Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy

Chin Yang Shapland[1,2], Qingyuan Zhao[3], and Jack Bowden[4,1,2]

[1] MRC Integrative Epidemiology Unit at the University of Bristol, U.K.
[2] Population Health Sciences, University of Bristol, U.K.
[3] Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, U.K.
[4] College of Medicine and Health at the University of Exeter, U.K.
chinyang.shapland@bristol.ac.uk

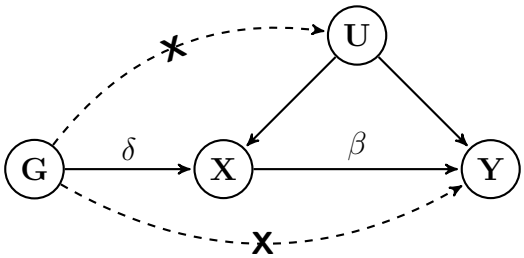
Abstract. Two-sample summary data Mendelian randomisation (MR) is a popular method for assessing causality in epidemiology, by using genetic variants as instrumental variables. If genes exert pleiotropic effects on the outcome not through the exposure of interest, this can lead to heterogeneous and (potentially) biased estimates of causal effect. We investigate the use of Bayesian model averaging (BMA) to preferentially search the space of models with the highest posterior likelihood. We develop a bespoke Metropolis-Hasting algorithm to perform the search using the recently developed Robust Adjusted Profile Likelihood (MR-RAPS) of Zhao et al as the basis for defining a posterior distribution that efficiently accounts for pleiotropic and weak instrument bias. We demonstrate how our general modelling approach can be extended from a standard one-parameter causal model to a two-parameter model, to allow a large proportion of SNPs to violate the Instrument Strength Independent of Direct Effect (InSIDE) assumption. We use Monte Carlo simulations to illustrate our methods and compare it to several related approaches. We finish by applying our approach in practice to investigate the changes in causal effect of their resulting high risk metabolite on the development age-related macular degeneration.

Keywords: Two-sample summary data Mendelian randomization · Bayesian Model Averaging, weak instruments, horizontal pleiotropy, InSIDE violation.

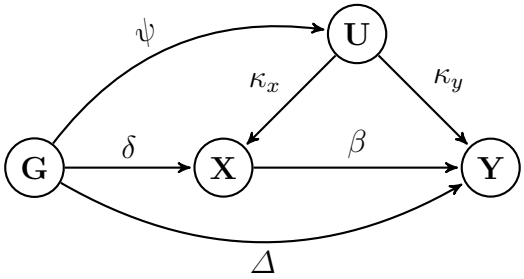
1 Introduction

The capacity of traditional observational epidemiology to reliably infer whether a health exposure causally influences a disease rests on its ability to appropriately measure and adjust for factors which jointly predict (or confound) the exposure-outcome relationship. Mendelian randomization (MR) [1] avoids bias

from unmeasured confounding by using genetic variants as instrumental variables (IVs) [2]. In order for the approach to be valid for testing causality, each specific IV must be robustly associated with the exposure (assumption IV1), independent of any confounders of the exposure and outcome (IV2) and be independent of the outcome given the exposure and the confounders (IV3). This is illustrated by the causal diagram in Figure 1a.



(a) General causal diagram



(b) Violations to IV assumptions

Fig. 1: *Causal diagrams representing the hypothesized relationship between genetic instrument (G), exposure (X), outcome (Y) and all unmeasured variables (U) which confound X and Y. β is the causal effect of X on Y.* (a) *δ is the genetic effect on X. Dashed lines and crosses indicate violations of the standard IV assumptions which can lead to bias.* (b) *Genetic instruments have a direct effect on Y (Δ), a phenomenon known as horizontal pleiotropy. Genetic instruments have a direct effect on U (ψ), an example of horizontal pleiotropy that violates the InSIDE assumption.*

The recent boom of genome-wide association studies (GWAS) [3] has triggered the development of MR approaches that utilise this widely available summary data source. Specifically, 'two-sample summary data' MR is a design that derives causal effect estimates with summary statistics obtained from two separate GWAS - one supplying the Single Nucleotide Polymorphism (SNP)-exposure as-

sociations and the other supplying the SNP-outcome associations [4–7] - a SNP being the most common type of genetic variation in the genome. If the chosen SNPs are valid IVs, and the causal effect of a unit increase in X on the mean value or risk of Y is approximately linear in the local region of X predicted by these variants [8] then a simple inverse-variance weighted (IVW) meta-analysis of SNP-specific causal estimates provides an approximately unbiased estimate of this causal effect. If sufficient heterogeneity exists between the MR estimates across a set of variants, this suggests that some of the SNPs may in fact violate the IV assumptions. This could be due to assumption IV1 being only weakly satisfied by the genetic variants (i.e. weak instrument bias), which can easily be accounted for [9, 8]. It is however more problematic when the heterogeneity is caused by violations of assumptions IV2 and IV3 [10, 11]. The latter violation is commonly known as "horizontal pleiotropy" [12], and hereafter referred to as pleiotropy for simplicity. Pleiotropy does not necessarily lead to biased causal estimation if it is balanced, in the sense that the average pleiotropic bias across SNPs is zero and the weight each SNP receives in the analysis is also independent of this bias. This latter condition is referred to as the Instrument Strength Independent of Direct Effect (InSIDE) assumption [13, 14]. However, this assumption is itself unverifiable.

Methods have been developed that are naturally robust to pleiotropy and InSIDE violation. For example, the weighted median estimator [15] provides a consistent estimate under the assumption that 50% of the SNPs are valid IVs (or not pleiotropic). Similarly, mode-based estimation strategies focus on identifying the largest subset of variants yielding a homogeneous causal estimate, and are consistent when this set is made up of valid IVs [16, 17]. These approaches do not make any assumptions the nature of the pleiotropy for invalid SNPs - it could violate InSIDE or not. Other approaches, such as MR-PRESSO [18] and Radial MR [9] attempt to detect and remove SNPs that are deemed responsible for bias and heterogeneity in an MR-analysis. They can in theory provide consistent estimates for the causal effect if the SNPs responsible for InSIDE violation can be removed from the analysis so that only balanced pleiotropy remains. Finally, the Robust Adjusted Profile Score (MR-RAPS) [8] takes a subtly different approach. It accounts for weak instruments and balanced pleiotropy using an adjusted profile likelihood, which penalizes outlying SNPs that may induce bias in the analysis using a robust loss function.

In this paper we develop a method for pleiotropy robust MR analysis with two-sample summary data using the general framework of Bayesian Model Averaging (BMA) [19]. BMA incorporates uncertainty about the effects of pleiotropy and weak instruments into MR estimate by pooling the causal effect estimates from all possible combinations of the genetic instruments with appropriate weights. In this paper, we adapt this general approach to the summary data setting where the SNPs are uncorrelated but potentially pleiotropic. Our approach uses the profile likelihood of MR-RAPS [8] as a basis for efficiently modelling the

summary data in the presence of weak instrument bias (IV1 violation) and pleiotropy (IV2-3 violation), but with the addition of an indicator function to denote whether an individual SNP is included or disregarded in the model. We develop a bespoke Metropolis-Hastings BMA algorithm to intelligently search the space models defined by all possible SNP subsets (i.e $\approx 2^L$ in the case of L SNPs) in order to decide which SNPs to include in the identified set of valid IVs within a given iteration of the markov chain. For this reason, we call our method BayEsian Set IDE Mendelian randomization (BESIDE-MR). BESIDE-MR naturally up-weights large sets of variants that furnish consistent, homogeneous estimates of causal effect, and down-weights sets of variants that provide heterogeneous estimates of causal effect. It also naturally accounts for uncertainty introduced by SNP selection across models, which we will show is important for preserving the coverage of resulting MR estimates.

In Section 2.1 and 2.2 we introduce the methodology behind our basic approach and in Section 3 assess its performance in Monte-Carlo simulations. In Section 4 we show how the basic one-parameter causal model can be extended to account for the case where a substantial proportion of the SNPs exhibit pleiotropy violating the InSIDE assumption. In Section 5 we apply our approach to investigate the causal role of high density lipoprotein cholestorol (XL.HDL.C) on the risk of age related macular degeneration (AMD) using data from the 2019 MR Data Challenge [20]. We conclude with a discussion and point to further research.

2 Motivation and Method

2.1 Description of the general model

Suppose that we have data from an MR study consisting of N individuals, where for each subject k we measure L independent genetic variants $(G_{k1} \dots G_{kL})$, an exposure (X_k) and an outcome (Y_k). U_k represents the shared residual error between X and Y due to confounding, which we wish to overcome using IV methods. We assume the following linear structural models [21] for U, X and Y consistent with Figure 1b:

$$U_k|G_k = \sum_{j=1}^{L} \psi_j G_{kj} + \epsilon_k^U,$$

$$X_k|U_k, G_k = \sum_{j=1}^{L} \delta_j G_{kj} + \kappa_x U_k + \epsilon_k^X,$$

$$Y_k|X_k, U_k, G_k = \sum_{j=1}^{L} \Delta_j G_{kj} + \beta X_k + \kappa_y U_k + \epsilon_k^Y,$$

where ϵ_k^U, ϵ_k^X and ϵ_k^Y are independent error terms for U, X and Y respectively. From this we can derive the approximate reduced form models for the G-X and

G-Y associations for SNP j :

$$X_k|G_{kj} \approx (\delta_j + \kappa_x\psi_j)G_{kj} + \epsilon_k'^X, \tag{1}$$

$$Y_k|G_{kj} \approx \Big[\Delta_j + \kappa_y\psi_j + \beta(\delta_j + \kappa_x\psi_j)\Big]G_{kj} + \epsilon_k'^Y. \tag{2}$$

We use 'approximate' here because the error terms $\epsilon_k'^X$ and $\epsilon_k'^Y$ not strictly constant or mutually independent - the jth residual error term in fact contains common contributions from all other genetic variants not equal to j. This approximation is very accurate in most settings because the genetic variants combined make a very small contribution to the total residual error in each model (e.g. typically of the order of 1-2%). For further justification see Zhao *et al.* [8]. Under this assumption the following models can then be justified for summary data estimates of the G-X and G-Y associations gleaned from fitting (1) and (2):

$$\hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{Xj}), \quad \hat{\Gamma}_j|\alpha_j, \gamma_j \sim N(\alpha_j + \beta\gamma_j, \sigma^2_{Yj}), \tag{3}$$

Here, $\alpha_j = \Delta_j + \kappa_y\psi_j$, and $\gamma_j = \delta_j + \kappa_x\psi_j$. Model (3) is typically applied in the two-sample summary data setting. Under this design it is assumed that the first study provides the G-X associations $\hat{\gamma}_j$ and standard errors σ_{Xj}, and the second study provides the G-Y associations $\hat{\Gamma}_j$ and standard errors σ_{Yj}. Both the standard errors are assumed to be fixed and known. The two-sample design implicitly assumes that SNP j has an identical association with the outcome in studies 1 and 2. Since the two studies are independent, it is also assumed that the uncertainty in $\hat{\gamma}_j$ is independent of the uncertainty in $\hat{\Gamma}_j$. For a detailed description of all the assumptions underlying the two-sample approach see Bowden *et al.* [11] and Zhao *et al.* [22].

The individual Wald ratio estimand for SNP j from model 3 is then

$$\beta_j = \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j} = \beta + \frac{\Delta_j + \kappa_y\psi_j}{\delta_j + \kappa_x\psi_j}$$

From this we see that:

- A SNP is invalid due to pleiotropy if $\alpha_j \neq 0$
- A SNP is invalid due to InSIDE respecting pleiotropy if $\alpha_j \neq 0$ but $\psi_j = 0$
- A SNP is invalid due to InSIDE violating pleiotropy if $\alpha_j \neq 0$ and $\psi_j \neq 0$.

InSIDE violation occurs in the last case because instrument strength and pleiotropic effects are functionally related due to a shared ψ_j component, so that the sample covariance $\widehat{Cov}(\alpha_j, \gamma_j) \neq 0$. For the case of InSIDE respecting pleiotropy we are able to assume the sample covariance is approximately zero for a sufficient number of instruments, since Δ_j and δ_j are imagined to be themselves generated via independent processes [11]. In Appendix 1, we show, under the simplifying assumption that the SNP-outcome standard errors are approximately

constant, that IVW estimator for the causal effect

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{L} \hat{\Gamma}_j \hat{\gamma}_j}{\sum_{j=1}^{L} \hat{\gamma}_j^2} \rightarrow \beta + \underbrace{\frac{\widehat{Cov}(\alpha_j, \gamma_j) + \bar{\alpha}\bar{\gamma}}{\widehat{Var}(\gamma_j) + \bar{\gamma}^2}}_{\text{bias term}} \tag{4}$$

as the sample size grows large. If all SNPs are pleiotropic, but satisfy the InSIDE assumption ($\widehat{Cov}(\alpha_j, \gamma_j) = 0$) and have mean of zero ($\bar{\alpha}$=0), then numerator of the bias term is zero and the standard IVW estimate provides a reliable way of estimating β. MR-Egger regression is an extension of the method that can work under the InSIDE assumption even if $\bar{\alpha} \neq 0$, which is referred to as 'directional' pleiotropy. It does this by estimating an intercept parameter in addition to the causal slope parameter. However, this approach has several downsides; its estimates are generally very imprecise and it is not invariant to allele recoding [23]. Lastly, it can not separate non-zero mean pleiotropy satisfying the InSIDE assumption from zero mean pleiotropy violating the InSIDE assumption. Its intercept reflects the numerator of the bias term, which is a combination of both. This motivates the use of methods that can attempt to detect and down-weight a small number of variants that may be responsible for either InSIDE violation or directional pleiotropy so that, for the remainder of SNPs left, model (3) holds with only InSIDE respecting balanced pleiotropy remaining. This is the approach taken by Zhao *et al.* [8] and Verbanck *et al.* [18], and is the approach we will initially pursue using BESIDE-MR.

2.2 Bayesian Model Averaging over the summary data model

We are interested searching over the space of all possible models defined by each of the 2^L subsets in the entire summary data. Let $I = (I_1, \ldots, I_L)$ be the L-length indicator vector denoting whether SNP G_j is included ($I_j = 1$) or not ($I_j = 0$) in the model. The model we want to 'force' our data to conform to is model (3) with the additional assumption that

$$\alpha_j \sim N(0, \tau^2)$$

The parameters of interest are then $\theta = (\beta, \tau^2, I)$ and with data, D, that consists of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$, and their standard errors σ_{Xj} and σ_{Yj} respectively. Then the joint posterior is

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

where $P(D|\theta)$ is the likelihood and $P(\theta)$ is user specified prior for each of the parameters.

As we rarely know the identity of the pleiotropic instruments, BMA offers the uncertainty about their identity and gives the inclusion posterior probability to quantify how likely an instrument is invalid. BMA achieves this by considering all possible models, where the models are different combinations of the potential

set of genetic instruments and averages the resulting causal effect estimate with appropriate weights (in the form of posterior probability of how often the instrument is included during MCMC). The selection of instruments is conditional on the likelihood of the data and the given priors. The prior in Bayesian analysis reduces the effect of weak instrument bias, even with noninformative prior. However, we must advice against this, as reasons discussed by Thompson *et al.* [24]. This method is particularly attractive as it has also been found to reduce bias from many weak instruments in econometric [25, 26] and in one-sample MR with highly valid but highly correlated genetic instruments [27]. For a comprehensive tutorial of BMA see Hoeting *et al.* [19].

2.3 The profile score likelihood

For $P(D|\theta)$, we use the profile log-likelihood score derived by Zhao *et al.* [8]. Specifically this is the likelihood for (β, τ^2) given the data $(\hat{\gamma}, \hat{\Gamma})$ profiled over the parameters $\gamma_1, ..., \gamma_L$. After the incorporation of our indicator vector I, the likelihood is modified to

$$l(\beta, \tau^2, I|\hat{\gamma}, \hat{\Gamma}) = -\frac{\sum_{j=1}^{L} I_j}{2} log(2\pi) - \frac{1}{2}\sum_{j=1}^{L} I_j \left\{ log(\sigma_{Yj}^2 + \tau^2) + \left(\frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\beta^2\sigma_{Xj}^2 + \sigma_{Yj}^2 + \tau^2} \right) \right\} \quad (5)$$

This likelihood allows for heterogeneity due to pleiotropy via τ^2, and weak instruments, via σ_{Xj}^2. Failure to account for weak instrument bias can lead to bias in the standard IVW estimate and inflation in related heterogeneity tests even under balanced pleiotropy [23]. Note that this likelihood is an increasing function of number of instruments included and decreasing function of τ^2. Hence, our BMA algorithm will naturally give more weight to I-vectors that include large sets of instruments with homogeneous causal effect estimates. This property is reminiscent of ZEro Modal Pleiotropy Assumption (ZEMPA) [16] or plurality rule that defines the two-stage hard thresholding (TSHT) approach of Guo *et al.* [28]. However, there is an important distinction. The TSHT approach explicitly aims to isolate the largest set of 'valid' instruments and base all inference on this single set, which is equivalent to giving a single I-vector a weight of 1 and all other vectors a weight of zero. Our approach is less aggressive, allowing as many distinct I-vectors as are supported by the data to be given weight in the analysis. This feature properly accounts for model uncertainty. Indeed, as subsequent simulations will demonstrate, this yields causal estimates and standard errors that are less prone to under-coverage than methods which incorporate instrument selection or penalization.

One such method of penalization, also proposed by Zhao *et al.* [8], is MR-RAPS. Instead of being based on likelihood function (5) which uses standard least squares or L_2 loss plus the addition of our indicator function, it uses a

robust L_1 function such as Huber or Tukey loss. They enable the contribution of large outliers to be penalized (i.e. reduced) compared to L_2 loss. Our BMA indicator function implemention of the standard profile likelihood can be viewed as an alternative way to achieving the robustness of MR-RAPS, because it will give more weight to model choices in which outliers are removed.

The profile likelihood is particularly well suited to a Bayesian implementation because it enables heterogeneity due to weak instrument bias and pleiotropy to be accounted for, whilst only having to update three parameters, (β, τ and I). Weak instrument bias is traditionally addressed using standard (non-profile) likelihood formulae (e.g. see Thompson *et al.* [24]), but this would require the posterior distribution of an additional L parameters $(\gamma_1, \ldots, \gamma_L)$ to be estimated, and is far more computationally intensive.

2.4 Choice of priors

In general we encourage the construction of priors to be based on previous epidemiological study or biological knowledge. For the purpose of elucidating our approach, we will use priors that ensure efficient mixing and rapid convergence. For the causal effect parameter β, we use a zero centered normal prior $P(\beta)$. For the pleiotropy variance we use a gamma prior $P(Prec)$ for the precision, where $Prec = 1/\tau^2$. For the indicator function prior, we will assume an uninformative Bernoulli prior $P(I)$ with probability $\frac{1}{2}$ for all I_j.

2.5 Metropolis-Hastings algorithm

We use a random walk Metropolis-Hastings (M-H) algorithm for updating the model parameter values. Unlike Gibbs sampling, the M-H algorithm does not directly sample from the conditional posterior distribution, but instead requires a proposal distribution for each parameter. Let $\theta_i = (\beta_i, \tau_i^2, I_i)$ be the current ith value of the parameter vector θ. θ_i is updated to θ_{i+1} one parameter at a time, by simulating a candidate value θ^* from proposal density, until it is accepted. Note that if the proposal density $C()$ for a given parameter is 'symmetric' - that is if $C(\theta_i|\theta_{i+1}) = C(\theta_{i+1}|\theta_i)$ then the proposal density can be omitted from the calculation of the acceptance probability. This is the case for β and I, but not τ^2. In Appendix 2, we give the specific details of the M-H algorithm.

2.6 An alternative implementation

It is well known that the estimation of τ^2 is challenging, even within a classical framework, see Zhao *et al.* [8] for further discussion. Therefore, we propose an alternative implementation of our M-H algorithm in which a plug-in estimate for

τ^2 is substituted at each iteration. For simplicity, we chose to use the closed-form DerSimonian-Laird estimate for τ^2 [29];

$$\hat{\tau}^2 = \max(0, (Q - (\sum_{j=1}^{L} I_j - 1))/W) \tag{6}$$

where

$$Q = \sum_{j=1}^{L} I_j w_j (\hat{\beta}_j - \beta_{IVW})^2, \quad \beta_{IVW} = \frac{\sum_{j=1}^{L} I_j w_j \hat{\beta}_j}{\sum_{j=1}^{L} I_j w_j}, \quad W = \sum_{j=1}^{L} I_j w_j - \frac{\sum_{j=1}^{L} I_j w_j^2}{\sum_{j=1}^{L} I_j w_j}$$

and $w_j = 1/Var(\hat{\beta}_j)$ respectively. Note that I_j should not be confused with Higgin's I^2 statistic used to quantify heterogeneity in Meta-analysis. In Appendix 2, we describe how the M-H algorithm is modified to implement this alternative approach. Hereafter, we will refer to the first method as the 'full Bayesian' approach and this latter method as the DerSimonian-Laird (DL) approach.

3 Monte Carlo simulation

3.1 Simulation strategy

In order to assess the performance of our BMA algorithm we simulate two-sample summary MR data sets with L=50 instruments from model (3). The parameters γ_j were generated from a Uniform $U(0.34, 1.1)$ distribution, σ_{Xj} was generated from a Uniform $U(0.06, UB)$ and σ_{Yj} was generated from a Uniform $U(0.015, 0.11)$ distribution. The upper bound on the G-X association standard error UB was used to determine mean instrument strength - with $0.095 \leq UB \leq 1$ giving mean F-statistics between 10 and 100 respectively. In this setting, the F-statistic for a single SNP can be approximated as $\hat{\gamma}_j^2/\sigma_{Xj}^2$.

We generated pleiotropic effects from a $N(\mu_\alpha, \tau^2)$ distribution, with the parameter μ_α being used to determine the mean bias induced by including the invalid instruments in the model. The task of our BMA algorithm in the presence of a non-zero μ_α is to give large weight to models which include SNPs for which $\mu_\alpha \approx 0$. Apart from a potential non-zero mean bias, the simulated pleiotropic effects satisfy the InSIDE assumption.

Convergence We test the convergence of our algorithm in different scenarios. From this we determined that it was necessary to use 50,000 iterations with 10,000 burn-ins for our algorithm to function effectively. Except for rare occasions, we removed results from data where number of iterations chosen was not sufficient for convergence. See details in Appendix 3 of the SM available at *Biostatistics* online.

A comparison with IVW, MR-APS and MR-RAPS We first compare our approach with the standard IVW method, MR-APS and MR-RAPS. The latter two are the classical counterpart that our approach sits between. We monitor the following quantities across our simulations:

- Mean bias of the causal parameter estimate. For our BMA algorithm we use the mean of the posterior distribution of β to assess this;
- Coverage: For IVW, MR-APS and MR-RAPS this is based on 95% symmetric confidence intervals assuming normality. For BESIDE-MR this is based on a 95% credibility interval;
- The inclusion probability for each SNP (BESIDE-MR only).

Four scenarios are considered; (1) balanced pleiotropy with strong instruments, (2) balanced pleiotropy with weak instruments, (3) directional pleiotropy with strong instruments and (4) directional pleiotropy with weak instruments (see Table 1 for a summary). Within each scenario, four sub-scenarios are considered where 0% to 100% of the SNPs are simulated as invalid/pleiotropic instruments. In order capture the amount of heterogeneity present in the data, we report the exact Q-statistic [9]:

$$Q = \sum_{j=1}^{L} w_j(\beta)(\hat{\beta}_j - \beta)^2 \tag{7}$$

Note that only invalid SNPs which have a non-zero pleiotropic effect make a non-nominal contribution, so that, for a fixed set of pleiotropy parameters $\alpha_1, \ldots, \alpha_L$:

$$E[Q] = \sum_{\alpha_j \neq 0} \frac{\alpha_j^2}{\beta^2 \sigma_{Xj}^2 + \sigma_{Yj}^2} + (L-1) \tag{8}$$

Table 2 shows the results.

Table 1: *Summary of simulation scenarios*

Scenario	Type of pleiotropy	$\bar{F}$	pleiotropic effect of invalid instruments
1	Balanced	100	$N(0, 0.04)$
2	Balanced	10	$N(0, 0.04)$
3	Directional	100	$N(0.05, 0.04)$
4	Directional	10	$N(0.05, 0.04)$

3.2 Results

bias and coverage Under scenario 1, all methods deliver approximately unbiased estimates. The IVW, MR-APS and MR-RAPS estimators achieve nominal

coverage when there are no pleiotropic instruments. However, as the proportion of pleiotropic instruments (and hence the heterogeneity) increases, their coverages can drop substantially, with the MR-APS and MR-RAPS estimators most affected. BESIDE-MR approach has conservative coverage under no heterogeneity but maintains far better coverage when this increases. Scenario 2 is the same as Scenario 1, except the SNPs are now weaker. The general pattern is the same, except the coverage of IVW is lower and its estimate is negatively biased. This is as expected because it uses inverse variance weights that assume σ^2_{Xj}=0 for all SNPs [9]. The results remains the same with further simulation of many weak instruments (L=100), see Appendix 3.

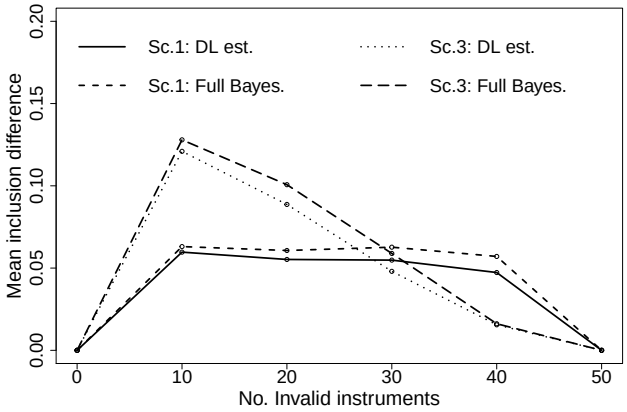
In scenarios 3 and 4, all the approaches deteriorate with increasing number of invalid instruments, but BMA has consistently the least bias and best coverage throughout. In Scenario 4, the IVW estimator is least biased, due to weak instrument bias cancelling out some of the pleiotropic bias. With 40% and 60% invalid instruments, full Bayesian BESIDE-MR struggled to converge within 50,000 iterations in a small number of cases.

SNP inclusion Figure 2a shows the difference between the mean probability of including valid (non-pleiotropic) SNPs and the mean probability of including invalid (pleiotropic) SNPs in our BMA likelihood for Scenarios 1 and 3. This difference is zero when there are no invalid instruments. Under Scenario 1 this difference is maximised (i.e. we get the best discrimination) when there are 20% invalid instruments, this difference steadily decreases to half its value as the number of invalid instruments increases further. Under Scenario 3 we see a smaller and more constant difference across different proportions of invalid instruments, indicating that the BMA likelihood generally struggles to deal with directional pleiotropy. There is still difference in inclusion posterior probability between valid and invalid instruments, however not in the same magnitude for weaker and many weak instruments, see Appendix 3.

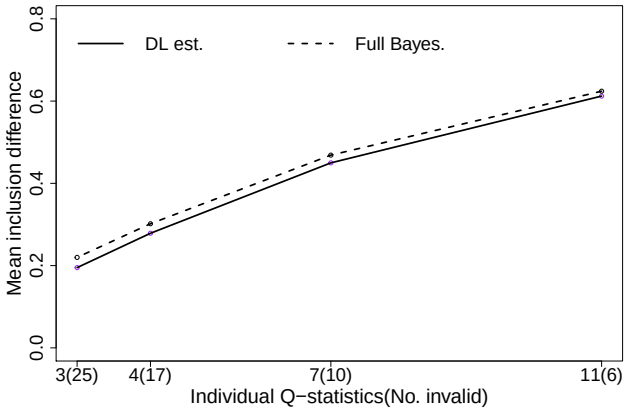
Additional simulations were performed to investigate the effect of different patterns of heterogeneity, but a fixed and borderline amount total heterogeneity, on the difference in inclusion probabilities of valid and invalid SNPs. Specifically, we generate summary data under balanced pleiotropy with Q fixed at 66, but varying the source of of this heterogeneity from 25 weakly pleiotropic SNPs (with individual Q contributions $\approx$ 3), to 11 strongly pleiotropic SNPs (with individual Q contributions $\approx$ 6). Figure 2b shows the results. As expected, the discrimination is best with small numbers of highly pleiotropic SNPs, and the worst with large numbers of weakly pleiotropic SNPs. However, the algorithm maintains its reliability in even in this case. For further results see Appendix 3.

Table 2: *Evaluation criteria for different types of pleiotropy and instrument strength (Table 1). 50 instruments in total. True β is 0.05. No. inv., Number of invalid instrument(s); Q, Q-statistics with exact weights; DL est., DL estimate; Full Bayes., Full Bayesian; bias, mean bias*

No. inv.	***Q***	**IVW**		**DL est.**		**Full Bayes.**		**MR-APS**		**MR-RAPS**	
		Scenario 1									
bias(i); coverage(ii)		i	ii	i	ii	i	ii	i	ii	i	ii
0	49.0	-0.001	96.40	-0.000	97.50	0.000	98.10	-0.000	94.40	-0.000	94.00
10	57.9	-0.001	93.20	0.000	97.50	0.000	97.70	-0.000	89.50	-0.000	92.10
20	66.4	-0.001	90.80	-0.000	95.40	-0.000	94.60	-0.000	83.90	-0.000	87.30
30	75.5	-0.000	88.30	0.001	94.20	0.001	92.00	0.001	77.30	0.001	80.80
40	84.0	-0.001	86.80	-0.000	95.80	-0.000	90.70	0.001	76.60	0.001	77.60
50	91.9	0.000	85.40	0.000	94.80	0.001	86.60	0.002	70.40	0.001	72.90
		Scenario 2									
0	48.7	-0.018	33.40	-0.001	97.10	0.002	96.10	-0.000	93.90	-0.000	92.90
10	54.4	-0.019	37.50	-0.000	97.10	0.005	93.70	0.003	91.80	0.003	92.10
20	59.2	-0.018	41.70	0.001	96.70	0.008	90.50	0.006	88.00	0.006	89.10
30	64.0	-0.018	44.60	0.001	96.70	0.011	87.80	0.009	83.20	0.008	84.90
40	68.8	-0.018	46.50	0.001	95.60	0.014	80.20	0.012	72.50	0.011	75.70
50	73.9	-0.019	47.80	0.002	94.60	0.017	73.40	0.015	68.80	0.015	70.10
		Scenario 3									
		i	ii	i	ii	i	ii	i	ii	i	ii
0	49.0	-0.001	96.40	-0.000	97.50	0.000	98.10	-0.000	94.40	-0.000	94.00
10	69.0	0.011	75.60	0.007	92.80	0.007	92.70	0.013	61.30	0.009	75.80
20	84.1	0.024	35.20	0.018	71.90	0.016	70.00	0.027	20.20	0.021	33.60
30	92.0	0.037	11.80	0.032	38.20	0.031	36.10	0.039	4.70	0.035	7.90
40	96.1	0.051	1.40	0.049	9.30	0.049	9.70	0.054	0.10	0.052	0.40
50	95.2	0.064	0.30	0.066	1.50	0.067	1.50	0.068	0.00	0.067	0.00
		Scenario 4									
		i	ii	i	ii	i	ii	i	ii	i	ii
0	48.7	-0.018	33.40	-0.001	97.10	0.002	96.10	-0.000	93.90	-0.000	92.90
10	58.8	-0.011	69.77	0.007	95.60	0.015	79.00	0.018	66.30	0.016	71.70
20	64.5	-0.003	84.70	0.017	84.60	0.028	46.20	0.035	23.70	0.034	29.60
30	66.5	0.006	82.60	0.028	64.60	0.040	21.70	0.050	5.10	0.048	7.00
40	66.2	0.014	70.10	0.040	35.60	0.049	9.90	0.064	0.40	0.063	0.60
50	65.3	0.022	53.90	0.050	18.90	0.057	5.20	0.075	0.10	0.074	0.10



(a) Scenario 1 and 3



(b) Individual Q

Fig. 2: (a) *The difference in mean inclusion probability between valid and invalid instruments for balanced and directional pleiotropy (scenario 1 and 3 respectively).* (b) *The difference in mean inclusion probability when a fixed amount of heterogeneity (Q=66) is due to many weakly pleiotropic or a small number of highly pleiotropic SNPs.*

4 An extended two parameter BMA model for substantial InSIDE violation

In Section 2.2 we introduced the BMA framework for summary data MR and applied it directly using the MR-APS likelihood of [8]. This method assumed that most SNPs were valid, but a small proportion could be invalid and directionally pleiotropic under the InSIDE assumption. The simulations in Section 3 showed that it performed well when this was the case, but like all other approaches, it suffered when a large number of SNPs were directionally pleiotropic, thus inducing both heterogeneity and bias into the results. We now consider the use of an extended BMA model to account for the extreme case where large numbers of SNPs may be invalid and in addition, violate the InSIDE assumption (Figure 1b).

Using the same underlying data generating model 3, suppose that we have two different groups of invalid instruments: in the first group, S_1 we have $\psi_j = 0$ for all SNPs and $\bar{\Delta} = \bar{\alpha} = 0$, shown in Section 3.1. That is, the SNPs in S_1 exhibit balanced pleiotropy under the InSIDE assumption. For illustrative purposes, suppose now that the remaining instruments are in a set S_2, defined by $\delta_j = 0$, $\Delta_j = 0$ and $\kappa_x = \kappa_y = 1$, but ψ_j have Uniform $U(0.34, 1.1)$ distribution. This means that that $\alpha_j = \gamma_j = \psi_j$, so that the InSIDE assumption is perfectly violated. Using the bias formulae in equation (4), it follows that

$$\text{For } j \in S_1 : \hat{\Gamma}_j = \alpha_j + \beta\gamma_j + \sigma_{Yj}\epsilon_j$$
$$\text{For } j \in S_2 : \hat{\Gamma}_j = \alpha_j + \beta^*\gamma_j + \sigma_{Yj}\epsilon_j$$

where $\beta^* = \beta + 1$. The set of SNPs in S_2 therefore identify a distinct, biased version of the causal effect. In the general case where the SNPs could be classified into an InSIDE-respecting set and an InSIDE-violating set, it would be more reasonable to assume that α_j, γ_j and Δ_j could all be non-zero. Although InSIDE would not then be maximally violated in S_2 we would still see two clusters in the data, albeit with a less defined separation. This motivates the development of an extended two-parameter version of our BMA algorithm to look for evidence of two clusters or slopes in the data.

4.1 A modified BMA algorithm

Under the generating model in (3), we further assume that the pleiotropic effects for valid SNPs in S_1 are generated from a $N(0, \tau_1^2)$ distribution and the set of invalid SNPs in S_2 are generated from a $N(0, \tau_2^2)$ distribution. Allowing these SNPs to violate InSIDE, and therefore identify a different slope parameter, our

total parameter space is modified to $\theta = (\beta_1, \tau_1^2, \beta_2, \tau_2^2, I_1, I_2)$, with likelihood:

$$\begin{aligned} l(\theta|\hat{\gamma}, \hat{\Gamma}) =& \text{Max}_{\boldsymbol{\gamma}} l(\beta_1, \tau_1^2, \beta_2, \tau_2^2 | \hat{\gamma}, \hat{\Gamma}) \\ =& \log f(\hat{\gamma}, \hat{\Gamma} | \beta_1, \tau_1^2, \beta_2, \tau_2^2) \\ =& -\frac{\sum_{j=1}^{L} I_{1j}}{2} log(2\pi) \\ & - \frac{1}{2} \sum_{j=1}^{L} I_{1j} \left\{ log(\sigma_{Yj}^2 + \tau_1^2) + \left(\frac{(\hat{\Gamma}_j - \beta_1 \hat{\gamma}_j)^2}{\beta_1^2 \sigma_{Xj}^2 + \sigma_{Yj}^2 + \tau_1^2} \right) \right\} \\ & - \frac{\sum_{j=1}^{L} I_{2j}}{2} log(2\pi) \\ & - \frac{1}{2} \sum_{j=1}^{L} I_{2j} \left\{ log(\sigma_{Yj}^2 + \tau_2^2) + \left(\frac{(\hat{\Gamma}_j - \beta_2 \hat{\gamma}_j)^2}{\beta_2^2 \sigma_{Xj}^2 + \sigma_{Yj}^2 + \tau_2^2} \right) \right\} \end{aligned} \tag{9}$$

Where the indicator functions I_{1j} and I_{2j} denote whether a SNP j is included in S_1 or S_2. We impose the condition that $I_{1j} + I_{2j} \leq 1$, which means that, at a given iteration of our BMA algorithm a SNP is either

- In S_1: $(I_{1j} = 1, I_{2j} = 0)$
- In S_2: $(I_{1j} = 0, I_{2j} = 1)$
- In neither S_1 or S_2 : $(I_{1j} = I_{2j} = 0)$

This gives the model the flexibility to assign a SNP to either set, or remove it from the model completely. In Appendix 4, we give further details on the M-H algorithm to update the parameter space of this extended model.

As with our one-parameter causal model, we propose a simplified implementation of the two-parameter model in which the variance component parameters τ_1^2 and τ_2^2 are replaced with plug-in data derived estimates using the DL formula.

4.2 Simulation study

We conduct a simulation study to test the ability of our new two-parameter model to correctly estimate causal effects allowing for InSIDE violation. Two-sample summary data are simulated with 50 SNPs under balanced pleiotropy but with progressively larger proportion of SNPs maximally violating the InSIDE assumption. This changes the proportion of SNPs that are in set S_1 and S_2. These data are simulated under a strong instrument scenario ($\bar{F} = 100$, Scenario 5) and a weak instrument scenario ($\bar{F} = 25$, Scenario 6). For precise details of the simulation parameters see Table 3. We also explore the performance of our two-parameter model under balanced pleiotropy with weak and strong instruments when there is no InSIDE violation. That is, under Scenario 1 and 2. This means that all SNPs are effectively in set S_1 the data can be described with a single causal slope parameter, not two. The full results are shown in Table 4 where we report the bias, root mean squared error, coverage and mean Q statistic of all approaches across 1000 simulations.

Table 3: *Summary of InSIDE simulation scenarios*

Scenario	$\bar{F}$ of $S_1 : S_2$	Type of pleiotropy	S_1	S_2
5	100:100	Balanced	$\psi_j = 0$, $\Delta_j \sim N(0, 0.04)$, $\delta_j \sim U(0.34, 1.1)$, $\sigma_{Xj} \sim U(0.06, 0.095)$, $\beta_1 = \beta$	$\psi_j \sim U(0.34, 1.1)$, $\Delta_j = 0$, $\delta_j = 0$, $\sigma_{Xj} \sim U(0.06, 0.095)$, $\beta_2 = \beta + 1$
6	25:25	Balanced	$\psi_j = 0$, $\Delta_j \sim N(0, 0.04)$, $\delta_j \sim U(0.34, 1.1)$, $\sigma_{Xj} \sim U(0.06, 0.4)$, $\beta_1 = \beta$	$\psi_j \sim U(0.34, 1.1)$, $\Delta_j = 0$, $\delta_j = 0$, $\sigma_{Xj} \sim U(0.06, 0.4)$, $\beta_2 = \beta + 1$

4.3 Results

For data generated under Scenario 1 and 2, and so in the complete absence of InSIDE-violating SNPs in set S_2, our two slope model correctly identifies β and does not try to estimate a second effect, i.e. $\beta_1 = \beta_2$.

When the data are generated under Scenario 5 we see that, when S_1 and S_2 have a similar number of instruments, both β_1 and β_2 can be estimated by the DL implementation of our two-parameter model. If the proportion of SNPs in either set is too small, then our algorithm tends to remove them completely and focus on estimating just one slope.

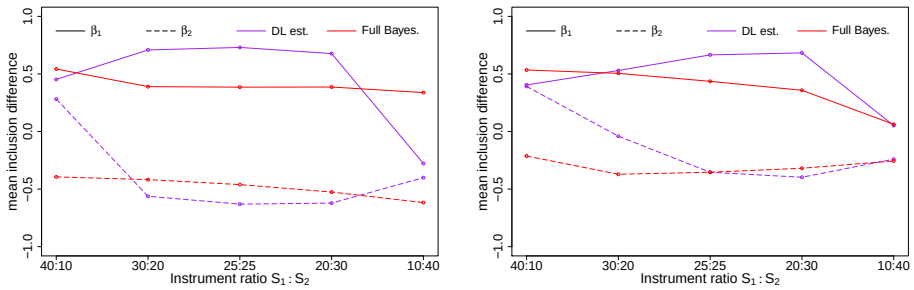
The full Bayesian implementation of our two slope model is shown to be more challenging to fit. It returns mean posterior estimates that are median unbiased but not mean unbiased. This demonstrates a lack of convergence, and indicates that longer iterations and a more sophisticated procedure for deciding on the tuning parameter may be required to properly fit the model.

When the data are generated with weaker instruments (Scenario 6), we see degrading in the performance of all approaches. In particular, see that the effect is worst for β_2. This is because, in our specific simulation, β_2 is larger in magnitude than β_1, which increases both the heterogeneity (as measured by Q) and the absolute magnitude of weak instrument bias relative to that experienced when estimating β_1. This in turn effected the coverage of the estimates. We further reduced the strength of the instruments, to have mean F-statistics of 10, our approach no longer assigns instruments to either set as the heterogeneity from the β_2 is too large to be pooled to a particular effect estimate, see Appendix 5.

When applying the full Bayesian approach in Scenario 6, we noticed an important feature most prominent when there was a large imbalance in the relative

sizes of S_1 and S_2. In this case, the M-H algorithm can switch from estimating the posterior for β_1 to estimating the posterior for β_2. This problem is referred to as "label switching" [30]. In our applied analysis in Section 5, we discuss this issue in more detail, and our proposal for addressing it.

Figure 3a (left and right) gives further insight into how well the DL and full Bayesian implementations can correctly partition the SNPs into their correct sets. The x-axis shows the true ratio of SNPs in S_1 and S_2 and the y-axis shows the difference in the mean probability of each SNP being assigned to S_1 and S_2. For the DL implementation, this probability is reassuringly high when the ratio is between 4:1 or 1:4, and is maximised when the ratio is 1:1. By contrast, the difference in mean inclusion probabilities for the full Bayesian approach are much more constant across all ratios and are also consistently lower.



(a) Scenario 5 (b) Scenario 6

Fig. 3: *Mean difference in inclusion probability between S_1 and S_2 as a function of the true ratio S_1:S_2 for Scenario 5* (a) *and Scenario 6* (b).

Table 4: *Evaluation criteria for estimating 2 causal parameter. 50 instruments in total. The true β is 0.05. Est., estimator; Inst., instrument(s); Q, exact Q-statistics; RMSE, root-mean-squared error.; DL est., DL estimate; Full Bayes., Full Bayesian*

Est.	**Inst.** $S_1 : S_2$	Q		mean bias		median bias		RMSE		coverage	
		S_1	S_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
		Scenario 1 *($\beta_1 = \beta_2 = \beta$)*									
DL est.	50:0	60.2	-	0.001	0.001	0.001	0.001	0.011	0.011	100.0	99.8
Full Bayes.	50:0	60.2	-	0.001	0.001	0.001	0.001	0.012	0.012	99.7	99.5
		Scenario 5 *($\beta_1 = \beta$, $\beta_2 = \beta + 1$)*									
DL est.	40:10	73.5	10.9	0.007	-0.876	0.001	-0.995	0.044	0.932	99.4	14.6
	30:20	55.1	23.8	0.003	-0.080	0.001	-0.013	0.040	0.264	95.9	92.2
	25:25	43.9	30.3	0.005	-0.009	0.001	-0.008	0.060	0.072	93.9	96.7
	20:30	35.3	36.9	0.053	-0.007	0.004	-0.007	0.211	0.051	90.9	95.5
	10:40	16.5	49.1	0.907	-0.009	0.988	-0.006	0.950	0.077	10.8	85.5
Full Bayes.	40:10	73.5	10.9	0.076	-0.287	0.003	-0.027	0.269	0.541	84.0	69.0
	30:20	55.1	23.8	0.230	-0.218	0.008	-0.009	0.481	0.471	69.4	76.2
	25:25	43.9	30.3	0.248	-0.182	0.011	-0.008	0.499	0.436	67.9	79.7
	20:30	35.3	36.9	0.254	-0.122	0.013	-0.002	0.519	0.360	66.8	86.1
	10:40	16.5	49.1	0.225	-0.041	0.017	0.003	0.738	0.283	62.4	95.4
		Scenario 2 *($\beta_1 = \beta_2 = \beta$)*									
DL est.	50:0	58.3	-	0.002	0.002	0.002	0.002	0.013	0.013	100.0	100.0
Full Bayes.	50:0	58.3	-	0.004	0.004	0.004	0.003	0.014	0.014	99.9	99.9
		Scenario 6 *($\beta_1 = \beta$, $\beta_2 = \beta + 1$)*									
DL est.	40:10	67.6	30.2	0.003	-0.985	0.002	-0.997	0.024	0.990	99.0	1.4
	30:20	50.3	65.7	0.038	-0.477	0.009	-0.391	0.111	0.643	97.5	59.4
	25:25	41.3	85.0	0.012	-0.099	0.006	-0.060	0.053	0.229	94.1	93.3
	20:30	32.8	102.6	0.007	-0.037	0.005	-0.033	0.034	0.120	94.6	96.8
	10:40	14.8	140.6	0.658	-0.080	0.785	-0.064	0.757	0.170	40.2	93.4
Full Bayes.	40:10	67.6	30.2	0.000	-0.336	0.003	-0.104	0.137	0.615	89.9	63.4
	30:20	50.3	65.7	0.023	-0.180	0.008	0.016	0.446	0.661	84.7	78.4
	25:25	41.3	85.0	0.035	-0.232	0.011	0.016	0.667	0.820	73.1	80.7
	20:30	32.8	102.6	0.006	-0.336	0.011	0.017	0.877	1.036	64.9	76.9
	10:40	14.8	140.6	-0.394	-0.591	0.003	0.012	1.548	1.461	32.7	76.2

5 Applied example: Age-related macular degeneration and cholesterol

Age-related macular degeneration (AMD) is a painless eye-disease that eventually leads to vision loss. Recent GWAS have identified several rare and common variants located in gene regions that are associated with lipid levels [31], fuelling speculation as to whether the relationship was causal [32, 33]. To this end, a multivariable MR analysis provided evidence to support a causal relationship between AMD and HDL cholesterol but not with LDL cholesterol and triglycerides [34]. In follow up work, Zuber *et al.* [35] fitted a multivariable MR model using Bayesian model averaging, with a total of 30 separate lipid fraction metabolites acting as the intermediate exposures. Out of the 30, large particle HDL cholesterol (XL.HDL.C) had the highest inclusion posterior probability as a risk factor for AMD.

Although multivariable MR approaches can remove bias due to pleiotropy via known pleiotropic pathways (in this case, other lipid fractions), they can be much more challenging to fit, especially when the correlation between the included exposures is high. For this reason we now revisit this data and use our univariate MR approaches to probe the causal relationship between XL.HDL.C and AMD.

We selected 54 genetic variants from Kettunen *et al.* [36] as instruments, due to having individual F-statistics for their association with XL.HDL.C greater than 2. Across all instruments this gave a mean F-statistic of 10. The summary scatter plot for these data is shown in Figure 4. The results for our various data analyses are given in Table 5.

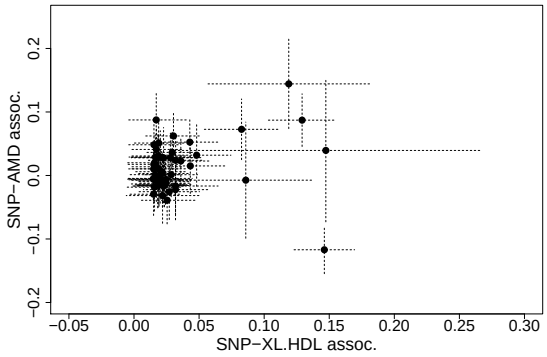
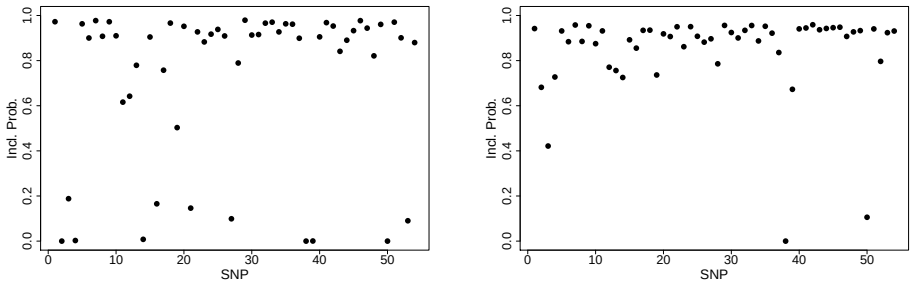


Fig. 4: *AMD and HDL: Scatter plot of the relationship between SNP-outcome and SNP-exposure association.*

Table 5: *Estimates for the causal effect of a unit increase in XL.HDL.C on the risk of AMD using a range of methods.*

Parameters	Estimator	Mean	95% Lower Interval	95% Upper Interval
Standard one-parameter approaches				
β	IVW	0.1576	-0.1003	0.4155
	MR-APS	0.2610	-0.0242	0.5462
	MR-RAPS	0.4705	0.2223	0.7187
BESIDE-MR: one-parameter model				
β	DL estimate	0.1908	-0.1661	0.6036
	Full Bayesian	0.6208	0.3595	0.8689
BESIDE-MR: two-parameter model				
β_1	DL estimate	0.8930	0.6123	1.2685
	Full Bayesian	0.8768	0.5318	1.2367
β_2	DL estimate	-0.6616	-0.9543	-0.2012
	Full Bayesian	-0.5135	-0.9355	0.5926

When one-parameter causal models are fitted to the data, all methods estimate a positive causal association, with full Bayesian BESIDE-MR and the MR-RAPS estimators giving the largest effects and the IVW method giving the smallest effect. This is not surprising because the IVW estimate is known to be vulnerable to weak instrument bias. Figure 5 shows the inclusion probability for each instrument, using our two implementations. The DL approach is seen to more aggressively select or de-select instruments than the full Bayesian approach.



(a) DL estimate (b) Full Bayesian

Fig. 5: *AMD and HDL: Inclusion probability for each instrument for DL estimate* (a) *and full Bayesian approach* (b).

Next, we fit our two-parameter causal model, which offers robustness to a substantial portion of the SNPs violating the InSIDE assumption. Interestingly,

we see that this estimates two distinct causal effects of opposite sign (Table 5). For the DL approach, approximately 17 SNPs have a strong inclusion probability (> 0.8) to each of the 2 clusters, with 1 SNP (rs103294) belonging to neither, (see Figure 6). For the full Bayesian approach, 12 instruments have a strong inclusion probability in the set identifying a positive relationship and only 2 instruments for the negative relationship (hence 0 is within the credible interval for this smaller set). SNP rs103294 also have low probability of being in either set. (Figure 7).

Our tentative conclusion here is that a small proportion of InSIDE-violating SNPs act to reduce the apparent causal effect of XL.HDL.C on AMD detectable by a one-parameter model. Once this set has been accounted for within a two-parameter model, this increases the evidence in favour of a causal role of XL.HDL.C on AMD further. Our results are consistent with Zuber *et al.* [35] who also found subsets of SNPs which suggested qualitatively different conclusions about the causal role of XL.HDL.C on AMD.

5.1 Detecting and adjusting for label switching in the full Bayesian model

The trace plots in Figure 8a and 8b show that the DL implementation consistently identifies two separate distributions for β_1 and β_2, which are centered around 0.89 and -0.66 respectively. This is not the case, however, under the full Bayesian implementation. Trace plots 8c and 8d show that the chains for β_1 and β_2 jumping between two distinct values pattern. This is commonly known as 'label switching'. It has been recommended that, instead of adjusting the MCMC algorithm itself, one can simply re-allocate iteration labels from the output [30] instead. To this end we performed a K-means clustering analysis [37] on the MCMC output. Before K-means correction, the mean posterior distribution of β_1 and β_2 gave 0.21 and 0.16 respectively. K-means analysis clustered 214,882 iterations centred at 0.88 and the second cluster contains 185,119 iterations with mean of -0.56. We re-assigned the estimates (to β_1 and β_2) accordingly (see Figure 8e) which gave new posterior distribution with mean and credible interval shown in Table 5. This issue further emphasizes the importance of carefully implementing the fully Bayesian approach, and for checking MCMC output for convergence issue.

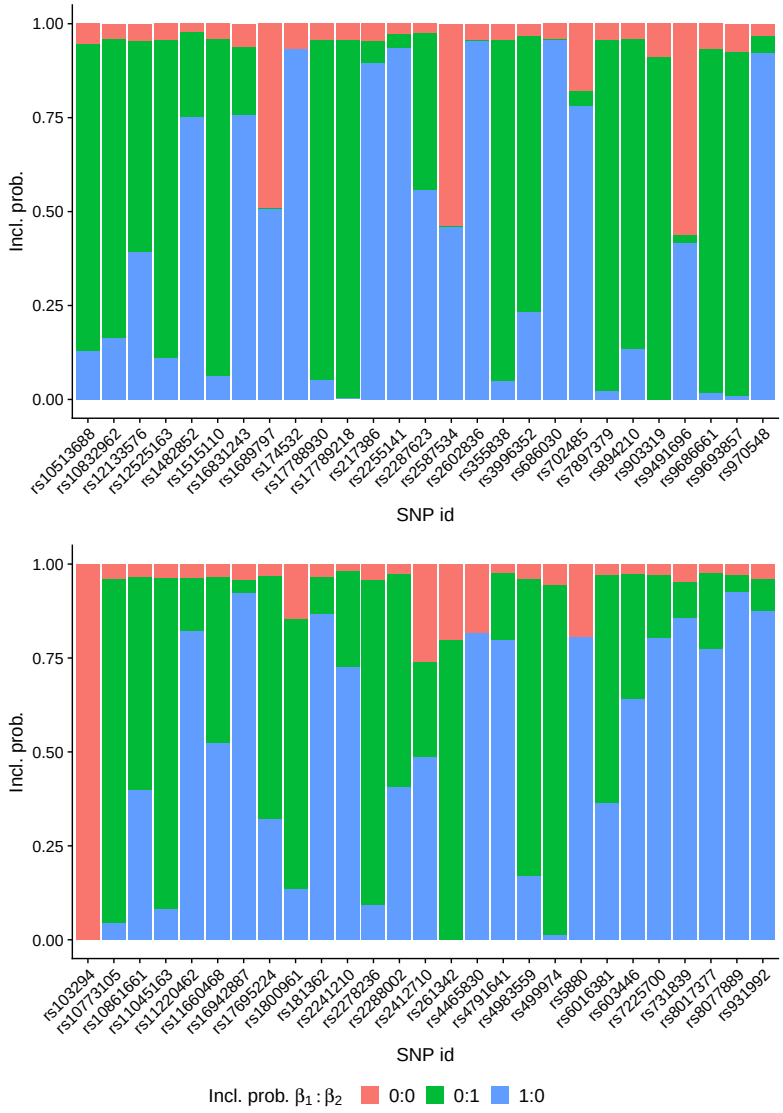


Fig. 6: *AMD and HDL: Inclusion probability for DL estimate, assuming InSIDE violation. As shown by legend; colour red, green and blue is for instrument in neither (0:0), instrument estimating β_2 (0:1) and β_1 (1:0) respectively.*

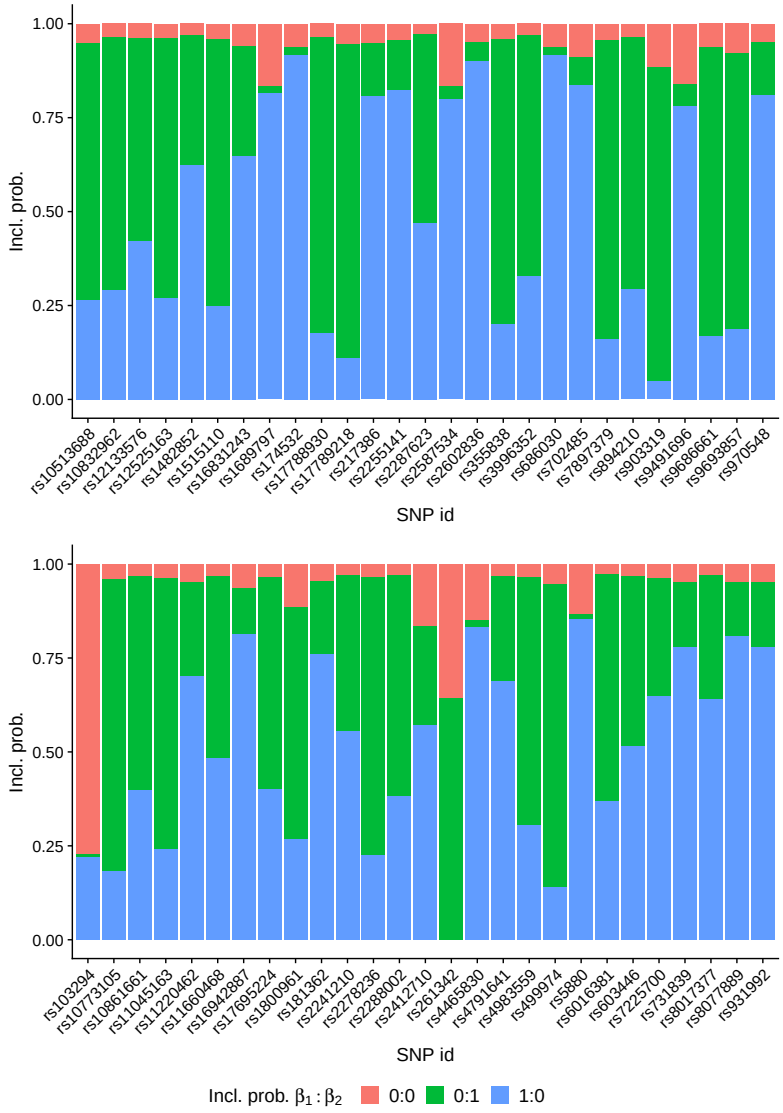
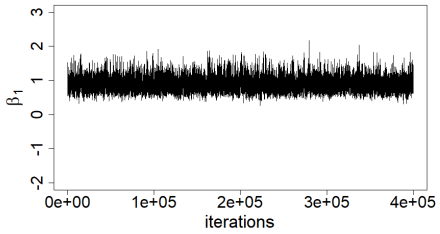
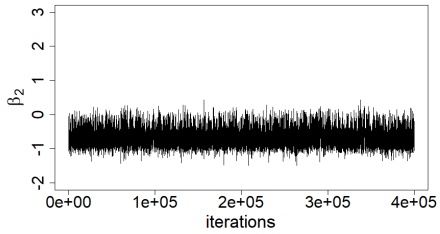


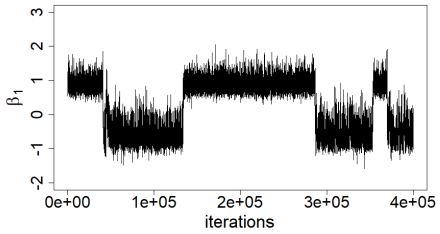
Fig. 7: *AMD and HDL: Inclusion probability for full Bayesian, assuming InSIDE violation. As shown by legend; colour red, green and blue is for instrument in neither (0:0), instrument estimating β_2 (0:1) and β_1 (1:0) respectively.*



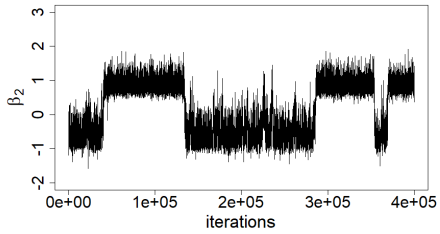
(a) DL estimate: β_1



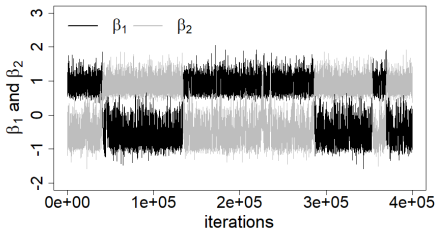
(b) DL estimate: β_2



(c) full Bayesian: β_1



(d) full Bayesian: β_2



(e) full Bayesian: β_1 and β_2

Fig. 8: AMD and HDL: Trace plots for β_1 and β_2 from the fully Bayesian and DL implementations of our algorithm.

6 Discussion

In this paper we propose a Bayesian Model Averaging approach that offers robustness pleiotropy and weak instruments. Our approach can be viewed as a Bayesian extension of the classical MR-RAPS approach, but with two advantages. The first is that SNPs deemed to be pleiotropic can be effectively downweighted, without dramatically affecting the coverage of the causal estimate. The second is robustness to large proportion of SNPs violating the InSIDE assumption. Rather than assuming the InSIDE violating SNPs are small in number and can be effectively penalized in the analysis, they can instead include be included using our two-parameter model. We were able to demonstrate the potential utility of this extended model in our applied example to uncover sub-signals in the data that would be missed by conventional methods. We explored two implementations of BESIDE-MR, namely the full Bayesian and the simplified DL implementation. Our simulations showed that the DL implementation generally performed well, and led to a more decisive selection of SNPs as either in or out of the model (or into set S_1 or S_2 in the two slope case) than the full Bayesian approach. It was also much more straightforward to fit and achieve convergence. Despite the fully Bayesian implementation requiring more computational time and careful consideration of the MCMC output, it is far better at detecting small effects and consistently identifying outlying instruments. We will attempt to improve the reliability of the full Bayesian approach. One aspect of this will be to create a label switching algorithm [38] for the output from full Bayesian model. Another would be to specify a more sophisticated procedure for optimising the tuning parameter for each model parameter separately. In the meantime, we urge users of the full Bayesian approach to manually adapt the tuning parameters and carefully monitor the mixing and convergence of the MCMC chains (especially for the latter approach), which are essential aspects of the analysis. As seen in Appendix 3, diagnostic tools such as performing multiple chains with different initial values and trace plots can be used in this regard. For a comprehensive tutorial see Albert [39] and Lunn *et al.* [40]. For our two-parameter model, it is also important to note that in the presence of weak instruments, the results from our approach must be interpreted with care and even more so when most instruments aren't assigned to either clusters.

A useful additional output from our BMA approach compared to classical approaches is the inclusion probability for each SNP. This of course necessitates the specification of a prior probability of inclusion, which we fixed at a constant value of $\frac{1}{2}$. Ideally, one should use informative priors where possible. Indeed, there are multiple sources of external information, e.g. epigenetic databases and bioinformatic webtools that could be used to achieve this. For example, a genetic variant that is located in a protein coding gene that is relevant to the pathway between exposure and outcome of interest can be given a higher inclusion prior probability. Conversely, we might give a much lower inclusion prior probability if the variant is located in a gene that is expressed in multiple tissues. This is again a topic for future research.

For individual-level data MR analyses, Berzuini *et al.* [41] have recently suggested a Bayesian approach that uses a horseshoe shrinkage prior on the possible pleiotropic effect of each instrument. When the pleiotropic effect is small, it is shrunk toward zero, thereby increasing the instrument's influence on the causal effect estimate. Their simulation showed their prior is reasonably robust to directional pleiotropy. Our approach is currently not robust to pure directional pleiotropy, although it is robust to apparent directional pleiotropy caused by violation of the InSIDE. As future work we intend to explore extending our approach to model pure directional pleiotropy using modifications to MR Egger regression Bowden *et al.* [13]. However, for the aforementioned reasons, the resulting estimator is likely to be imprecise and useful only in a limited number of circumstances.

We introduced our two-parameter BMA model for a univariate MR analysis. Zuber *et al.* [35] have already proposed a BMA implementation of multivariable MR [34, 42], which includes an algorithm for selecting both SNPs and exposure traits. Our model can in principle be extended to multivariable MR too. For a model with 10 exposure traits, this would necessitate the estimation of 20 causal parameters to account for InSIDE violation via unmeasured pathways. This is another topic for future research.

References

1. George Davey Smith and Shah Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 32(1):1–22, Feb 2003.
2. Sander Greenland. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*, 29(4):722–729, 2000.
3. Melinda C Mills and Charles Rahal. A scientometric review of genome-wide association studies. *Commun Biol*, 2(1):9, 2019.
4. Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *Rev Econ Stat*, 92(3):557–561, 2010.
5. Stephen Burgess, Adam Butterworth, and Simon G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*, 37(7):658–665, Nov 2013.
6. John R Thompson, Cosetta Minelli, and M Fabiola Del Greco. Mendelian randomization using public data from genetic consortia. *Int J Biostat*, 12(2), 2016.
7. Debbie A Lawlor. Commentary: Two-sample mendelian randomization: opportunities and challenges. *Int J Epidemiol*, 45(3):908, 2016.
8. Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score, 2018.
9. Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, Qingyuan Zhao, Debbie A Lawlor, Nuala A Sheehan, John Thompson, and George Davey Smith. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol*, 48(3):728–742, 12 2018.

10. Fabiola M Del Greco, Cosetta Minelli, Nuala A Sheehan, and John R Thompson. Detecting pleiotropy in mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*, 34(21):2926–2940, 2015.
11. Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala Sheehan, and John Thompson. A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. *Stat Med*, 36(11):1783–1802, 2017.
12. Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Hum Mol Genet*, 27(R2):R195–R208, 2018.
13. Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*, 44(2):512–525, 2015.
14. Michal Kolesár, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens. Identification and inference with many invalid instruments. *JBES*, 33(4):474–484, 2015.
15. Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*, 40(4):304–314, 2016.
16. Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*, 46(6):1985–1998, 2017.
17. Stephen Burgess, Verena Zuber, Apostolos Gkatzionis, and Christopher N Foley. 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. *Int J Epidemiol*, 47(4):1242–1254, 05 2018.
18. Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nat Genet*, 50(5):693, 2018.
19. Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Stat Sci*, 14(4):382–401, 1999.
20. Jack Bowden and Wes Spiller. The MR Data Challenge 2019. https://www.mendelianrandomization.org.uk/the-mr-data-challenge-2019/, 2019.
21. Judea Pearl. *Causality*. Cambridge university press, 2009.
22. Qingyuan Zhao, Jingshu Wang, Wes Spiller, Jack Bowden, Dylan S Small, et al. Two-sample instrumental variable analyses using heterogeneous samples. *Stat Sci*, 34(2):317–333, 2019.
23. Jack Bowden, Wesley Spiller, Fabiola Del Greco M, Nuala Sheehan, John Thompson, Cosetta Minelli, and George Davey Smith. Improving the visualization, interpretation and analysis of two-sample summary data mendelian randomization via the radial plot and radial regression. *Int J Epidemiol*, 47(4):1264–1278, 2018.
24. John R Thompson, Cosetta Minelli, Jack Bowden, Fabiola M Del Greco, Dipender Gill, Elinor M Jones, Chin Yang Shapland, and Nuala A Sheehan. Mendelian randomization incorporating uncertainty about pleiotropy. *Stat Med*, 36(29):4627–4645, 2017.
25. G. Koop, R. Leon-Gonzalez, and R. Strachan. Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics*, 171(2):237–250, 2012.

26. Alex Lenkoski, Anna Karl, and Andreas Neudecker. *ivbma: Bayesian Instrumental Variable Estimation and Model Determination via Conditional Bayes Factors*, 2014. R package version 1.05.
27. Chin Yang Shapland, John R Thompson, and Nuala A Sheehan. A Bayesian approach to mendelian randomisation with dependent instruments. *Stat Med*, 38(6):985–1001, 2019.
28. Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. Royal Stat. Soc: Series B (Statistical Methodology)*, 80(4):793–815, 2018.
29. Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.
30. Jean-Michel Marin and Christian P Robert. *Bayesian essentials with R*, volume 48. Springer, 2014.
31. Lars G Fritsche, Wilmar Igl, Jessica N Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L Bragg-Gresham, Kathryn P Burdon, Scott J Hebbring, Cindy Wen, Mathias Gorski, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*, 48(2):134, 2016.
32. Elisabeth M van Leeuwen, Eszter Emri, Benedicte MJ Merle, Johanna M Colijn, Eveline Kersten, Audrey Cougnard-Gregoire, Sascha Dammeier, Magda Meester-Smoor, Frances M Pool, Eiko K de Jong, et al. A new perspective on lipid research in age-related macular degeneration. *Progress in retinal and eye research*, 67:56–86, 2018.
33. JM Colijn, AI den Hollander, Ayse Demirkan, Audrey Cougnard-Grégoire, Timo Verzijden, Eveline Kersten, MA Meester, Benedicte MJ Merle, Grigorios Papageorgiou, Shahzad Ahmad, et al. Increased high density lipoprotein-levels associated with age-related macular degeneration. evidence from the eye-risk and e3 consortia. *Ophthalmology*, 126(3):393–406, 2018.
34. Stephen Burgess and George Davey Smith. Mendelian randomization implicates high-density lipoprotein cholesterol–associated mechanisms in etiology of age-related macular degeneration. *Ophthalmology*, 124(8):1165–1174, 2017.
35. Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nat. Commun*, 11(1):29, 2020.
36. Johannes Kettunen, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikäinen, Matti Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. *Nat. Commun*, 7:11122, 2016.
37. John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *J. Royal Stat. Soc: Series C (Applied Statistics)*, 28(1):100–108, 1979.
38. A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Stat Sci*, 20(1):50–67, 02 2005.
39. Jim Albert. *Bayesian computation with R*. Springer Science & Business Media, 2009.
40. D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012.

41. Carlo Berzuini, Hui Guo, Stephen Burgess, and Luisa Bernardinelli. A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*, 21(1):86–101, 08 2018.
42. Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int J Epidemiol*, 48(3):713–727, 12 2018.