

RESEARCH ARTICLE

Open Access



# Overlapping-sample Mendelian randomisation with multiple exposures: a Bayesian approach

Linyi Zou, Hui Guo\*  and Carlo Berzuini

## Abstract

**Background:** Mendelian randomization (MR) has been widely applied to causal inference in medical research. It uses genetic variants as instrumental variables (IVs) to investigate putative causal relationship between an exposure and an outcome. Traditional MR methods have mainly focussed on a two-sample setting in which IV-exposure association study and IV-outcome association study are independent. However, it is not uncommon that participants from the two studies fully overlap (one-sample) or partly overlap (overlapping-sample).

**Methods:** We proposed a Bayesian method that is applicable to all the three sample settings. In essence, we converted a two- or overlapping- sample MR to a one-sample MR where data were partly unmeasured. Assume that all study individuals were drawn from the same population and unmeasured data were missing at random. Then the missing data were treated au pair with the model parameters as unknown quantities, and thus, were imputed iteratively conditioning on the observed data and estimated parameters using Markov chain Monte Carlo. We generalised our model to allow for pleiotropy and multiple exposures and assessed its performance by a number of simulations using four metrics: mean, standard deviation, coverage and power. We also compared our method with classic MR methods.

**Results:** In our proposed method, higher sample overlapping rate and instrument strength led to more precise estimated causal effects with higher power. Pleiotropy had a notably negative impact on the estimates. Nevertheless, the coverages were high and our model performed well in all the sample settings overall. In comparison with classic MR, our method provided estimates with higher precision. When the true causal effects were non-zero, power of their estimates was consistently higher from our method. The performance of our method was similar to classic MR in terms of coverage.

**Conclusions:** Our model offers the flexibility of being applicable to any of the sample settings. It is an important addition to the MR literature which has restricted to one- or two- sample scenarios. Given the nature of Bayesian inference, it can be easily extended to more complex MR analysis in medical research.

**Keywords:** Mendelian randomization, Bayesian approach, Missing data, Overlapping-sample, Multiple exposures

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



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Many questions in medical research address putative causal nature of the relationship between a clinical outcome and a corresponding risk factor or exposure. Randomised controlled trials (RCTs) are ideal for this purpose, but are often infeasible due to cost or ethical considerations. With the aid of proper statistical methodology, causal inference can be made from observational studies. This is a context where Mendelian randomization (MR) [1–3] can play an important role. MR uses genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure on an outcome of interest [4, 5]. Without loss of generality, we assume the IVs are single nucleotide polymorphisms (SNPs).

MR analysis requires data from two association studies: IV-exposure and IV-outcome. Most of MR methods [6–9] have focussed on a two-sample scenario where the IV-exposure and IV-outcome associations were estimated separately from independent studies. In other words, there were no shared individuals between these two studies. A limited number of one-sample MR methods, if a single group of individuals come with a complete set of measurements (IVs, exposure and outcome), are also available in the literature [5, 10–12]. Little attention has, however, been devoted to overlapping samples where the two studies have a subset of individuals in common. For example, an investigator has identified a number of SNPs associated with diabetes from a genome-wide association study. Her interest is in whether expression levels of certain genes are causal risk factors of diabetes. But gene expression can be measured only from a subset of participants of the SNP-diabetes association study due to cost. Another example is that the investigator has collected data from two small independent IV-BMI and IV-diabetes association studies, with the aim to estimate the causal effect of BMI on diabetes using MR. To enhance statistical power, a natural way is to incorporate previous studies that have complete sets of observed data on IV, BMI and diabetes into her current studies.

To the best of our knowledge, few studies have investigated overlapping-sample MR. Burgess et al. [13] have shown that sample overlap increases type I error and leads to bias in classic MR methods. LeBlanc et al. [14] have developed a correction method of overlapping samples by decorrelation. There is a pressing need for the development of a flexible MR approach that can be used for one-, two- as well as overlapping- samples. Here we propose a novel MR method based on the Bayesian framework introduced by Berzuini et al [5]. Our method preserves data information while avoiding bias. It exploits two simple, but powerful, ideas: (i) the overlapping-sample problem is a special case of the more general situation where some or all the individuals have missing values in exposure or in

outcome; and (ii) a Bayesian approach can coherently deal with missing data by treating them as additional parameters which can be estimated from the data. In this paper we introduce our method and assess its performance by comparing it with classic MR in a simulation study. A further advantage of Bayesian MR is freedom to elaborate the model in various directions of interest. This feature we illustrate by moving away from the standard single-exposure MR model and considering multiple exposures instead.

## Methods

### Model

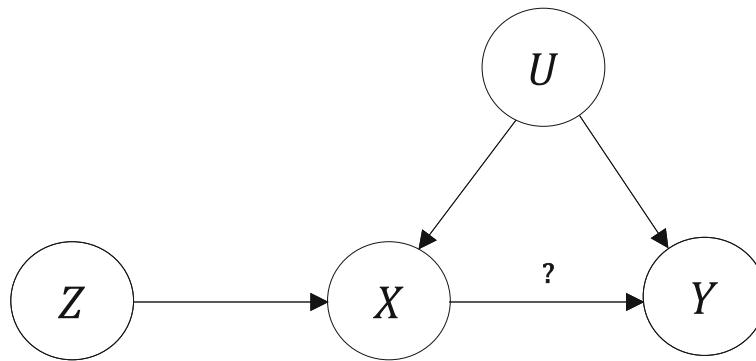
Let  $U$  denote a set of unobserved confounders which could possibly distort the causal relationship between the exposure  $X$  and the outcome  $Y$ . Let  $Z$  be an instrumental SNP (multiple instruments are considered later) which satisfies the following three assumptions ([10]):

- $A_1$ :  $Z$  is associated with  $X$ ;
- $A_2$ :  $Z$  is independent of  $U$ ;
- $A_3$ :  $Z$  is independent of  $Y$ , conditioning on  $(X, U)$ .

These assumptions are graphically expressed in Fig. 1. The  $Z \rightarrow X$  arrow represents a non-zero association between the IV and the exposure, in accord with  $A_1$ . Assumption  $A_2$  follows from the graph (it would not if there were an arrow directly connects  $Z$  and  $U$ ).  $A_3$  is also satisfied (it would not if an arrow pointed directly from  $Z$  to  $Y$ ). The  $X \rightarrow Y$  arrow represents the putative causal effect of  $X$  on  $Y$ , which is our primary interest.

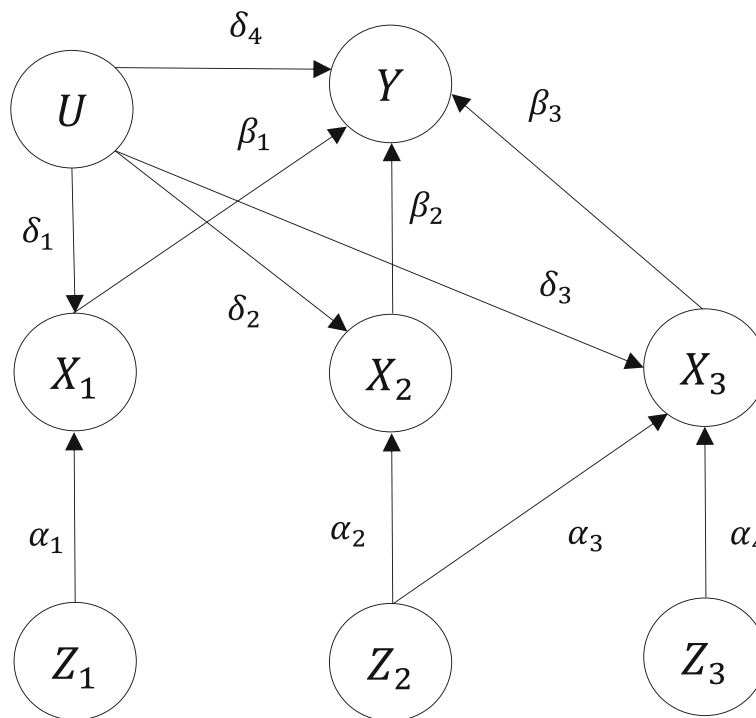
Consider a study with multiple exposures and each instrumental SNP associated with at least one exposure. Without loss of generality, we shall hereafter restrict attention to the case where there are three exposures  $\mathbf{X} = (X_1, X_2, X_3)$  and three IVs  $\mathbf{Z} = (Z_1, Z_2, Z_3)$ . Figure 2 depicts our model and parameter notations, with  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  representing the putative causal effects of the exposures  $X_1$ ,  $X_2$  and  $X_3$  on the outcome  $Y$ , respectively;  $\delta_1$ ,  $\delta_2$ ,  $\delta_3$  and  $\delta_4$  representing the effects of the confounder  $U$  on the three exposures and  $Y$ , respectively. For simplicity, we assume that there are no direct associations between the exposures and that the IVs are mutually independent. The  $Z_1 \rightarrow X_1$  association is quantified by parameter  $\alpha_1$  and the  $Z_2 \rightarrow X_2$  association by  $\alpha_2$ . Both  $Z_2$  and  $Z_3$  are associated with  $X_3$ , and the strengths of these associations are quantified by parameters  $\alpha_3$  and  $\alpha_4$  respectively.  $Z_1$  and  $Z_3$  are “valid instruments” as they satisfy all the above stated MR assumptions. However,  $Z_2$  violates assumption  $A_3$  because the effect of  $Z_2$  on  $Y$  is mediated by both  $X_2$  and  $X_3$  - a problem of horizontal pleiotropy [15].

By assuming linearity and additivity of the conditional dependencies, and in accord with the graph of Fig. 1, we specify our model as follows.



**Fig. 1** Schematic representation of the three assumptions required in Mendelian randomization. 1) Instrumental variable  $Z$  is associated with the exposure  $X$ ; 2)  $Z$  is independent of the confounder  $U$ ; 3)  $Z$  is independent of the outcome  $Y$ , conditioning on  $X$  and  $U$

- |                                                                                                           |     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $U \sim N(0, 0.1),$                                                                                       | (1) | where $N(a, b)$ stands for a normal distribution with mean $a$ and variance $b$ . The $\omega$ parameters are unknown intercepts and the $\sigma$ s are standard deviations of independent random noise terms. The parameters of primary inferential interest are the $\beta$ s, each representing the causal effect of a particular exposure on the outcome. The $\alpha$ parameters quantify the strengths of pairwise associations between the instruments and the exposures. They are often referred to as “instrument strengths”, and should |
| $X_1 Z_1, U \sim N(\omega_1 + \alpha_1 Z_1 + \delta_1 U, \sigma_1^2),$                                    | (2) |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| $X_2 Z_2, U \sim N(\omega_2 + \alpha_2 Z_2 + \delta_2 U, \sigma_2^2),$                                    | (3) |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| $X_3 Z_2, Z_3, U \sim N(\omega_3 + \alpha_3 Z_2 + \alpha_4 Z_3 + \delta_3 U, \sigma_3^2),$                | (4) |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| $Y X_1, X_2, X_3, U \sim N(\omega_Y + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \delta_4 U, \sigma_Y^2),$ | (5) |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |



**Fig. 2** Diagram of our model and parameter settings. There are three instrumental variables ( $Z_1, Z_2, Z_3$ ) and three exposures ( $X_1, X_2, X_3$ ).  $Z_1$  is associated with  $X_1$  only.  $Z_3$  is associated with  $X_3$  only.  $Z_2$  is associated with both  $X_2$  and  $X_3$ . For simplicity, we assume there is no direct associations among the exposures and the instrumental variables are mutually independent

be significantly different from zero to attenuate weak instrument bias [16]. Finally,  $U$  denotes a sufficient scalar summary of the unobserved confounders, which is set to be drawn from a  $N(0,0.1)$  distribution with its parameters not identifiable from the likelihood. This model is built on the basis of the Bayesian approach developed by Berzuini et al. [5], with an extension to multiple exposures and allowing for an IV to be associated with more than one exposure.

**Our method**

We propose a Bayesian MR method that works in all the sample settings (whether one-sample, two-sample or overlapping-sample), and therefore is more flexible than existing non-Bayesian methods in this respect. In essence, our method treats two- or overlapping- samples as one sample where the data of some (or all) of the individuals are incomplete, in the sense that these individuals have  $Z$  and  $X$  (but not  $Y$ ) measured or  $Z$  and  $Y$  (but not  $X$ ) measured. Clearly our method allows for quite general patterns of missingness, although it is not our purpose here to explicitly explore their full repertoire. In addition, for simplicity, we illustrate our method by restricting attention to the special case of a continuous outcome variable  $Y$ . We assume that all individuals are drawn from the same population and unmeasured data are missing at random. Then whatever are the unobserved variables  $X$  or  $Y$ , they are treated au pair with the model parameters as unknown quantities, and thus, can be imputed iteratively from their distributions conditioning on the observed variables and estimated parameters using Markov chain Monte Carlo (MCMC) [17] - the best Bayesian tradition.

We start with introducing three disjoint datasets as follows (see Fig. 3a):

- Dataset A: all individuals with observed values of  $Z$ ,  $X$  and  $Y$ ;
- Dataset B: all individuals with observed values of  $Z$  and  $X$  only;
- Dataset C: all individuals with observed values of  $Z$  and  $Y$  only.

No individuals are common in  $A$ ,  $B$  and  $C$  but they are drawn independently from the same population. As discussed earlier,  $A$  may be collected from a previous study;  $B$  and  $C$  may represent data from two independent IV-exposure and IV-outcome association studies. If we combine  $A$  with  $B$

$$D_1 = A \cup B$$

and  $A$  with  $C$

$$D_2 = A \cup C,$$

the two resulting datasets,  $D_1$  and  $D_2$ , will not be disjoint as they have dataset  $A$  in common- a typical overlapping-sample problem. If  $A$  is empty,  $B$  and  $C$  will form a two-sample problem.

Our method treats missing data as all other unknown quantities in the model. That is, we impute the missing values of  $Y$  in  $B$  by randomly drawing a value from the conditional distribution of  $Y$  given observed exposures and the estimated parameters at each iteration. Likewise, the missing values of  $X$  in  $C$  will be imputed by a randomly draw from the conditional distribution of  $X$  given the observed instruments  $Z$  and the estimated parameters at each iteration. Let  $Y^*$  be imputed values of  $Y$  and  $X^* = (X_1^*, X_2^*, X_3^*)$  be imputed values of  $X$ . Given all the datasets  $A, B$  and  $C$ , we proceed with the following Markov chain scheme.

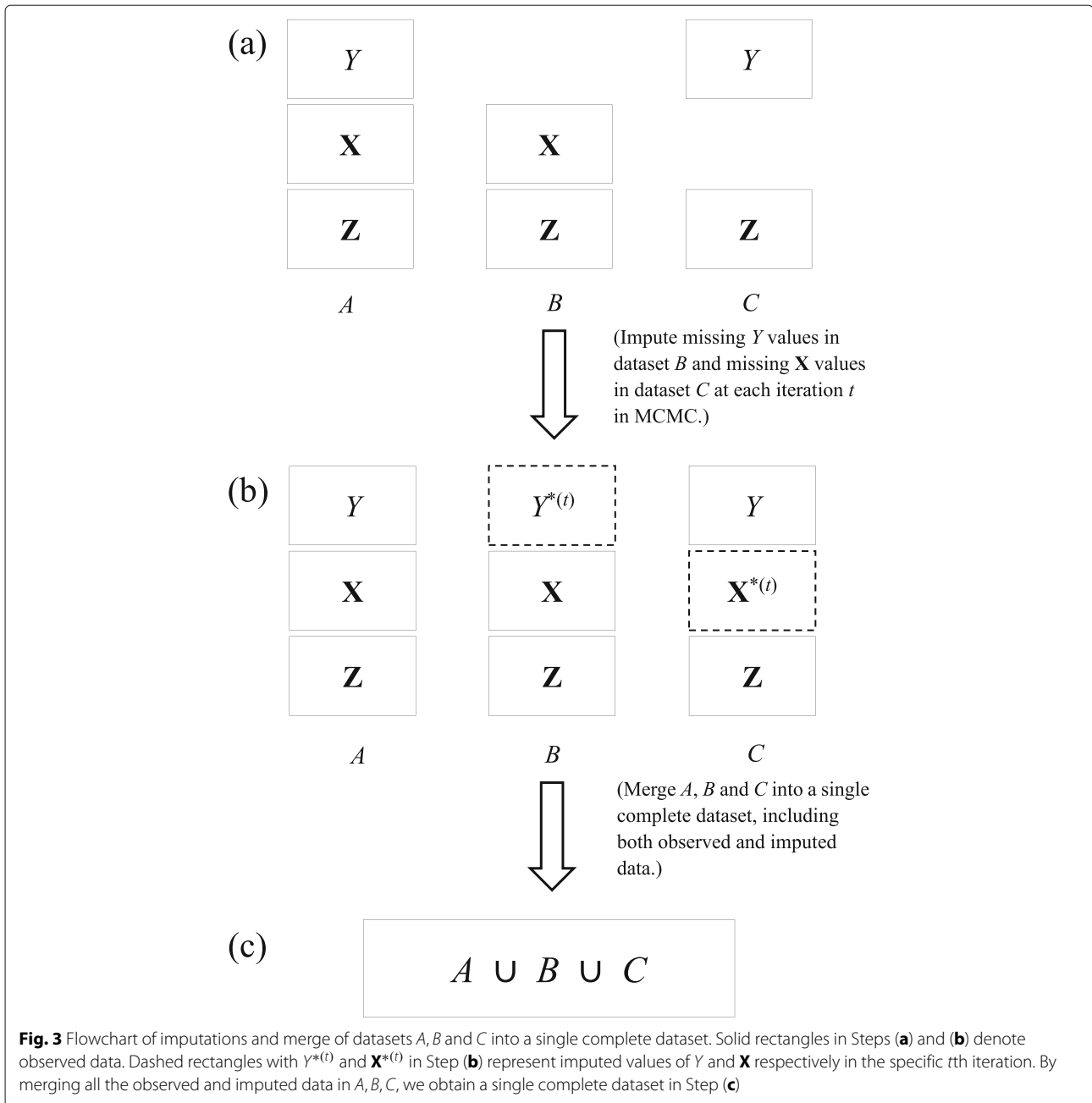
1. Set initial values for all the unknown parameters in the model. We also fix the desired number of Markov iterations, with iteration index  $t$ .
2. At the  $t$ th iteration, in accord with Fig. 3b, we replace the missing values of  $Y$  in  $B$  with  $Y^*$  which is drawn from a Normal distribution with mean  $\omega_Y^{(t)} + \beta_1^{(t)} X_1 + \beta_2^{(t)} X_2 + \beta_3^{(t)} X_3$  and standard deviation  $\sigma_Y^{(t)}$ , where  $(X_1, X_2, X_3)$  are observed values in dataset  $B$ . Similarly, we replace the missing values of  $X$  in  $C$  with  $(X_1^*, X_2^*, X_3^*)$  which are drawn from a Normal distribution with mean  $\omega_1^{(t)} + \alpha_1^{(t)} Z_1$ ,  $\omega_2^{(t)} + \alpha_2^{(t)} Z_2$  and  $\omega_3^{(t)} + \alpha_3^{(t)} Z_2 + \alpha_4^{(t)} Z_3$  respectively, where  $(Z_1, Z_2, Z_3)$  are observed values in dataset  $C$ . We use superscript  $(t)$  for the random parameters at the  $t$ th iteration, with  $t = 0$  representing their initial values. Because both  $U$  and the random errors have zero mean, they have vanished in this step.
3. Merge all the data, imputed and observed, into a new complete dataset (Fig. 3c).
4. Estimate model parameters based on the complete dataset obtained in Step 3 using MCMC and set  $t \leftarrow t + 1$ .
5. Repeat Steps 2-4 until  $t$  equals the number of iterations specified in Step 1.

**The priors**

Here we discuss our choice of prior distributions of the unknown parameters involved in Models (1)-(5). Let the priors for  $\beta = (\beta_1, \beta_2, \beta_3)$  be independently normally distributed with mean 0 and standard deviation 10

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^2 & 0 & 0 \\ 0 & 10^2 & 0 \\ 0 & 0 & 10^2 \end{pmatrix} \right],$$

and the IV strength parameters  $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$  be independently normally distributed with mean 1 and standard deviation 0.3



$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} \sim N \left[ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.3^2 & 0 & 0 & 0 \\ 0 & 0.3^2 & 0 & 0 \\ 0 & 0 & 0.3^2 & 0 \\ 0 & 0 & 0 & 0.3^2 \end{pmatrix} \right].$$

$U$  has been assumed to follow a normal distribution  $U \sim N(0, 0.1)$  in Model (1). The priors of the standard deviation parameters  $\sigma_1, \sigma_2, \sigma_3$  and  $\sigma_Y$  are assumed to follow a same inverse-gamma distribution  $\sigma(\cdot) \sim Inv\text{-Gamma}(3, 2)$ .

### Simulations

In our simulations, according to Models (1)-(5), we consider a total of 72 configurations including

- 6 sample overlapping rates (100%, 80%, 60%, 40%, 20%, 0%)
- 2 IV strengths:  $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \mathbf{0.5}$  and  $\mathbf{0.1}$
- 3 levels of the confounding effects of  $U$  on  $X$ :  $(\delta_1, \delta_2, \delta_3) = \mathbf{1, 0.5}$  and  $\mathbf{0.1}$
- 2 levels of causal effects of  $X$  on  $Y$ :  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0.3}$  and  $\mathbf{0}$ .

The effect of  $U$  on  $Y$  is set to 1 ( $\delta_4 = 1$ ). In each configuration, we simulated 200 datasets, step by step, as follows.

1. Generate a dataset  $\mathbf{H}$  which contains observed IVs ( $\mathbf{Z}$ ), exposures ( $\mathbf{X}$ ) and outcome ( $Y$ ) from 1000 independent individuals.
2. Randomly sample  $n_A$  individuals without replacement from  $\mathbf{H}$  and take their observations of ( $\mathbf{Z}, \mathbf{X}, Y$ ) as dataset  $A$ ;
3. Randomly Sample  $n_B$  individuals without replacement from  $\mathbf{H} - A = \{x \in \mathbf{H}, x \notin A\}$  and take their observations of ( $\mathbf{Z}, \mathbf{X}$ ) as dataset  $B$ ;
4. Randomly sample  $n_C$  individuals without replacement from  $\mathbf{H} - A - B = \{x \in \mathbf{H}, x \notin A \cup B\}$  and take their observations of ( $\mathbf{Z}, Y$ ) as dataset  $C$ .

The sample size of  $B$  was set to be the same as that of  $C$  (i.e.,  $n_B = n_C$ ) and the sample size of both  $D_1 = A \cup B$  and  $D_2 = A \cup C$  to 400. The overlapping rate was defined as the percentage of the number of individuals from  $A$  ( $n_A$ ) in  $D_1$  (or equivalently, in  $D_2$ ). For example, if overlapping rate was 80%, then  $n_A = 320$  and  $n_B = n_C = 80$ . Similarly, for an overlapping rate of 60%,  $n_A = 240$  and  $n_B = n_C = 160$ . When overlapping rate was 100%, we only used dataset  $A$  of sample size 400 (or equivalently,  $D_1 = D_2 = A$  where  $n_A = 400$ ). For an overlapping rate of 0%, we only used datasets  $B$  and  $C$  (i.e.,  $D_1 = B$  with  $n_B = 400$  and  $D_2 = C$  with  $n_C = 400$ ). Imputations of missing data and causal effect estimations were then performed simultaneously using MCMC in Stan [18, 19].

#### Assessments of the performance of our method

The performance of our method was evaluated using 4 metrics:

- mean (posterior mean)
- standard deviation
- coverage (the proportion of the times that the 95% credible interval contained the true value of the causal effect)
- power (the proportion of the times that the 95% credible interval did not contain value zero when the true causal effect was non-zero).

The causal effects of  $X_1$ ,  $X_2$  and  $X_3$  on the outcome  $Y$  (denoted as  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ ) were estimated from data simulated under the alternative hypothesis that  $\beta = \mathbf{0.3}$ , and from data simulated under the null hypothesis that  $\beta = \mathbf{0}$ . Note that the metric power was only applicable under the alternative hypothesis when  $\beta \neq \mathbf{0}$ , by its definition. A high value of power indicates high sensitivity of the model to deviations of the data from the null or, equivalently, a low expected probability of false negatives.

We compared our method with the two-stage least squares (2SLS) regression [10] for one-sample MR (100% overlap) and with the inverse-variance weighted (IVW) estimation [8] for two-sample MR (0% overlap). For overlapping samples, our method was compared with IVW, and in the latter, we only included data of non-overlapping samples to make it a two-sample MR.

#### Results

Table 1 displays a summary of results obtained from the simulated data under the alternative hypothesis ( $\beta = \mathbf{0.3}$ ). Each row of the table corresponds to a configuration of specified sample overlapping rate, instrument strength  $\alpha$  and degree of confounding  $\delta$ . Columns are values of the four metrics of the estimated causal effects  $\hat{\beta}$ s obtained from our method and classic MR. Let us first focus on our method (grey columns “Bayesian”). If overlapping rate was high (80% or over), both the coverage and the power of the  $\hat{\beta}$ s were consistently high ( $\geq 0.93$ ). When the sample overlap was 60% or under, strong IVs ( $\alpha = \mathbf{0.5}$ ) resulted in high statistical power (=1) while the impact of reduced IV strength ( $\alpha = \mathbf{0.1}$ ) became detrimental, even more so when overlapping rate was reduced. When IVs were relatively weak ( $\alpha = \mathbf{0.1}$ ), power increased as confounding effects  $\delta$  decreased. Overall, where there was pleiotropy (in estimating  $\beta_2$  because there were two competing paths from  $Z_2$  to  $Y$  via  $X_2$  and  $X_3$ ) the results were notably worse than those without pleiotropy (in estimating  $\beta_1$  and  $\beta_3$ ). As  $\alpha$  increased, the averages of  $\hat{\beta}$ s (mean) became closer to the true value 0.3 and the variations (sd) decreased, concluding that higher IV strength reduced bias and resulted in more precise estimates. We note that despite the greater posterior uncertainty for  $\beta_2$  across all the configurations (indicated in the “sd” column), the posterior mean for this parameter remained relatively close to the true value. This is evidence that our model is reliable and “honest about uncertainty”. Compared with the classic MR, our method showed higher precision in estimated causal effect in all configurations (“sd” columns). When IV strength became weaker ( $\alpha = \mathbf{0.1}$ ), our method led to higher power throughout. With highly overlapped samples (80%), even for strong IVs, power of the estimates in classic MR (IVW) was relatively low, partly due to smaller sample size after removing data of overlapping samples.

Table 2 shows the results of the simulated data under the null ( $\beta = \mathbf{0}$ ). Again, we first report results from our method (grey columns “Bayesian”). The coverages of the three estimated causal effects  $\hat{\beta}$ s were high ( $> 0.9$ ) in all the configurations, indicating that the method does not produce a large number of false positives - an important property of MR analysis. Therefore, none of the sample overlapping rate, IV strength and confounding effect has had much negative influence on the results, although there seemed to be an increasing trend in bias and varia-

**Table 1** Causal effects estimated from simulated data using our Bayesian method and classic methods (2SLS when sample overlap = 100% and IVW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0.3}$

overlap	$\alpha$	$\delta$	$\hat{\beta}_1$						$\hat{\beta}_2$						$\hat{\beta}_3$												
			Bayesian			Classic			Bayesian			Classic			Bayesian			Classic									
			mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power					
100%	0.5	1	0.300	0.001	0.96	1	0.301	0.020	0.96	1	0.300	0.002	0.97	1	0.298	0.025	0.99	1	0.300	0.001	0.99	1	0.301	0.019	0.99	1	
			0.301	0.006	0.96	1	0.300	0.017	0.98	1	0.301	0.004	0.99	1	0.302	0.024	0.97	1	0.300	0.002	0.99	1	0.299	0.018	0.97	1	
			0.301	0.006	0.96	1	0.300	0.015	0.95	1	0.301	0.009	0.98	1	0.300	0.022	0.97	1	0.300	0.006	0.95	1	0.300	0.016	0.96	1	
	0.1	1	0.302	0.006	0.97	1	0.286	0.100	0.98	0.72	0.303	0.010	0.97	1	0.309	0.141	0.98	0.47	0.300	0.006	0.98	1	0.292	0.106	0.97	0.76	
			0.306	0.009	0.95	1	0.302	0.087	0.96	0.92	0.309	0.015	0.98	1	0.305	0.130	0.96	0.65	0.301	0.009	0.99	1	0.294	0.092	0.98	0.86	
			0.319	0.029	0.95	1	0.301	0.071	0.97	0.97	0.324	0.050	0.94	1	0.304	0.104	0.97	0.79	0.300	0.029	0.98	1	0.296	0.079	0.95	0.96	
	80%	0.5	1	0.304	0.020	0.93	1	0.297	0.108	0.95	0.78	0.304	0.022	0.95	1	0.311	0.117	0.99	0.66	0.300	0.013	0.97	1	0.294	0.100	0.96	0.83
				0.302	0.010	0.96	1	0.302	0.099	0.97	0.87	0.302	0.015	0.96	1	0.300	0.101	1	0.71	0.299	0.011	0.96	1	0.306	0.098	0.97	0.89
		0.1	1	0.302	0.011	0.95	1	0.307	0.098	0.92	0.91	0.300	0.016	0.97	1	0.311	0.092	0.97	0.85	0.301	0.011	0.96	1	0.298	0.093	0.95	0.90
				0.306	0.048	0.98	1	0.319	0.393	0.96	0.11	0.307	0.072	0.98	0.98	0.317	0.577	0.96	0.11	0.302	0.060	0.97	1	0.284	0.406	0.94	0.12
60%	0.5	1	0.320	0.041	0.98	1	0.276	0.335	0.95	0.17	0.320	0.060	0.98	1	0.301	0.461	0.97	0.12	0.303	0.047	1	1	0.320	0.355	0.96	0.15	
			0.320	0.050	0.97	1	0.296	0.296	0.94	0.22	0.320	0.075	0.97	1	0.310	0.437	0.93	0.14	0.301	0.052	0.98	1	0.295	0.301	0.94	0.18	
	0.1	1	0.302	0.020	0.95	1	0.295	0.078	0.94	0.97	0.299	0.027	0.95	1	0.310	0.081	0.99	0.93	0.302	0.018	0.94	1	0.295	0.073	0.94	0.99	
			0.299	0.014	0.97	1	0.305	0.068	0.95	1	0.301	0.018	0.96	1	0.303	0.070	1	0.99	0.301	0.013	0.96	1	0.298	0.069	0.96	1	
	0.1	1	0.300	0.012	0.98	1	0.304	0.060	0.96	1	0.303	0.018	0.96	1	0.301	0.070	0.98	0.98	0.298	0.012	0.95	1	0.298	0.065	0.96	1	
			0.283	0.066	0.96	0.98	0.304	0.283	0.94	0.22	0.285	0.106	0.95	0.77	0.336	0.376	0.96	0.14	0.307	0.086	0.95	0.97	0.299	0.298	0.95	0.20	
0.5	1	0.300	0.057	0.97	1	0.304	0.236	0.92	0.27	0.307	0.085	0.97	0.93	0.324	0.319	0.95	0.17	0.304	0.066	0.98	1	0.287	0.229	0.96	0.26		
		0.301	0.061	0.97	1	0.306	0.203	0.94	0.41	0.303	0.085	0.97	0.95	0.318	0.246	0.97	0.20	0.300	0.059	0.96	1	0.280	0.183	0.94	0.32		

**Table 1** Causal effects estimated from simulated data using our Bayesian method and classic methods (2SLS when sample overlap = 100% and IVW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0.3}$  (Continued)

overlap	$\alpha$	$\delta$	$\hat{\beta}_1$						$\hat{\beta}_2$						$\hat{\beta}_3$											
			Bayesian			Classic			Bayesian			Classic			Bayesian			Classic								
			mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power	mean	sd	coverage	power				
40%	0.5	0.1	0.300	0.025	0.91	0.299	0.060	0.96	1	0.298	0.033	0.91	1	0.304	0.059	1	0.301	0.021	0.93	1	0.298	0.055	0.96	1		
			0.301	0.017	0.93	1	0.297	0.058	0.94	1	0.300	0.023	0.95	1	0.303	0.055	0.99	1	0.298	0.015	0.94	1	0.297	0.057	0.94	1
		0.300	0.013	0.97	1	0.300	0.049	0.97	1	0.300	0.019	0.95	1	0.308	0.055	0.99	1	0.300	0.014	0.95	1	0.296	0.058	0.93	1	
		0.270	0.085	0.94	0.90	0.265	0.200	0.97	0.24	0.277	0.143	0.97	0.57	0.298	0.323	0.94	0.18	0.301	0.111	0.92	0.82	0.291	0.237	0.95	0.29	
	0.5	0.2	0.292	0.073	0.95	0.99	0.299	0.174	0.96	0.38	0.296	0.112	0.94	0.74	0.321	0.276	0.95	0.27	0.298	0.085	0.94	0.95	0.273	0.192	0.95	0.37
			0.301	0.063	0.98	1	0.313	0.154	0.97	0.55	0.296	0.097	0.98	0.84	0.302	0.213	0.94	0.27	0.301	0.067	0.97	1	0.307	0.155	0.96	0.51
		0.306	0.028	0.86	1	0.304	0.045	0.99	1	0.297	0.040	0.86	1	0.301	0.058	1	1	0.301	0.025	0.91	1	0.301	0.051	0.95	1	
		0.301	0.018	0.96	1	0.298	0.050	0.95	1	0.301	0.029	0.93	1	0.304	0.050	0.99	1	0.299	0.021	0.93	1	0.299	0.044	0.97	1	
	0.1	0.3	0.302	0.015	0.96	1	0.304	0.041	0.96	1	0.300	0.021	0.94	1	0.298	0.047	0.99	1	0.299	0.016	0.92	1	0.304	0.047	0.95	1
			0.281	0.097	0.93	0.82	0.267	0.171	0.96	0.28	0.273	0.147	0.95	0.49	0.298	0.235	0.98	0.21	0.286	0.124	0.93	0.74	0.308	0.177	0.97	0.41
		0.283	0.092	0.94	0.86	0.294	0.153	0.95	0.47	0.291	0.127	0.97	0.63	0.276	0.201	0.97	0.24	0.292	0.097	0.92	0.89	0.306	0.151	0.95	0.49	
		0.316	0.073	0.95	0.99	0.304	0.132	0.97	0.64	0.308	0.099	0.99	0.82	0.307	0.178	0.96	0.38	0.308	0.071	0.97	0.99	0.290	0.130	0.97	0.57	
0%	0.5	0.303	0.028	0.88	1	0.299	0.049	0.95	1	0.295	0.037	0.91	1	0.296	0.055	0.97	1	0.302	0.028	0.87	1	0.303	0.050	0.93	1	
		0.304	0.021	0.93	1	0.301	0.042	0.97	1	0.302	0.030	0.91	1	0.300	0.047	0.98	1	0.300	0.020	0.94	1	0.308	0.043	0.95	1	
	0.302	0.015	0.97	1	0.301	0.039	0.97	1	0.301	0.025	0.92	1	0.299	0.041	1	1	0.300	0.015	0.96	1	0.308	0.041	0.95	1		
	0.323	0.135	0.93	0.74	0.289	0.158	0.98	0.41	0.289	0.188	0.91	0.43	0.294	0.280	0.89	0.31	0.309	0.140	0.90	0.74	0.311	0.194	0.91	0.46		
0.5	0.3	0.319	0.110	0.90	0.91	0.290	0.135	0.96	0.58	0.303	0.153	0.91	0.62	0.293	0.229	0.89	0.35	0.308	0.110	0.90	0.85	0.309	0.162	0.91	0.62	
		0.307	0.081	0.95	0.97	0.316	0.120	0.95	0.74	0.286	0.108	0.93	0.75	0.312	0.184	0.92	0.45	0.307	0.082	0.94	0.95	0.299	0.135	0.89	0.73	

Mean, standard deviation (sd), coverage and power are displayed for the estimated causal effects of the exposures  $X_1, X_2$  and  $X_3$  on the outcome  $Y$ , denoted as  $\hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\beta}_3$  respectively. There are 36 configurations containing six sample overlapping rates (100%, 80%, 60%, 40%, 20%, 0%), two levels of IV strengths ( $\alpha = \mathbf{0.5}$  and  $\mathbf{0.1}$ ) and three levels of effects of the confounder  $U$  on the exposures ( $\delta = (\delta_1, \delta_2, \delta_3) = \mathbf{1.05}$  and  $\mathbf{0.1}$ ). The effect of  $U$  on  $Y$  ( $\delta_4$ ) is set to 1



**Table 2** Causal effects estimated from simulated data using our Bayesian method and classic methods (2SLS when sample overlap = 100% and IWW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0}$

overlap	$\alpha$	$\delta$	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$												
			Bayesian		Classic	Bayesian		Classic	Bayesian		Classic										
			mean	sd	coverage	mean	sd	coverage	mean	sd	coverage	mean	sd	coverage							
100%	0.5	1	0.000	0.001	0.98	0.000	0.015	0.94	0.000	0.002	0.99	-0.002	0.020	0.95	0.000	0.001	0.99	0.001	0.014	0.97	
		0.5	0.000	0.003	0.98	0.000	0.014	0.97	0.001	0.005	0.96	0.002	0.020	0.95	0.000	0.003	0.99	-0.001	0.014	0.95	
	0.1	1	0.001	0.006	0.97	0.000	0.014	0.95	0.001	0.010	0.94	0.000	0.021	0.96	0.001	0.006	0.99	0.000	0.015	0.96	
		0.5	0.002	0.006	0.98	-0.009	0.072	0.95	0.003	0.010	0.98	0.005	0.105	0.95	0.000	0.006	0.98	-0.005	0.074	0.94	
	80%	0.5	1	0.007	0.009	0.94	0.004	0.073	0.95	0.011	0.015	0.93	0.000	0.114	0.91	-0.001	0.009	0.99	-0.002	0.076	0.95
			0.1	0.019	0.029	0.95	-0.001	0.069	0.97	0.024	0.050	0.95	0.004	0.101	0.94	0.000	0.029	0.98	-0.003	0.075	0.94
60%	0.5	1	0.005	0.012	0.97	-0.001	0.047	0.97	0.005	0.011	0.98	0.000	0.063	0.95	0.001	0.006	0.99	-0.002	0.044	0.95	
		0.5	0.003	0.010	0.98	0.003	0.048	0.94	0.004	0.011	0.98	-0.002	0.067	0.95	0.000	0.007	0.97	0.002	0.050	0.95	
	0.1	1	0.001	0.009	0.98	0.004	0.047	0.96	0.002	0.016	0.97	0.003	0.060	0.97	0.000	0.010	0.98	0.000	0.043	0.97	
		0.5	0.028	0.026	0.95	-0.008	0.248	0.96	0.027	0.042	0.96	0.002	0.372	0.96	0.002	0.033	0.99	-0.018	0.263	0.94	
	0.1	1	0.029	0.030	0.94	-0.035	0.261	0.94	0.032	0.049	0.95	0.000	0.336	0.96	0.002	0.036	0.98	0.011	0.246	0.96	
		0.5	0.024	0.049	0.97	-0.002	0.254	0.94	0.025	0.077	0.96	0.003	0.376	0.92	0.001	0.049	1	-0.020	0.249	0.96	
60%	0.5	1	0.004	0.011	0.97	0.001	0.034	0.95	0.004	0.015	0.98	0.005	0.045	0.96	0.000	0.009	0.98	-0.003	0.035	0.94	
		0.5	0.002	0.010	0.96	0.003	0.034	0.96	0.002	0.014	0.98	0.000	0.043	0.96	0.000	0.010	0.99	0.001	0.033	0.95	
	0.1	1	0.000	0.010	0.99	0.001	0.036	0.91	-0.001	0.017	0.97	-0.002	0.048	0.94	0.001	0.012	0.97	0.000	0.030	0.97	
		0.5	0.034	0.037	0.92	-0.008	0.160	0.96	0.032	0.057	0.97	0.019	0.258	0.93	-0.001	0.045	0.99	-0.015	0.185	0.97	
	0.1	1	0.033	0.039	0.95	-0.003	0.173	0.94	0.030	0.061	0.96	0.002	0.252	0.93	-0.001	0.049	0.97	-0.009	0.169	0.95	
		0.5	0.010	0.061	0.95	0.002	0.179	0.94	0.009	0.089	0.94	0.021	0.223	0.96	0.002	0.059	0.95	-0.022	0.162	0.95	

**Table 2** Causal effects estimated from simulated data using our Bayesian method and classic methods (2SLS when sample overlap = 100% and IVW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0}$  (Continued)

overlap	$\alpha$	$\delta$	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$											
			Bayesian		Classic		Bayesian		Classic		Bayesian		Classic							
			mean	sd	coverage	sd	coverage	sd	coverage	sd	coverage	sd	coverage	sd	coverage					
40%	0.5	1	0.001	0.014	0.96	0.000	0.026	0.95	0.003	0.017	0.97	0.001	0.036	0.97	0.000	0.011	0.98	-0.001	0.025	0.96
		0.5	0.002	0.011	0.97	-0.003	0.029	0.93	0.002	0.017	0.97	0.001	0.036	0.97	0.000	0.012	0.95	0.001	0.026	0.95
		0.1	0.001	0.013	0.99	0.000	0.027	0.95	0.000	0.019	0.95	0.006	0.036	0.97	0.000	0.012	0.96	-0.004	0.028	0.94
	0.1	1	0.013	0.045	0.96	-0.021	0.135	0.95	0.019	0.077	0.96	-0.010	0.204	0.92	-0.001	0.060	0.96	-0.003	0.152	0.97
		0.5	0.013	0.051	0.98	0.002	0.131	0.96	0.013	0.078	0.97	0.009	0.215	0.94	-0.001	0.059	0.96	-0.016	0.150	0.95
		0.1	0.005	0.064	0.95	0.008	0.130	0.97	-0.002	0.082	0.98	-0.004	0.190	0.94	0.000	0.061	0.97	0.003	0.141	0.93
20%	0.5	1	0.002	0.014	0.95	0.002	0.021	0.96	0.004	0.019	0.95	0.001	0.033	0.94	-0.001	0.012	0.99	0.000	0.023	0.96
		0.5	-0.001	0.015	0.92	-0.002	0.023	0.95	0.001	0.020	0.97	0.002	0.032	0.98	0.001	0.015	0.95	0.000	0.023	0.96
		0.1	0.000	0.014	0.98	0.001	0.022	0.98	0.001	0.019	0.97	-0.002	0.033	0.94	-0.001	0.014	0.96	0.002	0.023	0.95
	0.1	1	0.012	0.057	0.96	-0.018	0.106	0.97	-0.002	0.083	0.96	0.000	0.158	0.97	0.002	0.064	0.96	-0.001	0.116	0.97
		0.5	0.010	0.061	0.98	-0.004	0.115	0.95	0.007	0.091	0.97	-0.010	0.157	0.95	0.000	0.068	0.96	0.001	0.114	0.96
		0.1	0.013	0.069	0.95	0.007	0.111	0.97	0.005	0.092	0.98	0.003	0.160	0.97	0.007	0.066	0.98	-0.006	0.117	0.94
0%	0.5	1	0.003	0.014	0.96	0.001	0.020	0.97	0.001	0.020	0.95	0.000	0.028	0.95	0.000	0.014	0.96	0.002	0.021	0.94
		0.5	0.002	0.013	0.97	-0.001	0.021	0.96	0.001	0.023	0.93	-0.002	0.029	0.95	0.000	0.014	0.96	0.000	0.021	0.94
		0.1	0.001	0.015	0.96	0.001	0.022	0.91	0.000	0.022	0.94	-0.002	0.029	0.95	0.001	0.014	0.95	0.001	0.022	0.93
	0.1	1	0.011	0.080	0.92	-0.005	0.100	0.97	-0.008	0.109	0.93	0.003	0.158	0.93	0.008	0.075	0.96	0.000	0.108	0.96
		0.5	0.009	0.073	0.95	-0.003	0.096	0.96	-0.005	0.102	0.94	-0.003	0.169	0.92	0.005	0.071	0.98	0.006	0.121	0.91
		0.1	0.010	0.072	0.96	-0.004	0.106	0.96	-0.003	0.105	0.97	-0.012	0.149	0.95	0.000	0.077	0.96	0.001	0.106	0.94

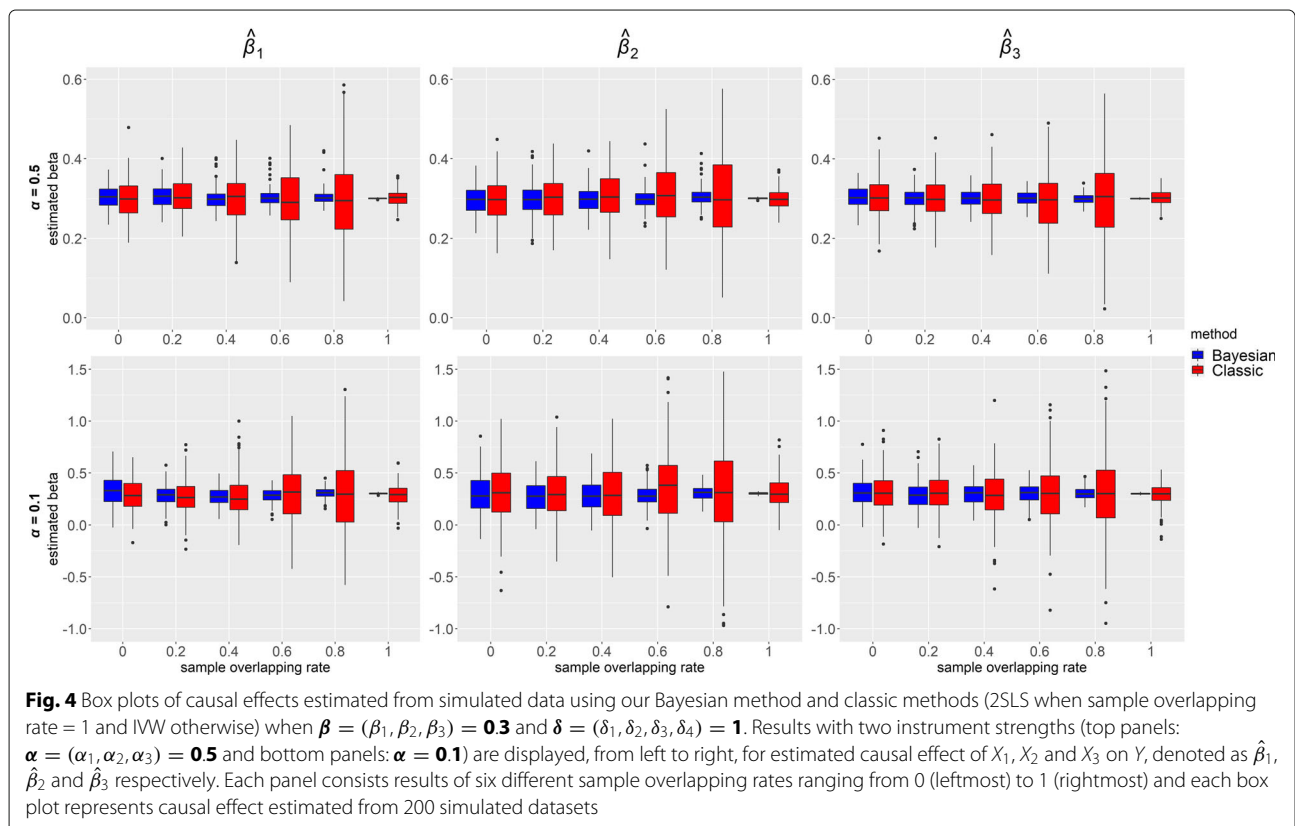
Mean, standard deviation (sd) and coverage are displayed for the estimated causal effects of the exposures  $X_1, X_2$  and  $X_3$  on the outcome  $Y$ , denoted as  $\hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\beta}_3$  respectively. There are 36 configurations containing six sample overlapping rates (100%, 80%, 60%, 40%, 20%, 0%), two levels of IV strengths ( $\alpha = (\alpha_1, \alpha_2, \alpha_3) = \mathbf{0.5}$  and  $\mathbf{0.1}$ ) and three levels of effects of the confounder  $U$  on the exposures ( $\delta = (\delta_1, \delta_2, \delta_3) = \mathbf{1.05}$  and  $\mathbf{0.1}$ ). The effect of  $U$  on  $Y$  ( $\delta_4$ ) is set to 1

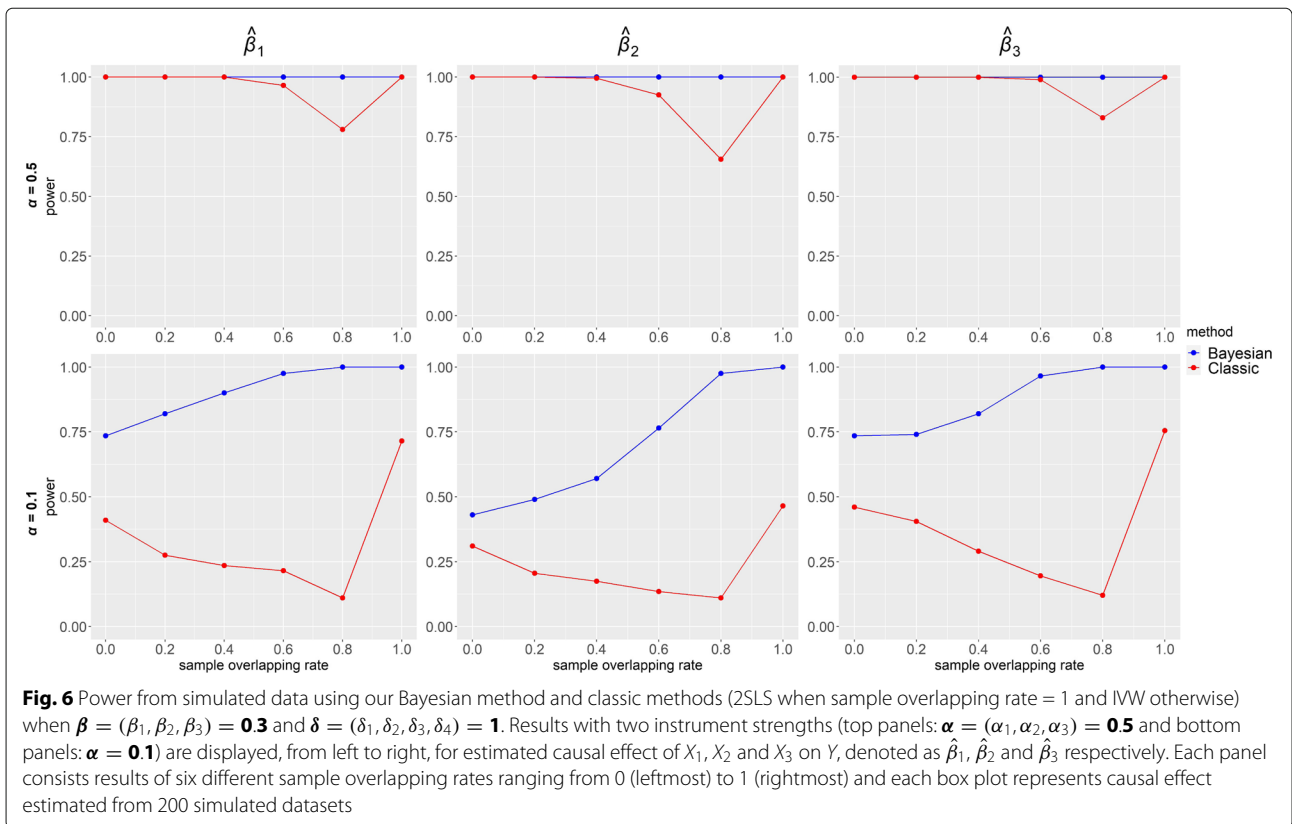
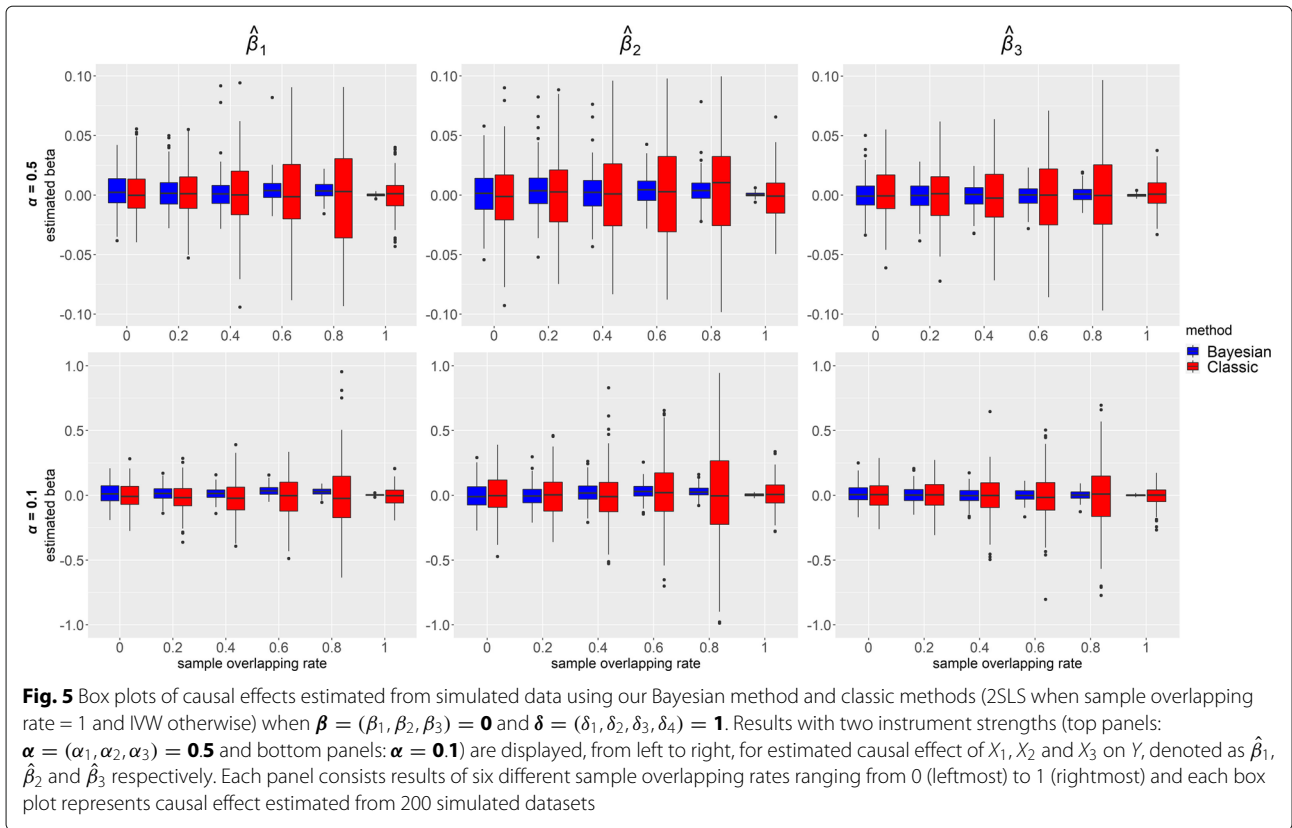
tion when IV strength decreased. In comparison with classic methods, our method provided estimates with higher precision (smaller values of “sd”), but similar coverage.

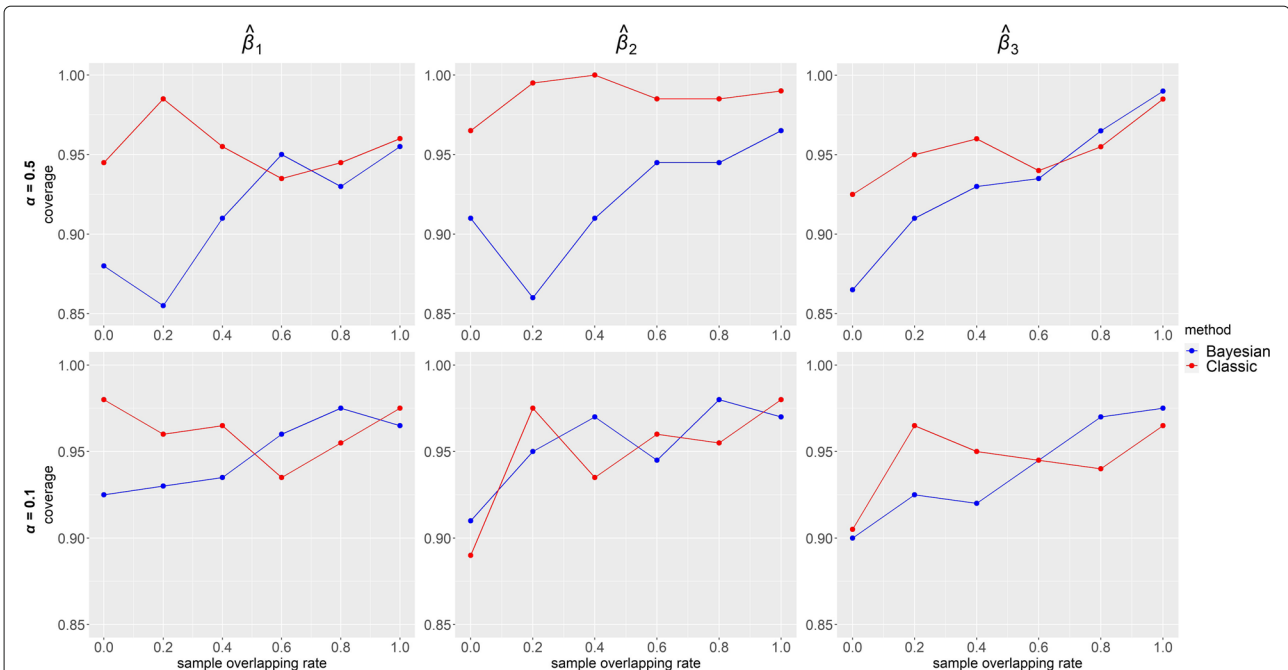
Figures 4 and 5 depict, respectively, distributions of  $\hat{\beta}$ s for each of the IV-strength and overlapping rate configurations when the confounding effects  $\delta = 1$ , for  $\beta = 0.3$  and  $\beta = 0$ . Each box represents distribution of the estimated causal effect  $\beta$  based on 200 simulated datasets. Higher overlapping rate led to a gain in precision from our Bayesian method, which can be observed in all panels across the figures as the blue boxes became narrower in height as the sample overlap increased and reached minimum in the one-sample scenario (overlapping rate: 1). It is also shown that estimates were more precise when instruments were stronger (top panels vs bottom panels). Blue boxes were narrower than red boxes in height in all the panels, indicating that our method consistently outperformed classic MR in precision. Interestingly, in classic methods, variations of the estimates in one sample were smaller than those in two- and overlapping- samples, which is probably because no data were excluded in the one-sample 2SLS MR. For two- and overlapping- samples, we adopted two-sample IVW which used summary statistics and required data removal of overlapping samples. Consequently, the sample size in IVW was reduced

to some extent for overlapping-samples, and the higher the overlap rate, the lower the precision.

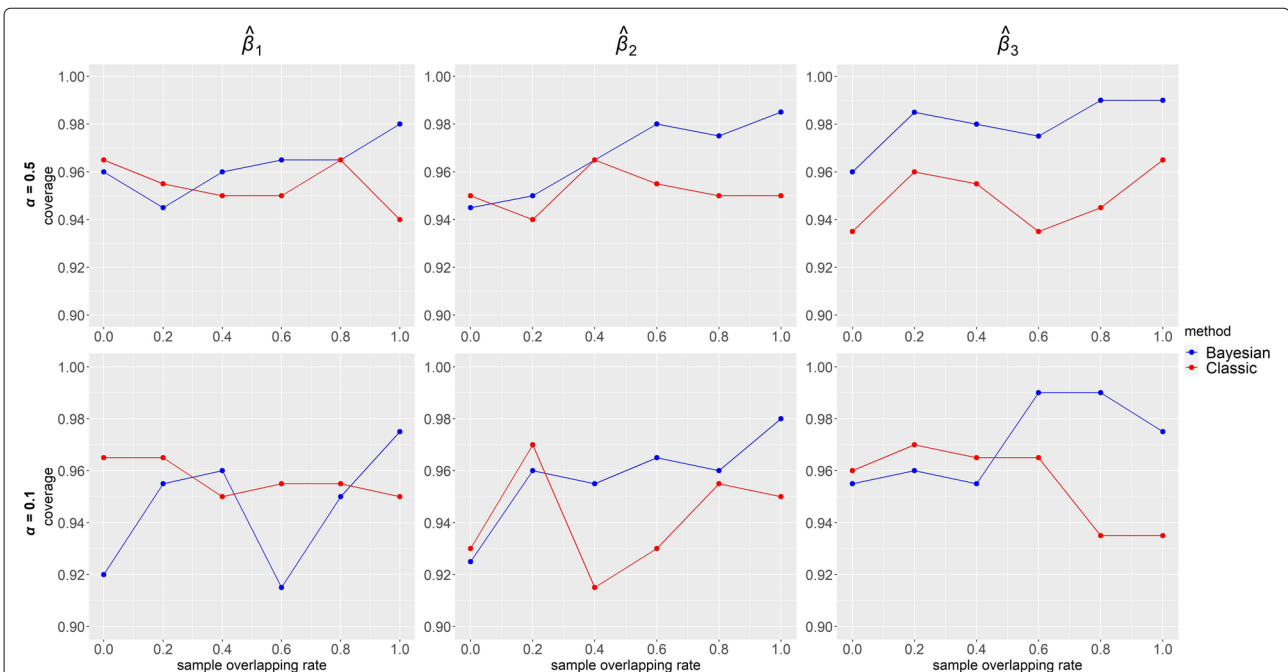
Power of  $\hat{\beta}$  for each IV-strength and overlapping rate configurations when the confounding effects  $\delta = 1$  under the alternative  $\beta = 0.3$  is presented in Fig. 6. In Bayesian method, power was always equal to 1 for strong IVs (top panels) and gradually became lower as overlapping rate decreased for weak IVs (bottom panels). Compared with our method, estimates in classic MR had lower power, especially when IVs were weak. In classic MR, similar to the trend of precision in Fig. 4, power was the highest for one sample and lowest for overlapping samples with highest overlap 80%. Figure 7 displays the coverage of  $\hat{\beta}$  for the same configurations and confounding effects as in Fig. 6 under the alternative  $\beta = 0.3$ . Overall, coverage was high ( $>0.85$ ) in both of the MR methods. Bayesian method performed better for high overlapping rate while classic MR was better for low overlapping rate. Under the alternative hypothesis, however, it would be sensible to focus on power rather than coverage. When  $\beta = 0$ , it is seen from Fig. 8 that the coverage of  $\hat{\beta}$  was high ( $>0.91$ ) in both Bayesian and classic MR. Our method consistently outperformed classic MR in power but not in coverage. This is partly because our method provides very precise estimates (narrow 95% credible intervals). Thus, even for







**Fig. 7** Coverage from simulated data using our Bayesian method and classic methods (2SLS when sample overlapping rate = 1 and IVW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0.3}$  and  $\delta = (\delta_1, \delta_2, \delta_3, \delta_4) = \mathbf{1}$ . Results with two instrument strengths (top panels:  $\alpha = (\alpha_1, \alpha_2, \alpha_3) = \mathbf{0.5}$  and bottom panels:  $\alpha = \mathbf{0.1}$ ) are displayed, from left to right, for estimated causal effect of  $X_1, X_2$  and  $X_3$  on  $Y$ , denoted as  $\hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\beta}_3$  respectively. Each panel consists results of six different sample overlapping rates ranging from 0 (leftmost) to 1 (rightmost) and each box plot represents causal effect estimated from 200 simulated datasets



**Fig. 8** Coverage from simulated data using our Bayesian method and classic methods (2SLS when sample overlapping rate = 1 and IVW otherwise) when  $\beta = (\beta_1, \beta_2, \beta_3) = \mathbf{0}$  and  $\delta = (\delta_1, \delta_2, \delta_3, \delta_4) = \mathbf{1}$ . Results with two instrument strengths (top panels:  $\alpha = (\alpha_1, \alpha_2, \alpha_3) = \mathbf{0.5}$  and bottom panels:  $\alpha = \mathbf{0.1}$ ) are displayed, from left to right, for estimated causal effect of  $X_1, X_2$  and  $X_3$  on  $Y$ , denoted as  $\hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\beta}_3$  respectively. Each panel consists results of six different sample overlapping rates ranging from 0 (leftmost) to 1 (rightmost) and each box plot represents causal effect estimated from 200 simulated datasets

a slightly biased estimate, it is likely the credible interval does not include the true value, which will affect its performance in coverage.

## Discussion

In this paper we have proposed a Bayesian method that effectively enables a one-sample MR whether there are two overlapping samples or disjoint samples. It is noteworthy that our method has the best performance in one-sample setting, which is unsurprising because 1) our method has been developed on the basis of a Bayesian approach tailored for one-sample MR [5]; 2) the percentage of imputed data becomes lower as overlapping rate increases, and consequently, the uncertainty in data decreases. As discussed, our method provides the flexibility of combining data from previous studies with current ones to enhance statistical power, which is particularly advantageous because MR studies are often underpowered partly due to small sample sizes. It is also shown from our simulation results that pleiotropy (effect of  $Z_2$  on  $Y$  mediated by both  $X_2$  and  $X_3$ ) resulted in the worst estimated causal effect, because it involves two completing paths which led to higher uncertainty in estimation. This is, however, an issue commonly seen in MR studies [20, 21].

The precision of the estimates from our proposed method was higher than that from the classic MR. Our method also resulted in higher power of the estimates under the alternative hypothesis where the true causal effects were different from zero. Coverage of the estimates from both methods was similar under the null as well as the alternative hypotheses.

Our method is limited by the fact that we have focussed on a simple model with only three IVs and three exposures and a moderate number of configurations, although the configurations can never be exhaustive. In the real world, we often encounter much more complex data in which there are possibly many (weak and/or correlated) IVs and (correlated) exposures and outcomes, which triggers a research topic of variable selections in future MR methodology [22]. Bayesian approach, however, is well known for its flexibility of building various models to address different scientific questions.

## Conclusions

We have developed a Bayesian MR that can be applied to any of the sample settings by means of treating missing data as unknown parameters which can then be imputed using MCMC. This is an important addition to the existing MR literature where some methods can only be used for one-sample and others for two-sample settings. Because of the nature of Bayesian inference, our model can be easily extended to tackle more complex MR analysis in medical research.

## Abbreviations

RCT: Randomised controlled trial; MR: Mendelian randomization; IV: Instrumental variable; SNP: Single nucleotide polymorphism; BMI: Body mass index; MCMC: Markov chain Monte Carlo; IWW: Inverse-variance weighted; 2SLS: Two-stage least squares

## Acknowledgements

Not applicable.

## Authors' contributions

HG and CB conceived and 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.

## Ethics approval and consent to participate

Not applicable

## Consent for publication

Not applicable

## Competing interests

The authors declare that they have no competing interests.

Received: 6 August 2020 Accepted: 19 November 2020

Published online: 07 December 2020

## References

- Katan MB. Apolipoprotein e isoforms, serum cholesterol, and cancer. *Lancet*. 1986;327:507–8.
- Smith GD, Ebrahim S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1):1–22.
- Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27(8):1133–63.
- Robinson PC, Choi HK, Do R, Merriman TR. Insight into rheumatological cause and effect through the use of mendelian randomization. *Nat Rev Rheumatol*. 2016;12(8):486–96.
- Berzuini C, Guo H, Burgess S, Bernardinelli L. A bayesian approach to mendelian randomization with multiple pleiotropic variants. *Biostatistics*. 2020;21(1):86–101.
- Johnson T, Uk GS. Efficient Calculation for Multi-SNP Genetic Risk Scores. <https://cran.r-project.org/src/contrib/Archive/gtx>. Accessed 7 Jul 2020.
- 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.
- 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(4):304–14.
- Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical Inference in Two-sample Summary-data Mendelian Randomization Using Robust Adjusted Profile Score. <https://arxiv.org/abs/1801.09652>.
- Burgess S, Thompson SG. *MENDELIAN RANDOMIZATION Methods for Using Genetic Variants in Causal Estimation*. London: Chapman & Hall/CRC Press; 2015.
- Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variable regression. *J Econ*. 2003;114(1):29–72.
- 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.

13. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample mendelian randomization. *Genetic Epidemiology*. 2016;40(7): 597–608.
14. LeBlanc M, Zuber V, Thompson WK, Andreassen OA, Schizophrenia and Bipolar Disorder Working Groups of the Psychiatric Genomics Consortium, Frigessi A, Andreassen BK. A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics*. 2018;19(494): <https://doi.org/10.1186/s12864-018-4859-7>.
15. Jordan DM, Verbanck M, Do R. Hops: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biol*. 2019;20(222):. <https://doi.org/10.1101/311332>.
16. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in mendelian randomization studies. *Int J Epidemiol*. 2011;40(3):755–64.
17. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21(6):1087–92.
18. Stan Development Team. STAN: A C++ Library for Probability and Sampling, Version 2.18.2. <https://mc-stan.org>. Accessed 15 Sep 2020.
19. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. *Found Trends Mach Learn*. 2008;1(1-2):1–305.
20. Hemani G, Bowden J, Smith GD. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human Mol Genet*. 2018;27(R2):195–208.
21. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nat Genet*. 2018;50:693–8.
22. Zuber V, Colijn JM, Klaver C, Burgess S. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nat Commun*. 2020;11(29):. <https://doi.org/10.1038/s41467-019-13870-3>.

### Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

