated									
Inverse-variance weighted	0.298	0.131	63.5	0.343	0.140	66.6	0.426	0.146	74.4
MR-Egger	0.396	0.449	42.8	0.473	0.454	48.4	0.586	0.415	51.9
Simple median	0.232	0.125	42.7	0.252	0.139	45.7	0.304	0.176	53.2
Weighted median	0.285	0.156	62.1	0.338	0.175	71.5	0.444	0.184	85.4
Simple MBE	0.212	0.145	22.0	0.237	0.155	25.2	0.290	0.175	37.2
Weighted MBE	0.245	0.173	37.1	0.293	0.195	46.8	0.383	0.202	65.4
Model averaging	0.226	0.137	40.5	0.257	0.167	42.7	0.348	0.217	52.3

method was close to unbiased, and Type 1 error rates were at or below nominal levels.

Applied examples

We provide further illustration of the proposed model-averaging method and other robust methods in two applied examples. In the first example, all the variants have similar ratio estimates whereas, in the second example, there is marked heterogeneity in the variant-specific ratio estimates. Further detail about the applied examples is given in Supplementary Material A.5, available as Supplementary data at *IJE* online.

Low-density lipoprotein cholesterol and coronary artery disease (CAD) risk

We consider the causal relationship between low-density lipoprotein (LDL) cholesterol and CAD risk based on eight genetic variants having strong biological links with LDL-cholesterol. Each of these variants is located in a gene region that either encodes a biologically relevant compound to LDL-cholesterol or is a proxy for an existing or proposed LDL-cholesterol-lowering drug. Genetic associations with LDL-cholesterol were obtained from the Global Lipids Genetics Consortium's 2013 data release[19] and associations with CAD risk from CARDIoGRAMplusC4D's 2015 data release.[20] These associations are displayed graphically in Figure 1 (left panel). Weights for the variants and subsets of variants are displayed in Supplementary Figure A1, available as Supplementary data at *IJE* online.

Inflammation and CAD risk

We also consider the causal relationship between inflammation and CAD risk based on 17 genetic variants previously demonstrated to be associated with C-reactive protein (CRP) at a genome-wide level of statistical significance.[21] The biological rationale for this analysis is not to evaluate the causal role of CRP, as several of these genetic variants are not specifically associated with CRP and hence are not valid instruments as they violate the exclusion restriction assumption (they have an effect on the outcome not via CRP). The causal role of CRP can be evaluated in a Mendelian randomization analysis using genetic variants in the *CRP* gene region—the region that encodes CRP.[22] Rather, the biological rationale for this analysis considers CRP as a proxy measure for inflammation more generally and investigates whether there are any consistent causal relationships between inflammation and CAD risk. Genetic associations with CRP are obtained from Dehghan *et al.*[21] and associations with CAD risk from the CARDIoGRAM consortium.[23] These associations are displayed graphically in Figure 1 (right panel).

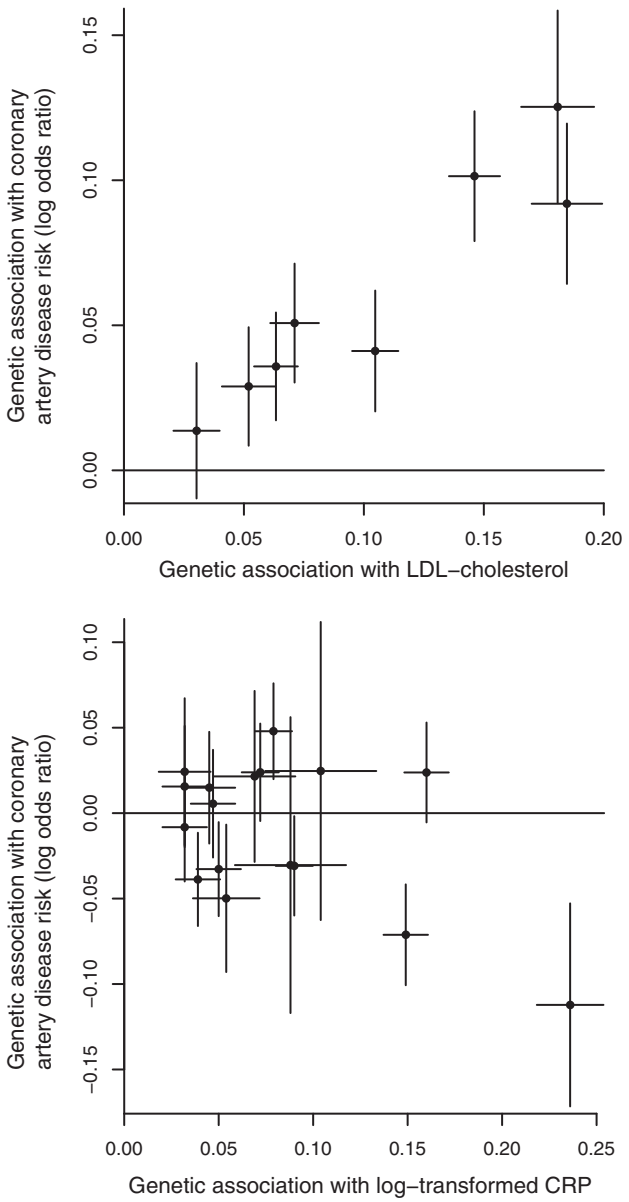

Figure 1. Genetic associations with risk factor and outcome (lines are 95% confidence intervals) for: (left) 8 genetic variants having biological links to LDL-cholesterol; (right) 17 genetic variants associated with C-reactive protein (CRP) at a genome-wide level of significance.

Results

Results for both examples are presented in Table 3. Estimates represent log odds ratios for CAD per 1-mmol/L increase in LDL-cholesterol or per unit increase in log-transformed CRP. For the MBE method, we present estimates for a range of values of the bandwidth in the kernel-density estimator representing the suggested bandwidth from the modified Silverman rule ($\phi = 1$), half the

Table 3. Estimates (standard errors, SE) and 95% confidence intervals (CI) from a variety of methods for applied examples. MBE, mode-based estimate of Hartwig *et al.*[10]

Risk factor:	LDL-cholesterol		C-reactive protein	
Method	Estimate (SE)	95% CI	Estimate (SE)	95% CI
Inverse-variance weighted	0.585 (0.044)	0.499, 0.671	–0.135 (0.102)	–0.334, 0.065
MR-Egger	0.611 (0.100)	0.415, 0.807	–0.223 (0.198)	–0.611, 0.165
Simple median	0.561 (0.067)	0.429, 0.693	0.118 (0.155)	–0.187, 0.422
Weighted median	0.585 (0.057)	0.473, 0.697	–0.303 (0.108)	–0.515, –0.092
Simple MBE ($\phi = 1$)	0.522 (0.105)	0.316, 0.727	0.295 (0.372)	–0.433, 1.023
Simple MBE ($\phi = 0.5$)	0.700 (0.136)	0.434, 0.966	0.285 (0.502)	–0.698, 1.269
Simple MBE ($\phi = 0.25$)	0.699 (0.147)	0.411, 0.987	0.306 (0.510)	–0.694, 1.305
Weighted MBE ($\phi = 1$)	0.686 (0.096)	0.498, 0.875	–0.407 (0.152)	–0.705, –0.108
Weighted MBE ($\phi = 0.5$)	0.697 (0.140)	0.423, 0.971	–0.458 (0.112)	–0.678, –0.238
Weighted MBE ($\phi = 0.25$)	0.696 (0.140)	0.421, 0.970	–0.472 (0.218)	–0.898, –0.045
Heterogeneity-penalized model averaging[a]	0.598	0.475, 0.718	–0.441	–0.602, –0.257 and 0.038, 0.352[b]

[a]The heterogeneity-penalized model-averaging method does not estimate a standard error. For the risk factor LDL-cholesterol, and assuming normality, the standard error would be 0.062.

[b]The confidence interval in this case is the union of two disjoint ranges.

suggested bandwidth ($\phi = 0.5$) and one-quarter of the suggested bandwidth ($\phi = 0.25$), as well as for simple and weighted versions of the method.

In the first example, all of the methods suggest a positive causal effect. In the model-averaging method, the weight of the estimate including all eight variants is 12.1% and estimates with seven or more variants comprise 42.1% of the total weight (compared with 0.4% and 3.6% of the weight with no heterogeneity penalization—equal weights). The width of the confidence interval from the model-averaging method is similar to that from the weighted median method, and narrower than that from all other methods except for the standard IVW method. Confidence intervals from the MBE method are considerably wider than those from other methods, and vary in size by up to 40% for the different choices of bandwidth considered here. The improvement in efficiency of our method compared with the best-case estimate from the MBE method is a 1.54-fold reduction in the standard error. Assuming that the standard error decreases proportionally as the square root of the sample size, this improvement would correspond to including an additional 98 000 cases and 154 000 controls in the analysis. In the second example, the methods give varied estimates. In particular, the simple MBE method gives a positive estimate, whereas the weighted MBE method gives a negative estimate with a confidence interval that excludes 0. In contrast, the model-averaging method gives a negative estimate, but a confidence interval that includes both negative and positive values, although excludes 0—it includes two disjoint ranges of values. Again, the precision of the MBE estimates varied for different choices of bandwidth, in the most extreme comparison by almost a factor of two.

Figure 2 shows the mixture distributions of the IVW estimates based on all subsets of genetics variants using both equal weights (dashed line) and heterogeneity-penalized weights (solid line) from the model-averaging method. For the first example, the equally and penalized weighted distributions are similar, as the IVW estimates based on all subsets of variants are similar. For the second example, the heterogeneity-penalized distribution differs substantially from distribution using equal weights and is bimodal, indicating that there are groups of variants having similar weight of evidence supporting both a positive and a negative causal effect, and suggesting that there are causal mechanisms linked with inflammation that have both protective and harmful effects on CAD risk. These results could be driven by different inflammatory risk factors that are causally upstream of CRP and have different directions of effect on the outcome. This explains the composite confidence interval including both positive and negative values. Only the model-averaging method is able to capture this feature of the data.

Discussion

The aim of this manuscript was to develop a mode-based estimation method that provides a consistent estimate of the causal effect under the assumption that a plurality of the genetic variants are valid instruments. Although our method is not the first to provide consistent estimates under this assumption, we believe that our method has several technical advantages over previously proposed methods. In comparison with the MBE method proposed by Hartwig *et al.*, our method: (i) does not rely on the specification of a bandwidth parameter; (ii) makes inferences that

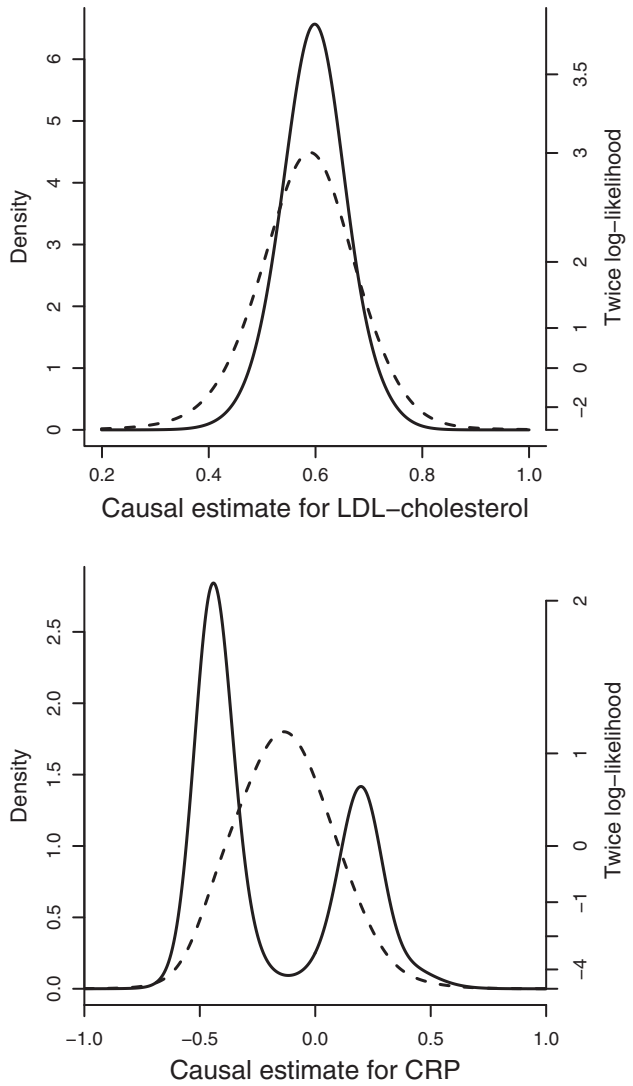

Figure 2. Mixture distributions of IVW estimates using equal (dashed line) and penalized (solid line) weights from model-averaging method for: (left) LDL-cholesterol; (right) C-reactive protein (CRP). The right-hand axis is twice the log-likelihood—the 95% confidence interval contains all points within a vertical distance of 3.84 units on this scale (3.84 is the 95th percentile of a chi-squared distribution on one degree of freedom).

do not rely on resampling methods; (iii) makes no asymptotic assumption about the distribution of the causal estimate for making inferences, in particular allowing confidence intervals to be asymmetric and to span multiple ranges; (iv) is asymptotically efficient, and should be efficient in finite samples, as the method seeks to up-weight the IVW estimate based on the largest number of variants with homogeneous ratio estimates. One particular concern with the MBE method is that the precision of the estimate is highly variable, depending on the choice of bandwidth parameter. There would be a great temptation as an applied researcher to perform the method for a variety of values of the bandwidth parameter and choose the bandwidth parameter corresponding to the most desirable estimate.

The proposed heterogeneity-penalized model-averaging method also outperformed Hartwig's method in the simulation study, and in the applied examples. No sizeable inflation in Type 1 error rates was observed across the simulation scenarios when 2 or 3 of the 10 genetic variants were invalid, and bias and Type 1 error rates were generally either better or no worse than for other robust methods. The method was also at least as efficient as other robust methods when all variants were valid instruments and had reasonable power to detect a causal effect throughout.

One deficiency of the proposed method is computational time. Whereas the method was substantially quicker than that of Hartwig *et al*. with 10 genetic variants, the run-time of our method doubles with each additional variant. In the applied example with 17 genetic variants, $2^{17} - 1 = 131\,071$ weights were calculated. The method calculated weights in 0.7 seconds on a single 2.60-GHz central processing unit (CPU). The grid search algorithm took a further 34 seconds. However, with 30 genetic variants, over 1 billion weights would need to be calculated. Reducing the computational burden may be possible—e.g. models including genetic variants with highly discrepant ratio estimates would receive low weights and could be dropped with little loss of accuracy. Alternatively, an algorithm such as shotgun stochastic search[24] may be able to explore the parameter space in an efficient way. However, solving this computational challenge in general is left as a problem for future work.

A particular novel feature of the method is its ability to identify multiple causal effects. Two categories of heterogeneity in the ratio estimates based on different variants can be conceived: 'random-effects' heterogeneity and heterogeneity from variants linked with different causal mechanisms. As in meta-analysis, it is likely that there will be some heterogeneity between ratio estimates from different variants arising due to slight differences in causal mechanisms, non-linearity of effects or non-homogeneity of effects across individuals. This is dealt with in the model-averaging model by allowing for over-dispersion in the standard errors from the IVW method. Another type of heterogeneity would occur if some genetic variants are invalid instruments and have incompatible ratio estimates; this is dealt with in the model-averaging model by upweighting evidence from the largest subset of variants with mutually compatible ratio estimates. An interesting case is if two or more sets of genetic variants have mutually similar but distinct ratio estimates (as in the example of CRP in the paper). This could occur for a complex risk

factor. For example, some genetic variants associated with body mass index (BMI) affect metabolism, whereas others may affect appetite. These two distinct biological processes may have different magnitudes of causal effect on the outcome. Future work would be beneficial to identify clusters of genetic variants having similar causal estimates that may reflect distinct causal mechanisms.

The heterogeneity-penalized model-averaging method is likely to be affected by weak instruments in a similar way to the IVW method, as it is based on a mixture of distributions centred on the IVW estimates. A weak instrument is one that does not have a statistically strong association with the risk factor.[25] When genetic associations with the risk factor and with the outcome are estimated in the same individuals (a one-sample investigation), the IVW estimate is biased by weak instruments in the direction of the observational association between the risk factor and outcome, and Type 1 error rates are inflated. However, if genetic variants are associated with the risk factor at a genome-wide level of significance, bias should be minimal.[26] When genetic associations with the risk factor and with the outcome are estimated in non-overlapping sets of individuals (a two-sample investigation), as is common in Mendelian randomization, bias due to weak instruments is in the direction of the null and does not lead to inflated Type 1 error rates.[27] Hence we would not expect weak instrument bias to adversely affect Mendelian randomization investigations using the model-averaging method in practice.

An extension of the method that could be valuable in applied practice is the use of prior information on particular variants. This can be achieved by multiplying the unnormalized weights w_k by a prior weight $\pi_0(k)$ before normalizing. For example, if an investigator is particularly confident that a genetic variant is likely to be a valid instrument, then models containing this variant can be up-weighted. Alternatively, prior weightings of models containing specific variants could be based on biological characteristics of the variants. For example, exonic and/or non-synonymous variants could be up-weighted or variants with functional information relating them to the risk factor. If these variants truly are more likely to be valid instruments, then this prior weighting would add to the robustness of the method. Additionally, a prior weighting could be set to more strongly up-weight less parsimonious models (i.e. up-weight models based on more genetic variants). This could add efficiency to the analysis, as models including more genetic variants will have more precise IVW estimates. Equal prior weights corresponds to a prior belief that 50% of genetic variants are valid instruments. If one instead believed that (say) 80% of genetic variants were valid instruments, then the prior for subset k could be set to $\pi_0(k) = 0.8^K \times 0.2^{J-K}$, where J is the total number of genetic variants and K is the number of variants in subset k. The option to set this prior probability is included in the software code.

Conclusion

In conclusion, the heterogeneity-penalized model-averaging procedure introduced in this paper will be a worthwhile contribution to the Mendelian randomization literature both in providing an additional robust method for causal estimation and testing the causal null hypothesis when some genetic variants may not be valid instruments and for revealing features in the data such as the presence of multiple causal mechanisms.

Supplementary Data

Supplementary data are available at *IJE* online.

Acknowledgements

This work was supported by the UK Medical Research Council (MC_UU_00002/7). S.B. and V.Z. are supported by Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 204623/Z/16/Z). A.G. is supported by a Medical Research Council Methodology Research Panel grant (Grant Number RG88311).

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. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Boca Raton, FL: Chapman & Hall, 2015.
3. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* 2007;**16**:309–30.
4. Wooldridge JM. *Introductory Econometrics: A Modern Approach. Chapter 15: Instrumental Variables Estimation and Two Stage Least Squares*. Nashville, TN: South-Western, 2009.
5. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med* 2016;**35**:1880–906.
6. 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.
7. Kang H, Zhang A, Cai T, Small D. Instrumental variables estimation with some invalid instruments, and its application to Mendelian randomisation. *J Am Stat Assoc* 2016;**111**:132–44.
8. 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.
9. Guo Z, Kang H, Cai TT, Small DS. Confidence intervals for causal effects with invalid instruments using two-stage hard thresholding with voting. *J Royal Stat Soc B* 2018; doi: 10.1111/rssb.12275.

10. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *Int J Epidemiol* 2017;**46**:1985–98.
11. Burgess S, Scott R, Timpson N, Davey Smith G, Thompson SG; EPIC-InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* 2015;**30**:543–52.
12. Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. *Stat Sci* 2011;**26**:403–22.
13. Burgess S; CHD CRP Genetics Collaboration. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Stat Med* 2013;**32**:4726–47.
14. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* 2017;**36**:1783–802.
15. Burgess S, Butterworth AS, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013;**37**:658–65.
16. Bowden J, Del Greco F, Minelli C *et al*. Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. *bioRxiv* 2017;159442.
17. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;**18**:2693–708.
18. 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* 2003;**73**:899–912.
19. The Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–83.
20. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;**47**:1121–30.
21. Dehghan A, Dupuis J, Barbalic M *et al*. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* 2011;**123**:731–38.
22. CRP CHD Genetics Collaboration. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ* 2011;**342**:d548.
23. Schunkert H, König I, Kathiresan S *et al*. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;**43**:333–38.
24. Hans C, Dobra A, West M. Shotgun stochastic search for 'large p' regression. *J Am Stat Assoc* 2007;**102**:507–16.
25. Burgess S, Thompson SG. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat Med* 2011;**30**:1312–23.
26. Burgess S, Thompson SG; CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol* 2011;**40**:755–64.
27. Pierce B, Burgess S. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *Am J Epidemiol* 2013;**178**:1177–84.