BMC Medical Research Methodology

RESEARCH Open Access

Bayesian mendelian randomization with study heterogeneity and data partitioning for large studies

Linyi Zou, Hui Guo* and Carlo Berzuini

Abstract

Background: Mendelian randomization (MR) is a useful approach to causal inference from observational studies when randomised controlled trials are not feasible. However, study heterogeneity of two association studies required in MR is often overlooked. When dealing with large studies, recently developed Bayesian MR can be computationally challenging, and sometimes even prohibitive.

Methods: We addressed study heterogeneity by proposing a random effect Bayesian MR model with multiple exposures and outcomes. For large studies, we adopted a subset posterior aggregation method to overcome the problem of computational expensiveness of Markov chain Monte Carlo. In particular, we divided data into subsets and combined estimated causal effects obtained from the subsets. The performance of our method was evaluated by a number of simulations, in which exposure data was partly missing.

Results: Random effect Bayesian MR outperformed conventional inverse-variance weighted estimation, whether the true causal effects were zero or non-zero. Data partitioning of large studies had little impact on variations of the estimated causal effects, whereas it notably affected unbiasedness of the estimates with weak instruments and high missing rate of data. For the cases being simulated in our study, the results have indicated that the "divide (data) and combine (estimated subset causal effects)" can help improve computational efficiency, for an acceptable cost in terms of bias in the causal effect estimates, as long as the size of the subsets is reasonably large.

Conclusions: We further elaborated our Bayesian MR method to explicitly account for study heterogeneity. We also adopted a subset posterior aggregation method to ease computational burden, which is important especially when dealing with large studies. Despite the simplicity of the model we have used in the simulations, we hope the present work would effectively point to MR studies that allow modelling flexibility, especially in relation to the integration of heterogeneous studies and computational practicality.

Keywords: Mendelian randomization, Bayesian inference, Study heterogeneity, Data partitioning

Background

Mendelian randomization (MR) [1–3] is a useful approach to causal inference from observational studies when randomised controlled trials are not feasible. It uses genetic variants as instrumental variables (IVs) to explore putative causal relationship between an exposure and an outcome. Conventional MR methods [4–11] have mainly used summary statistics of IV-exposure association and IV-outcome association analyses, from a single study (one-sample) or two independent studies (two-sample). Among recent developments of MR methods, a Bayesian approach [8, 12] has been proposed to tackle overlapping samples in which a subset of participants are common in the two association studies. This comes from the idea that overlapping- and two- sample settings can be treated as cases of missing data which can be imputed through Markov chain Monte Carlo (MCMC) while estimating

*Correspondence: hui.guo@manchester.ac.uk
Centre for Biostatistics, School of Health Sciences, The University of Manchester, Oxford Road, M13 9PL Manchester, UK

causal effects of interest. This way, we take full advantage of all the observed and imputed data. Bayesian MR also offers great flexibility of modelling complex data structure and explicitly quantifies uncertainties of model parameters.

It is not uncommon that studies from different research groups are designed to address similar (but not exactly the same) scientific questions. For example, in a genome-wide association study (*Study* 1), data of genetic variants and hypertension status (outcome) are collected to identify outcome-associated genetic variants. In another independent study (*Study* 2), besides this aim, the investigator is also interested in causal effect of blood pressure medication (exposure) on hypertension. Therefore, exposure information is also recorded. To investigate the exposure-outcome causal relationship, a conventional option would be one-sample MR using data from *Study* 2 only. Another option would be a two-sample MR which will use genetic variants and the outcome data from *Study* 1, and genetic variants and the exposure data from *Study* 2. In other words, the outcome data of *Study* 2 will be discarded. Both of the options involve removal of data which, in our view, is not necessary. In fact, we can combine observed data from the two studies, and impute exposure data for *Study* 1 in a Bayesian MR model. However, it is well possible that the two studies are not homogenous, which should be taken into consideration in our modelling.

Another important aspect of Bayesian MR analysis (in fact, all kinds of data analysis) is tractability of computation, as we are in the era of big data. MCMC requires a large number of iterations and a complete scan of data for each iteration [13]. Thus, it can be computationally challenging, and sometimes even prohibitive. An intuitive solution would be dividing data into a number of subsets and enabling data analysis in parallel.

This paper aims to address study heterogeneity and data partitioning for large studies in Bayesian MR. First, we build a Bayesian MR model including multiple IVs, exposures and outcomes based on two independent studies, of which one has exposure data completely missing. To account for study heterogeneity, we propose a random effect model. Second, a data partitioning and subset posterior aggregation method [13] is adopted for analysis of large studies. Third, simulation experiments are carried out for different configurations of IV strength and missing rate of exposure data, followed by evaluation of our proposed method.

Methods

Bayesian MR with study heterogeneity

Let X denote the exposure, Y the outcome, and U a scalar variable summarising the set of unobserved confounders of the relationship between X and Y. Traditional MR [9] requires that an IV (denoted by Z) is : *i*) associated with the exposure X, *ii*) not associated with the confounders U, and *iii*) associated with the outcome Y only through the exposure X. These three assumptions can be graphically expressed as Fig. 1 in which our interest is whether X causes Y (the $X \rightarrow Y$ arrow). For the purpose of illustration, we consider the data generating process shown in Fig. 2. $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ are vectors consisting of L, K and M independent IVs respectively. Random scalar variables X_1 and X_2 represent two exposures. Random scalar variables Y_1 and Y_2 represent two outcomes.

In a two-sample (or equivalently, two-study) MR setting with or without overlapping individuals, it has been shown that, compare to conventional MR analysis, a Bayesian approach may lead to more precise estimates of the causal effect by treating it as a case of incomplete data which may be dealt with through iterative imputations using MCMC [12]. Here, we further generalize the approach by allowing for some degree of heterogeneity between different studies.

Suppose we have data collected from two independent studies:

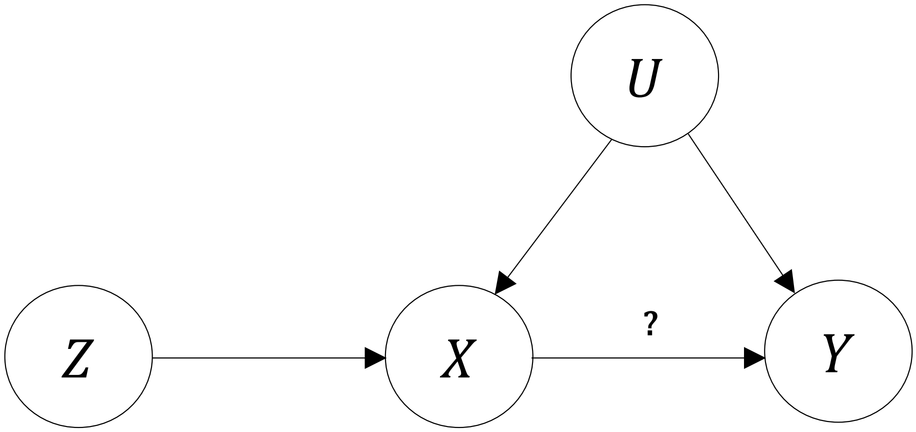

Fig. 1 Schematic representation of the three assumptions required in Mendelian randomization

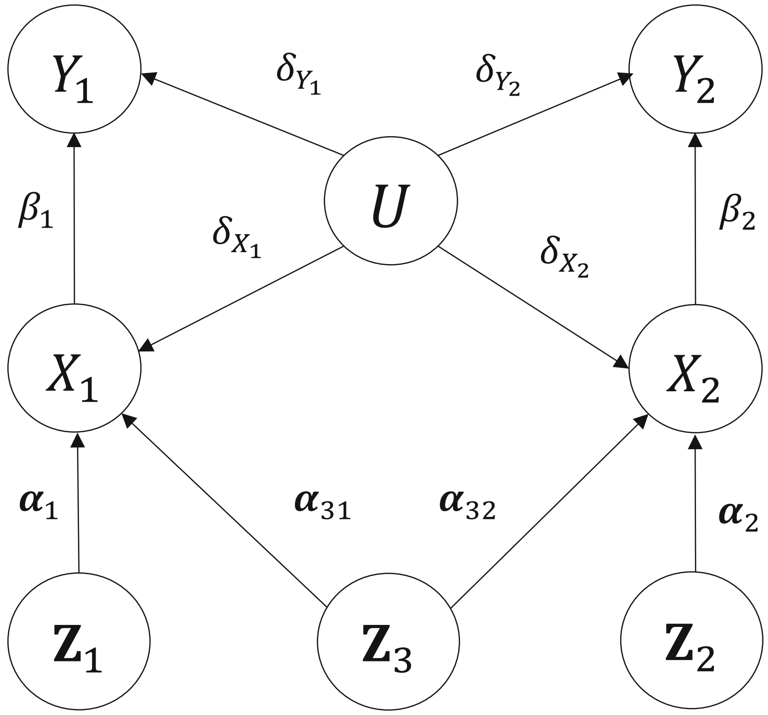

Fig. 2 Graphical model of Mendelian randomisation with outcomes Y_1 and Y_2, exposures X_1 and X_2 and unobserved confounder U. $\mathbf{Z}_1$ consists of L instrumental variables of X_1 and $\mathbf{Z}_2$ consists of K instrumental variables of X_2. In addition, $\mathbf{Z}_3$ consists of M instrumental variables shared between X_1 and X_2. The instrumental variables are assumed to be mutually independent

- *Study A* - observed data for IVs, exposures and outcomes $\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, X_1, X_2, Y_1, Y_2\}$.
- *Study B* - observed data for IVs and outcomes $\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, Y_1, Y_2\}$ only.

Study A includes fully observed data for MR, whereas *Study B* has exposure data completely missing. We shall include random effect terms in our MR model to capture study heterogeneity. By assuming standardised observed variables and linear additivity, according to Fig. 2, our models are constructed as follows.

For *Study A*,

$$U \sim N(0,1), \tag{1}$$

$$X_1|\mathbf{Z}_1, \mathbf{Z}_3, U \sim N\left(\boldsymbol{\alpha}_1\mathbf{Z}_1 + \boldsymbol{\alpha}_{31}\mathbf{Z}_3 + \delta_{X_1}U, \sigma^2_{X_{1A}}\right), \tag{2}$$

$$X_2|\mathbf{Z}_2, \mathbf{Z}_3, U \sim N\left(\boldsymbol{\alpha}_2\mathbf{Z}_2 + \boldsymbol{\alpha}_{32}\mathbf{Z}_3 + \delta_{X_2}U, \sigma^2_{X_{2A}}\right), \tag{3}$$

$$Y_1|X_1, U \sim N\left(\beta_1 X_1 + \delta_{Y_1}U, \sigma^2_{Y_{1A}}\right), \tag{4}$$

$$Y_2|X_2, U \sim N\left(\beta_2 X_2 + \delta_{Y_2}U, \sigma^2_{Y_{2A}}\right). \tag{5}$$

For *Study B*,

$$U \sim N(0,1), \tag{6}$$

$$X_1|\mathbf{Z}_1, \mathbf{Z}_3, U \sim N\left(V_{X_1} + \boldsymbol{\alpha}_1\mathbf{Z}_1 + \boldsymbol{\alpha}_{31}\mathbf{Z}_3 + \delta_{X_1}U, \sigma^2_{X_{1B}}\right), \tag{7}$$

$$X_2|\mathbf{Z}_2, \mathbf{Z}_3, U \sim N\left(V_{X_2} + \boldsymbol{\alpha}_2\mathbf{Z}_2 + \boldsymbol{\alpha}_{32}\mathbf{Z}_3 + \delta_{X_2}U, \sigma^2_{X_{2B}}\right), \tag{8}$$

$$Y_1|X_1, U \sim N\left(V_{Y_1} + \beta_1 X_1 + \delta_{Y_1}U, \sigma^2_{Y_{1B}}\right), \tag{9}$$

$$Y_2|X_2, U \sim N\left(V_{Y_2} + \beta_2 X_2 + \delta_{Y_2}U, \sigma^2_{Y_{2B}}\right). \tag{10}$$

In the above pre-specified models, $\boldsymbol{\alpha}$s are instrument strength parameters, and δs are effects of U on Xs or Ys. Causal effects of Xs on Ys are denoted by βs. The study heterogeneity is accounted for by Vs. Note that X_1 and X_2 do not have observed data in *Study B*, but they are part of data generating process, and thus, should be included in the model. U is a sufficient scalar summary of the unobserved confounders. We assume that $U \sim N(0,1)$.

The combined dataset of *Studies A* and *B* ($\mathcal{D}$, say) will contain fully observed data for the instruments and the outcomes. However, all participants in *Study B* have missing data of X_1 and X_2 which will be treated as unknown quantities and imputed from their conditional distributions given the observed data and current estimated parameters using MCMC. Let X^* be imputed values of X. Our approach involves the following sequence of five steps.

1. Specify initial values for unknown parameters and the number of Markov iterations T.
2. At the tth iteration, where $0 \leq t < T$, let missing values of X_1 and X_2 in *Study B* be filled with X_1^* drawn from $N\left(V_{X_1}^{(t)} + \boldsymbol{\alpha}_1^{(t)}\mathbf{Z}_1 + \boldsymbol{\alpha}_{31}^{(t)}\mathbf{Z}_3 + \delta_{X_1}^{(t)}U, {\sigma^2_{X_{1B}}}^{(t)}\right)$ and X_2^* drawn from $N\left(V_{X_2}^{(t)} + \boldsymbol{\alpha}_2^{(t)}\mathbf{Z}_2 + \boldsymbol{\alpha}_{32}^{(t)}\mathbf{Z}_3 + \delta_{X_2}^{(t)}U, {\sigma^2_{X_{2B}}}^{(t)}\right)$,

respectively. $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ are observed values of IVs in *Study B*.

3. Create a single complete dataset including both the observed and the imputed data.
4. Estimate model parameters using MCMC based on the complete dataset and set $t \leftarrow t+1$.
5. Repeat Steps 2-4 until $t = T$.

Now we specify priors in the Bayesian model (2)-(10). Previous GWAS studies show that individual SNPs explain a tiny proportion of exposure variance [14–16], corresponding to small magnitudes of the $\boldsymbol{\alpha}$ parameters in our model. In accord with this finding and previous MR simulation studies [17], we set IV strength parameters $\boldsymbol{\alpha}$s to be independent and identically distributed with mean zero and a small variance: $\boldsymbol{\alpha}_1 \sim N_L\left(\mathbf{0}, 0.3^2\mathbf{I}\right)$, $\boldsymbol{\alpha}_2 \sim N_K\left(\mathbf{0}, 0.3^2\mathbf{I}\right)$, $\boldsymbol{\alpha}_{31} \sim N_M\left(\mathbf{0}, 0.3^2\mathbf{I}\right)$, and $\boldsymbol{\alpha}_{32} \sim N_M\left(\mathbf{0}, 0.3^2\mathbf{I}\right)$. The priors of both β_1 and β_2 are set to a same distribution $N(0, 10^2)$. Finally, we assign the priors of the standard deviations σs to a same inverse-gamma distribution *Inv-Gamma*$(3, 2)$, and random effects Vs to $N(0, 1)$ in the Model (7)-(10) for *Study B*.

Bayesian MR for large studies

Advantages of an MCMC-powered Bayesian approach to MR are counterpoised by a relatively higher computational burden and a possibly large memory requirement. A natural way of dealing with this problem would be to divide data $\mathcal{D}$ into a number (J, say) of subsets $D_1, D_2, ..., D_J$ that we assume to contain an equal number (q, say) of individuals for simplicity. By running separate Bayesian MR analyses in parallel on the subsets, we will obtain J subset-specific posteriors which can then be aggregated in various ways. In this study, we adopt a "divide-and-combine" approach proposed by Xue and Liang [13] .

Let $\boldsymbol{\theta}$ denote the entire set of unknown quantities in the model. For subset D_j, where $j = 1, 2, ..., J$, let $\pi(\boldsymbol{\theta}|D_j)$ denote the joint posterior distribution of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\mu}}_j = \widehat{E}(\boldsymbol{\theta}|D_j)$ the corresponding estimated mean vector. Let $\widehat{\boldsymbol{\mu}} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\boldsymbol{\mu}}_j$ be the average of the $\widehat{\boldsymbol{\mu}}_j$s. According to [13], the posterior based on full data, $\pi(\boldsymbol{\theta}|\mathcal{D})$, can be estimated as the average of the recentred subset posteriors.

$$\widetilde{\pi}(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{J}\sum_{j=1}^{J}\widetilde{\pi}(\boldsymbol{\theta} - \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\mu}}_j|D_j). \tag{11}$$

And it has been proved that ([13])

$$E_{\widetilde{\pi}}(\boldsymbol{\theta}) - E_{\pi}(\boldsymbol{\theta}) = O_p\left(q^{-1}\right), \tag{12}$$

and

$$Var_{\widetilde{\pi}}(\boldsymbol{\theta}) - Var_{\pi}(\boldsymbol{\theta}) = o_p\left(n^{-1}\right), \tag{13}$$

where q is the sample size of the subsets and n the sample size of the full dataset. $E_{\widetilde{\pi}}(\boldsymbol{\theta})$ and $E_{\pi}(\boldsymbol{\theta})$ are expectations of the posteriors of $\boldsymbol{\theta}$ aggregated from subsets and obtained from full data respectively. $Var_{\widetilde{\pi}}(\boldsymbol{\theta})$ and $Var_{\pi}(\boldsymbol{\theta})$ are their variances. It is easily seen that the difference in expectation depends on the sample size of the subsets and the difference in variation depends on the sample size of the full dataset.

Simulations - Bayesian MR with study heterogeneity

We used simulated data to evaluate our Bayesian MR model with study heterogeneity in comparison with a conventional MR method. In particular, we considered 12 configurations including

- 3 missing rates of the exposures: 20%, 50%, 80%
- 2 degrees of the IV strength $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_{31}, \boldsymbol{\alpha}_{32})$: $\mathbf{0.1}$ and $\mathbf{0.3}$
- Zero and non-zero causal effects of the exposures on the outcomes (β_1, β_2): 0 and 0.3.

The number of IVs was set to 15, 15 and 5 for Z_1,Z_2 and Z_3 respectively. Data of each IV were randomly drawn from a binomial distribution $B(2, 0.3)$ independently. The specified values of the effects of U on the exposures $(\delta_{X_1}, \delta_{X_2})$ and on the outcomes $(\delta_{Y_1}, \delta_{Y_2})$ were set to 1. Standard deviations σs were set to 0.1. We simulated 200 datasets for each configuration.

For each dataset, we

- simulated a dataset of sample size n_A which contains observations of the IVs, exposures and outcomes (dataset A, denoted by $\mathcal{D}_A$);
- simulated a dataset of sample size n_B which contains observations of the IVs, exposures and outcomes, then included data of the IVs and outcomes only as if the exposure data were missing (dataset B, denoted by $\mathcal{D}_B$).

Sample size of $\mathcal{D}$, the combined data of $\mathcal{D}_A$ and $\mathcal{D}_B$, was set to 400 in all configurations, i.e., $n = n_A + n_B = 400$. The missing rate of the exposures was defined as $\frac{n_B}{n} \times 100\%$. For example, if missing rate was 50%, we simulated $\mathcal{D}_A$ of sample size 200 and $\mathcal{D}_B$ of sample size 200. To allow for different degrees of study heterogeneity in different datasets, random effects Vs in study B were randomly drawn from a uniform distribution $U(-0.5, 0.5)$ independently. Imputations of missing data and estimations of model parameters were then performed simultaneously using MCMC in `Stan` [18, 19]. $\hat{R}$ was used to check convergence of the Markov chains [20].

Estimated causal effects obtained from our Bayesian MR and two-sample inverse-variance weighted (IVW) estimation [6] were compared using 4 metrics: mean, standard deviation (sd), coverage (proportion of the times that the 95% credible/confidence intervals contained the true value of the causal effect) and power (proportion of the times that the 95% credible/confidence intervals did not contain zero when the true causal effect was non-zero, only applicable when $\beta_1 = \beta_2 = 0.3$ by defination). Higher power indicates lower chance of getting false negative results. In IVW estimation, we used observed IV and exposure data from $\mathcal{D}_A$ and observed IV and outcome data from $\mathcal{D}_B$.

Simulations - Bayesian MR with study heterogeneity for large studies

We also assessed the performance of dividing a big dataset into subsets in our Bayesian MR with study heterogeneity in simulation experiments. The simulation scheme was the same as above. However, the sample size of $\mathcal{D}$ was set to a much larger value 50,000. For each configuration, a single dataset was simulated by combining $\mathcal{D}_A$ and $\mathcal{D}_B$. We randomly divided data into 5 subsets of equal sample size, separately, for $\mathcal{D}_A$ ($\mathcal{D}_{A_1}, ..., \mathcal{D}_{A_5}$) and for $\mathcal{D}_B$ ($\mathcal{D}_{B_1}, ..., \mathcal{D}_{B_5}$). Subset $\mathcal{D}_i$ was then constructed by combining $\mathcal{D}_{A_i}$ and $\mathcal{D}_{B_i}$, where $i = 1, ..., 5$. This is to ensure that subset $\mathcal{D}_i$ had the same missing rate as that of the full data $\mathcal{D}$. Causal effects were estimated using $\mathcal{D}$, and using the 5 subsets in Bayesian MR. To explore the impact of different data partitioning strategies on estimated causal effects, we carried out the same analysis by also dividing data into 50 subsets of sample size 1,000.

Results

$\hat{R}$ values of all the parameters in the models (2)-(5) and (7)-(10) were greater than 1 and less than 1.1 across the simulations.

Simulation results - Bayesian MR with study heterogeneity

Table 1 displays simulation results when the true causal effects were non-zero ($\beta_1 = \beta_2 = 0.3$). Each row of the table corresponds to a configuration of a specified missing rate and a degree of IV strength $\boldsymbol{\alpha}$. Columns correspond to the estimated causal effects of X_1 on Y_1 ($\hat{\beta}_1$) and of X_2 on Y_2 ($\hat{\beta}_2$) from our Bayesian method and from the IVW method evaluated using the four metrics. Unsurprisingly, the estimated causal effect of X_1 on Y_1 was very similar to that of X_2 on Y_2 in each configuration from Bayesian MR, because their true values were set to be the same and the model had a symmetric structure as shown in Fig. 2. This was also observed in the results from the IVW method. However, Bayesian MR outperformed IVW uniformly across all the configurations, with less bias, higher precision, coverage and power. The impact of low missing rate was positive on coverage but negative on power in IVW. However, such impact was negligible in Bayesian MR. This was mainly due to much higher variations of the estimates, and consequently, much wider confidence intervals in IVW estimation. Weaker IVs had little influence on unbiasedness of the estimates and power, but resulted in slightly lower precision and coverage in Bayesian MR. However, there was a remarkable decrease in unbiasedness, precision and power in IVW as IV strength decreased.

Table 2 presents simulation results when the true causal effects were zero ($\beta_1 = \beta_2 = 0$). Again, the results of $\hat{\beta}_1$ was very similar to those of $\hat{\beta}_2$ in each configuration, separately, from Bayesian MR and from IVW. Overall, both methods performed well. However, Bayesian MR still outperformed IVW across all the configurations, with higher coverage and precision and less biased estimates. In both MR methods, missing rate did not have a notable effect on the estimates, whereas weaker IVs led to lower precision.

Simulation results - Bayesian MR with study heterogeneity for large studies

Figure 3 depicts the joint posterior distributions of $\hat{\beta}_1$ (horizontal axis) and $\hat{\beta}_2$ (vertical axis) based on simulated data when the true causal effects were non-zero. Columns corresponds to three missing rates and rows two levels of IV strength. In each panel, the black dot denotes the values of true causal effects ($\beta_1 = \beta_2 = 0.3$). The red, orange and blue contours are 2-dimensional Gaussian kernel density estimation of the joint posterior (GKDEJP) from the full dataset, aggregated GKDEJP from five subsets and aggregated GKDEJP from fifty subsets respectively. When IVs were strong in Bayesian MR analysis (top panels), estimated causal effects were close to their true values, with or without data partitioning. When IVs became weaker (bottom panels), the results from the full data were concordant with those from 5 subsets, but notably different from those based on 50 subsets. The impact of data partitioning was substantial with weak IVs and high missing rate. This could be explained by Equation (12), in which difference in mean of the GKDEJPs depends on the subset sample size q. Difference in variance of the GKDEJPs was, however, not evident in the three sets of contours in each configuration, because it only depends on the sample size of the full data (Equation (13)) which was a fixed value 50,000. Our simulation results suggest that, in Bayesian MR with a large sample size, there is a trade-off between data partitioning for more efficient computations, and large enough sample size of each subset for preventing estimates from a decrease in unbiasedness.

The same plots were presented in Fig. 4 when the true causal effects were zero. The performances of the three data partition strategies were very similar to those when the true causal effects were non-zero.

Discussion and conclusions

Numerous MR methods have been developed in recent years. To the best of our knowledge, little attention has been focused on study heterogeneity. In this study, we

Table 1 Causal effects estimated from 200 simulated datasets for each configuration from two MR methods (Bayesian, IVW) when $\beta_1 = \beta_2 = 0.3$, using four metrics: mean, standard deviation (sd), coverage and power. The six configurations were generated from three missing rates of the exposures (80%, 50%, 20%) and two levels of IV strength ($\boldsymbol{\alpha} = \mathbf{0.3}$ and $\mathbf{0.1}$). $\hat{\beta}_1$: estimated causal effect of X_1 on Y_1, $\hat{\beta}_2$: estimated causal effect of X_2 on Y_2

		$\widehat{\beta_1}$								$\widehat{\beta_2}$							
Missing rate	α	**Bayesian**				**IVW**				**Bayesian**				**IVW**			
		mean	**sd**	**coverage**	**power**	**mean**	**sd**	**coverage**	**power**	**mean**	**sd**	**coverage**	**power**	**mean**	**sd**	**coverage**	**power**
80%	0.3	0.299	0.005	0.980	1	0.217	0.101	0.790	0.685	0.298	0.005	0.970	1	0.209	0.086	0.765	0.665
	0.1	0.298	0.015	0.975	1	0.081	0.141	0.695	0.065	0.299	0.015	0.985	1	0.071	0.146	0.690	0.045
50%	0.3	0.300	0.004	0.975	1	0.245	0.118	0.920	0.580	0.299	0.004	0.980	1	0.265	0.113	0.935	0.595
	0.1	0.302	0.013	0.960	1	0.169	0.277	0.925	0.115	0.302	0.013	0.955	1	0.122	0.268	0.900	0.075
20%	0.3	0.299	0.004	0.970	1	0.260	0.203	0.915	0.255	0.299	0.004	0.970	1	0.276	0.185	0.955	0.285
	0.1	0.303	0.012	0.955	1	0.193	0.439	0.945	0.050	0.302	0.012	0.950	1	0.181	0.469	0.945	0.070

Table 2 Causal effects estimated from 200 simulated datasets for each configuration from two MR methods (Bayesian, IVW) when $\beta_1 = \beta_2 = 0$, using four metrics: mean, standard deviation (sd), coverage and power. The six configurations were generated from three missing rates of the exposures (80%, 50%, 20%) and two levels of IV strength ($\boldsymbol{\alpha} = \mathbf{0.3}$ and $\mathbf{0.1}$). $\hat{\beta}_1$: estimated causal effect of X_1 on Y_1, $\hat{\beta}_2$: estimated causal effect of X_2 on Y_2

		$\widehat{\beta_1}$						$\widehat{\beta_2}$					
Missing rate	α	**Bayesian**			**IVW**			**Bayesian**			**IVW**		
		mean	**sd**	**coverage**	**mean**	**sd**	**coverage**	**mean**	**sd**	**coverage**	**mean**	**sd**	**coverage**
80%	0.3	-0.001	0.005	0.960	0.007	0.061	0.955	-0.001	0.005	0.955	-0.005	0.062	0.960
	0.1	0.004	0.016	0.960	-0.010	0.112	0.965	0.004	0.015	0.960	-0.001	0.130	0.960
50%	0.3	0.000	0.005	0.975	-0.014	0.087	0.935	0.000	0.005	0.955	-0.002	0.090	0.955
	0.1	0.004	0.013	0.970	0.005	0.188	0.960	0.004	0.013	0.955	-0.011	0.202	0.950
20%	0.3	0.000	0.004	0.950	0.010	0.148	0.930	0.000	0.004	0.965	-0.003	0.152	0.935
	0.1	0.003	0.012	0.965	0.012	0.394	0.920	0.003	0.012	0.965	0.020	0.361	0.945

further elaborated our Bayesian MR method [8, 12] by including random effects to explicitly account for study heterogeneity. We also adopted a subset posterior aggregation method [13] to address the computational challenge of MCMC, which is important especially when dealing with large studies. For the cases being simulated in our study, the results have indicated that the "divide (data) and combine (estimated subset causal effects)" can help improve computational efficiency, for an acceptable cost in terms of bias in the causal effect estimates, as long as the size of the subsets is reasonably large. However, when instruments are weak and data missing rate is high, the results obtained using data partitioning are noticeably different from those obtained using full data. Hence, there is room for further development of robust and computationally efficient methods for Bayesian MR.

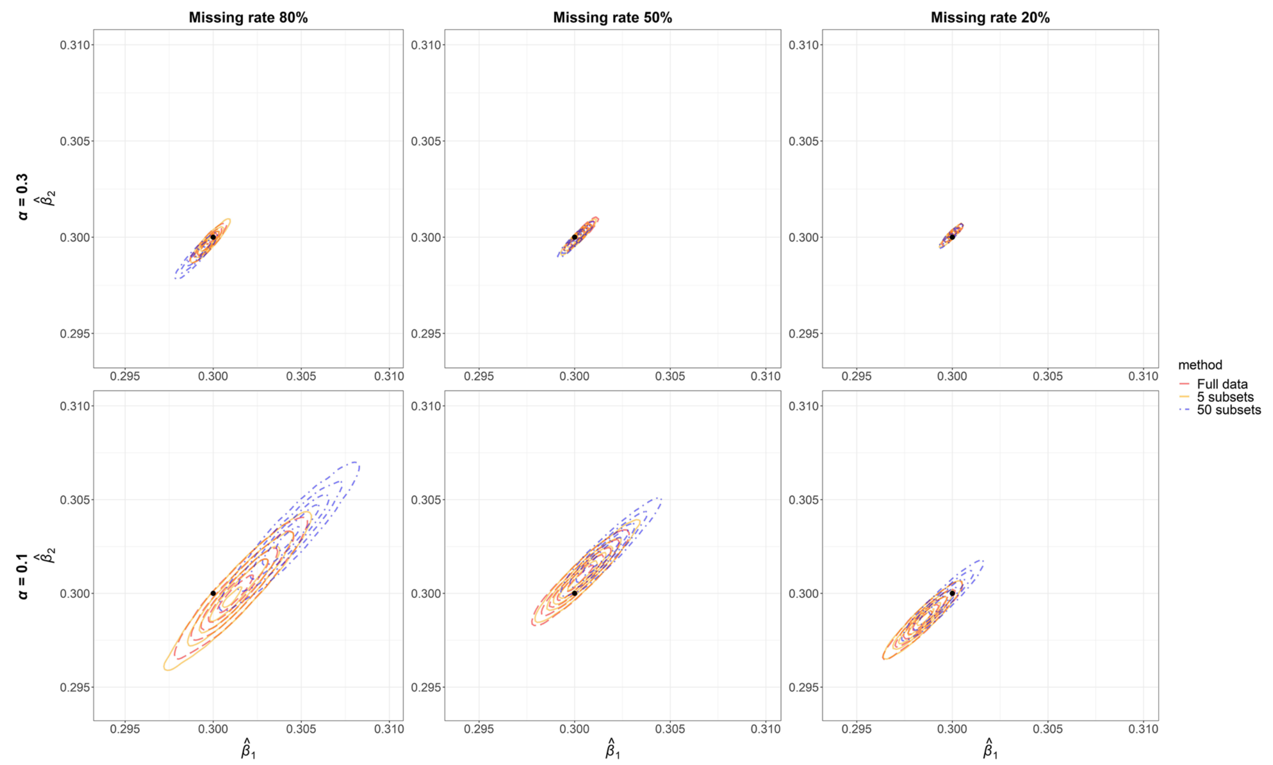

Fig. 3 Joint posterior distributions of causal effects of X_1 on Y_1 ($\hat{\beta}_1$, horizontal axis) and X_2 on Y_2 ($\hat{\beta}_2$, vertical axis) obtained from 2-dimensional Gaussian kernel density estimation in Bayesian Mendelian randomisation when the true causal effects $\beta_1 = \beta_2 = 0.3$. Results were based on full data (red), 5 subsets with equal sample size (orange) and 50 subsets with equal sample size (blue). $\boldsymbol{\alpha}$: instrument strength

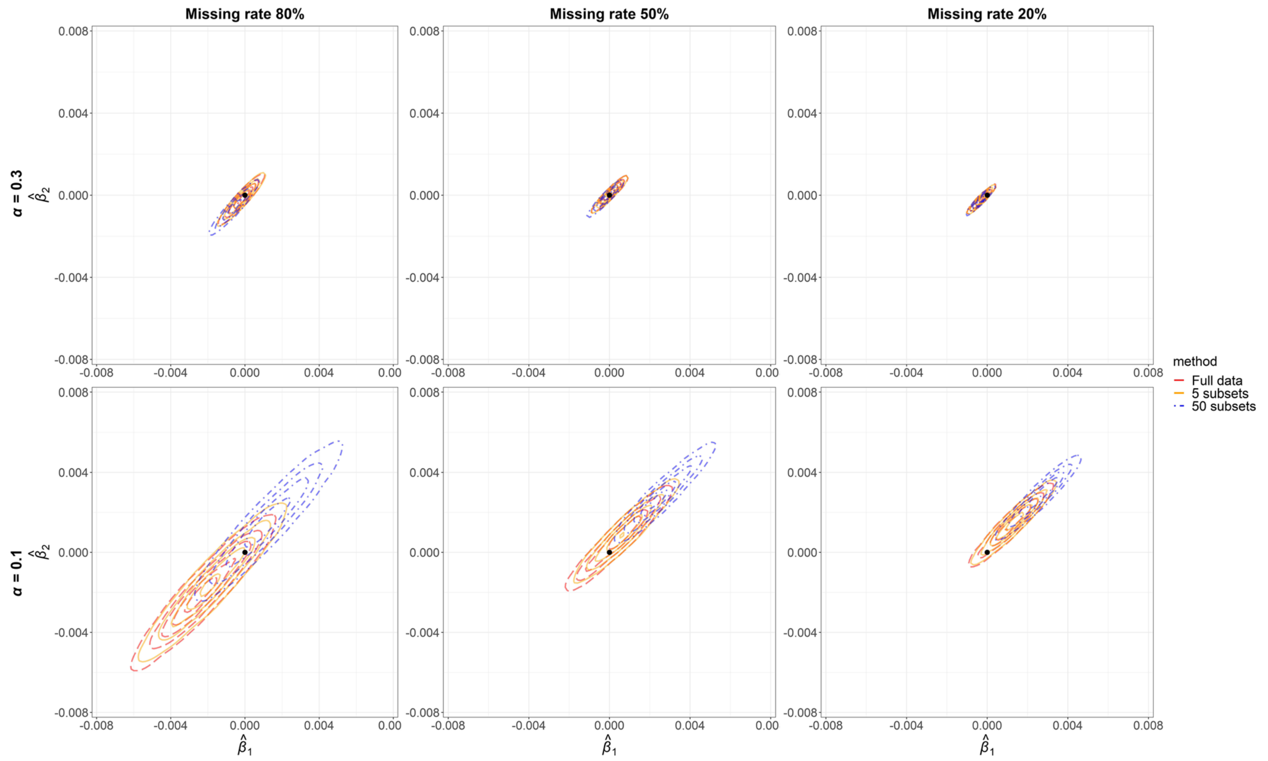

Fig. 4 Joint posterior distributions of causal effects of X_1 on Y_1 ($\hat{\beta}_1$, horizontal axis) and X_2 on Y_2 ($\hat{\beta}_2$, vertical axis) obtained from 2-dimensional Gaussian kernel density estimation in Bayesian Mendelian randomisation when the true causal effects $\beta_1 = \beta_2 = 0$. Results were based on full data (red), 5 subsets with equal sample size (orange) and 50 subsets with equal sample size (blue). $\boldsymbol{\alpha}$: instrument strength

Despite the simplicity of the model we have used in the simulations, we hope the present work would effectively point to MR studies that allow modelling flexibility, especially in relation to the integration of heterogeneous studies and computational practicality.

Abbreviations

MR: Mendelian randomization; IV: Instrumental variable; MCMC: Markov chain Monte Carlo; IVW: Inverse-Variance Weighted; GKDEJP: Gaussian kernel density estimation of the joint posterior

Acknowledgements

We thank Professor Jeanine Houwing-Duistermaat for insightful discussion on study heterogeneity.

Authors' contributions

HG conceived the study. HG and CB supervised the study. LZ performed simulations and statistical analysis. LZ, HG and CB interpreted statistical results and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by Manchester-CSC. The funder had no role in study design, data generation and statistical analysis, interpretation of data or preparation of the manuscript.

Availability of data and materials

The code of data simulations is available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Received: 21 December 2021 Accepted: 20 April 2022
Published online: 03 June 2022

References

1. Katan MB. Apolipoprotein e isoforms, serum cholesterol, and cancer. Lancet. 1986;327:507–8.
2. Smith GD, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1–22.
3. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Int J Epidemiol. 2008;27:1133–63.
4. Johnson T. Efficient calculation for multi-snp genetic risk scores. Technical report. 2013. http://cran.r-project.org/web/packages/gtx/vignettes/ashg2012.pdf.
5. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. Int J Epidemiol. 2015;44(2):512–25.

6. Bowden J, Smith GD, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol. 2016;40:304–14.
7. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical Inference in Two-sample Summary-data Mendelian Randomization Using Robust Adjusted Profile Score. Ann Statist. 48(3):1742–69.
8. Berzuini C, Guo H, Burgess S, Bernardinelli L. A bayesian approach to mendelian randomization with multiple pleiotropic variants. Biostatistics. 2018;21(1):86–101.
9. Burgess S, Thompson SG. MENDELIAN RANDOMIZATION Methods for Using Genetic Variants in Causal Estimation. London: Chapman & Hall/CRC Press; 2015.
10. Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variable regression. J Econ. 2003;114(1):29–72.
11. Jones EM, Thompson JR, Didelez V, Sheehan NA. On the choice of parameterisation and priors for the bayesian analyses of mendelian randomisation studies. Stat Med. 2012;31(14):1483–501.
12. Zou L, Guo H, Berzuini C. Overlapping-sample mendelian randomisation with multiple exposures: a bayesian approach. BMC Med Res Methodol. 2020;20:295.
13. Xue J, Liang F. Double-parallel monte carlo for bayesian analysis of big data. Stat Comput. 2019;29(1):23–32.
14. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, Song K, Yuan X, Johnson T, Ashford S, Inouye M, Luben R, Sims M, Hadley D, McArdle W, Barter P, Kesäniemi YA, Mahley RW, McPherson R, Grundy SM, Consortium WTCC, Bingham SA, Khaw K-T, Loos RJF, Waeber G, Barroso I, Strachan DP, Deloukas P, Vollenweider P, Wareham NJ, Mooser V. Ldl-cholesterol concentrations: a genome-wide association study. Lancet (London, England). 2008;371(9611):483–91. https://doi.org/10.1016/S0140-6736(08)60208-1.
15. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang H-Y, Demirkan A, Den Hertog HM, Do R, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–83. https://doi.org/10.1038/ng.2797.
16. Adams B, Jacocks L, Guo H. Higher bmi is linked to an increased risk of heart attacks in european adults: a mendelian randomisation study. BMC Cardiovasc Disord. 2020;20(1):258. https://doi.org/10.1186/s12872-020-01542-w.
17. Burgess S. Sample size and power calculations in mendelian randomization with a single instrumental variable and a binary outcome. Int J Epidemiol. 2014;43(3):922–9. https://doi.org/10.1093/ije/dyu005.
18. Stan Development Team. STAN: A C++ Library for Probability and Sampling, Version 2.2. 2014. http://mc-stan.org/.
19. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Found Trends Mach Learn. 2008;1:1–305.
20. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7(4):457–72. https://doi.org/10.1214/ss/1177011136.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.