Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Mendelian randomization incorporating uncertainty about pleiotropy

John R Thompson[1]* Cosetta Minelli[2], Jack Bowden[3], Fabiola Del Greco M[5], Dipender Gill[6], Elinor M Jones[7], Chin Yang Shapland[8], Nuala A Sheehan[1]

Mendelian randomization (MR) requires strong assumptions about the genetic instruments, of which the most difficult to justify relate to pleiotropy. In a two-sample MR different methods of analysis are available if we are able to assume, M_1: no pleiotropy (fixed effects meta-analysis), M_2: that there may be pleiotropy but that the average pleiotropic effect is zero (random effects meta-analysis), M_3: that the average pleiotropic effect is non-zero (MR-Egger). In the latter two cases we also require that the size of the pleiotropy is independent of the size of the effect on the exposure. Selecting one of these models without good reason would run the risk of misrepresenting the evidence for causality. The most conservative strategy would be to use M_3 in all analyses as this makes the weakest assumptions, but such an analysis gives much less precise estimates and so should be avoided whenever stronger assumptions are credible. We consider the situation of a two-sample design when we are unsure which of these three pleiotropy models is appropriate. The analysis is placed within a Bayesian framework and Bayesian model averaging (BMA) is used. We demonstrate that even large samples of the scale used in genome-wide meta-analysis may be insufficient to distinguish the pleiotropy models based on the data alone. Our simulations show that BMA provides a reasonable trade-off between bias and precision. BMA is recommended whenever there is uncertainty about the nature of the pleiotropy.

Keywords: Mendelian randomization; pleiotropy; MR-Egger; meta-analysis; Bayesian model averaging

1. Introduction

Mendelian randomization (MR) is an epidemiological instrumental variable analysis that uses genetic variants as the instruments and as such it provides a way of investigating causal exposure-outcome relationships using observational data [1]. Many early applications of MR used a single genetic variant with a well-established association with the exposure.

[1] *Department of Health Sciences, University of Leicester, Leicester, UK*
[2] *Population Health and Occupational Disease, NHLI, Imperial College London, London, UK*
[3] *MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK*
[5] *Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC), Bolzano/Bozen, Italy*
[6] *Department of Clinical Pharmacology and Therapeutics, Imperial College London, London, UK*
[7] *Department of Statistical Science, University College London, London, UK*
[8] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
**Correspondence to: John Thompson. E-mail: john.thompson@le.ac.uk*

To ensure that such an analysis is valid, researchers have to be confident that the genetic variant satisfies three core assumptions; (1) the variant is related to the exposure, (2) the variant is independent of the confounders and (3) the variant is independent of the outcome given the exposure and the confounders [2]. These assumptions imply that the variant will not have a pleiotropic effect on the outcome, that is, there will be no effect of the variant that acts on the outcome through a pathway that does not pass through the exposure [3]. The lack of pleiotropy is not testable from the data and so has to be assessed on biological grounds. As a consequence, the most reliable MR studies are restricted to using genetic variants with a well-understood biological function.

In a classic MR study the exposure, outcome and genetic instrument are measured on the same sample of subjects but such studies are usually limited in sample size with the result that the test of causality has low power and the estimates of the size of the causal effect are not very precise [4, 5]. When the sample size cannot be increased, an alternative way to improve precision is to find an instrument that has a stronger association with the exposure. To this end, researchers have used genetic risk scores based on several genetic variants chosen for their known association with the exposure [6, 7]. Assuming that the risk score is more predictive of the exposure than is any one instrument, then the precision of the MR estimate will improve, although it will be necessary to understand the biology of all of the variants.

In the era of genome-wide association studies (GWAS) there has been a dramatic rise in the number of potential instruments that could be used in MR studies. For some easily measured and strongly genetic traits, such as height, there are now hundreds of genetic variants that could be suitable [8]. To harness these data, researchers have started to adopt a two-sample design in which the variant-exposure information comes from one GWAS and the variant-outcome information comes from a second GWAS [9, 10, 11, 12]. Such data provide valid MR estimates provided that the variant-exposure relationship is the same in both study populations and, of course, that the other MR assumptions hold for all of the variants. This means that we need to be confident that none of the variants has a pleiotropic effect.

In a two-sample design, instead of creating a risk score that combines the variants into a single instrument, separate MR analyses can be performed with each variant in turn and then the MR estimates can be combined in a meta-analysis over the variants. Assuming no heterogeneity, each MR analysis will estimate the same causal exposure-outcome effect and a pooled estimate can be obtained by an inverse variance weighted fixed effects meta-analysis. What is more, any evidence of heterogeneity in the meta-analysis would suggest that some or all of the variants produce biased estimates and a possible explanation for this is pleiotropy [13]. A meta-analysis thus offers a mechanism for testing homogeneity, used as a proxy for the no pleiotropy assumption, although it will not tell us which variants are valid instruments and which are not.

With the growth in the number of available instruments, it has become more important to look for statistical methods that are robust to pleiotropy. One possible argument is that in a set of well-selected variants, pleiotropy is likely to be small and to act randomly, sometimes increasing the MR estimate and sometimes decreasing it. The overall impact would be to increase the heterogeneity between variant-specific estimates but on average to give the correct result. When this assumption is reasonable, the correct two-sample MR analysis will take the form of a random effects meta-analysis over the variants [14].

When the effects of pleiotropy do not cancel, a random effects meta-analysis would not be valid but as Bowden et al. [15] noted, an analysis that mirrors Egger regression, as used in meta-analysis to detect publication bias, would provide valid MR estimates under the weaker assumption that the sizes of the pleiotropic effects are independent of the sizes of the corresponding variant-exposure effects; they called this the Instrument Strength Independent of Direct Effect (InSIDE) assumption and the analysis was named MR-Egger regression. Effectively this analysis performs a weighted regression of the variant-outcome effect estimates on the variant-exposure estimates, so that the intercept reflects the average pleiotropy over the set of variants and the slope provides an unbiased MR estimate.

When there is no strong evidence to favour one set of pleiotropy assumptions over another, it would be misleading to select one form of analysis and only to present its results. The evidence for causality can be very different under the three forms of analysis and reporting the results under a single model for pleiotropy ignores our model uncertainty.

Unfortunately, MR-Egger regression analysis usually provides much less precision than a random effects meta-analysis and in turn a random effects meta-analysis is less precise than a fixed effects meta-analysis. The issue of model choice is therefore critically important and it is not sensible always to use MR-Egger regression on the grounds that it makes the weakest assumptions. Rarely will there be sufficient biological or external experimental information about the extent of pleiotropy to enable an MR model to be chosen with complete confidence and to select the model using the data that will subsequently be analysed can lead to bias and over-optimistic inferences [16, 17]. For this reason, many researchers have advocated averaging the estimates over a set of plausible models, see [18] for a Bayesian review or [19] for a discussion from a frequentist perspective. One advantage of the Bayesian approach to model averaging is that it gives the posterior model probabilities, which provide a natural set of weights for the model averaging. This approach has found wide-spread application in many fields, including epidemiology and genetics. For instance, Stephens and Balding recommended averaging over additive and non-additive genetic models [20], and model averaging has also been proposed for prioritizing genetic associations [21], combining models of gene-gene or gene-environment interaction [22, 23], for analysing microarray experiments [24], and for handling different covariate patterns [25].

When model averaged estimates are used, we have to accept that point estimates will provide a biased estimate of the causal effect because the method averages over different models that cannot all be correct. Although the relevance of frequentist properties to Bayesian analysis is debatable [26], the gain in model averaging comes from a bias-variance trade-off that produces better prediction with smaller root mean square error [27, 18]. The idea of a bias-variance trade-off underlies the benefits of k-class estimators for instrumental variable analysis [28] and has recently been used in a non-Bayesian context to combine two-stage least squares and ordinary least squares estimates [29].

In this paper we demonstrate how Bayesian model averaging enables us to make an appropriate adjustment for our uncertainty about the pattern of pleiotropy. In section 2, we introduce our notation and describe the three MR models that we will consider. These are placed in a Bayesian context in sections 3 and 4. The simulations in section 5 illustrate the benefits of model averaging and its ability to quantify the impact of our uncertainty about the pattern of pleiotropy and also demonstrates the impact of assuming the wrong model. In section 6, Bayesian model averaging is applied to an MR of the effect of age at menarche on lung function.

2. Meta-analysis of variant-specific estimates

Suppose that we have selected M independent genetic variants G_j, $j = 1, \ldots, M$ to act as instruments in a two-sample Mendelian randomization. The first study provides estimates, $\hat{\gamma}_j$, and standard errors s_j for the variant-exposure effects γ_j and the second study provides estimates, $\hat{\Gamma}_j$, and standard errors S_j for the variant-outcome effects Γ_j.

Assuming that the true sizes of the variant-exposure effects, γ_j, are the same in both study populations, that the second study has N subjects and that in that study the unobserved exposure, X, and observed outcome, Y, follow linear models with error terms ϵ and υ then we have,

$$\begin{aligned} x_i &= \alpha_{0j} + \gamma_j g_{ij} + \epsilon_{ij} \qquad i = 1, \ldots, N \quad j = 1, \ldots, M \\ y_i &= \alpha_1 + \beta x_i + \upsilon_i. \end{aligned} \tag{1}$$

where g_{ij} is the value of genetic variant j for subject i.

Least squares regression with the second equation would give a biased estimate of β because, although the core MR assumptions imply that the errors are independent of the genetic variants, they are not necessarily independent of X. Substituting one equation into the other,

$$y_i = \alpha^*_{1j} + \beta\gamma_j g_{ij} + \upsilon^*_{ij} \tag{2}$$

where $\alpha^*_{1j} = \alpha_1 + \beta\alpha_{0j}$ and $\upsilon^*_{ij} = \upsilon_i + \beta\epsilon_{ij}$. The error term in this derived relationship is assumed independent of the genetic variant so that regression analysis provides unbiased estimates of the coefficient of G_j. What is more, since $\Gamma_j = \beta\gamma_j$ we can obtain M Wald or ratio estimates of the causal exposure-outcome effect β as,

$$\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j \qquad j = 1, \ldots, M \tag{3}$$

where estimates $\hat{\gamma}_j$ are taken from the first study.

We will refer to this model in which there is no pleiotropy as model one (M_1) and under this model we can obtain a pooled estimate of β using inverse variance weighted fixed effects meta-analysis [30].

$$\hat{\beta} = \frac{\sum_{j=1}^{M} \hat{\beta}_j / V_j}{\sum_{j=1}^{M} 1/V_j}. \tag{4}$$

When s_j, the standard error of $\hat{\gamma}_j$, is negligible the variance, V_j, of $\hat{\beta}_j$ can be estimated by $S_j^2/\hat{\gamma}_j^2$. The delta method can be used to obtain estimates of V_j that allow for the uncertainty in $\hat{\gamma}_j$ [31], but these have to be used with care as such estimates are correlated with the estimate of $\hat{\beta}_j$ and this correlation can introduce its own bias into $\hat{\beta}$ [11].

We consider pleiotropy in the form of a direct effect on the outcome via secondary pathways so that,

$$y_i = \alpha_{1j} + \beta x_i + \psi_j g_{ij} + \upsilon_{ij}. \tag{5}$$

Now substitution gives,

$$y_i = \alpha^*_{1j} + (\beta\gamma_j + \psi_j) g_{ij} + \upsilon^*_{ij} \tag{6}$$

and the Wald estimator will estimate $\beta + \psi_j/\gamma_j$ rather than β. If we are to meta-analyse the Wald estimators then we need to make the assumption that errors due to pleiotropy, ψ_j/γ_j, will cancel so that

$$E\left\{\frac{\psi_j}{\gamma_j}\right\} = 0 \tag{7}$$

which is slightly different from the more natural assumption of cancelling pleiotropy $E\left\{\psi_j\right\} = 0$.

We will refer to the model in which the pleiotropy is assumed to cancel (i.e. $E\left\{\psi_j\right\} = 0$) as model M_2 and the model in which the errors in the Wald estimators cancel (i.e. $E\left\{\psi_j/\gamma_j\right\} = 0$) as model M_{2a}. Under M_2a the pooled estimate of

β can now be obtained by random effects meta-analysis,

$$\hat{\beta} = \frac{\sum_{j=1}^{M} \hat{\beta}_j/(V_j + \tau^2)}{\sum_{j=1}^{M} 1/(V_j + \tau^2)} \tag{8}$$

where τ^2 is the variance of the terms ψ_j/γ_j. This estimator will be consistent under model M_{2a} and consistent under model M_2 provided that the pleiotropy, ψ_j, is independent of the size of the genetic effect, γ_j. This is the InSIDE assumption introduced by Bowden et al. in the context of MR-Egger regression [15], see also [32, 14].

When we are unwilling to assume that the pleiotropy cancels, we are left with the relationship,

$$\Gamma_j = \beta\gamma_j + \psi_j \tag{9}$$

which we might fit via a regression of the estimates,

$$\hat{\Gamma}_j = \mu + \beta\hat{\gamma}_j + \zeta_j \tag{10}$$

where μ is the average of the ψ_j and the zero-centred error term ζ_j depends on the estimation error in $\hat{\Gamma}_j$ and the amount of pleiotropy, ψ_j. The error terms will vary in size from variant to variant, so we might represent this by writing $Var(\zeta_j) = \sigma_j^2$.

A useful diagnostic plot is obtained by plotting $\hat{\Gamma}_j$ against $\hat{\gamma}_j$ [15], in this plot the slope will represent β and the intercept will represent μ. Alternatively we can divide through by $\hat{\gamma}_j$ and obtain,

$$\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} = \hat{\beta}_j = \beta + \frac{\mu}{\hat{\gamma}_j} + \frac{\zeta_j}{\hat{\gamma}_j} \tag{11}$$

and a second diagnostic plot is obtained by plotting $\hat{\beta}_j$ against $1/\hat{\gamma}_j$. In this plot the intercept represents β and any slope represents directional pleiotropy, μ. Similar plots have been suggested for MR when a single variant is measured in multiple studies [33, 34].

Bowden et al. [15] noticed that this regression has the same structure as was assumed by Egger and colleagues when testing for publication bias [35]. We will get a consistent estimate of β from the regression provided that estimates of the variant-exposure effect $\hat{\gamma}_j$ are independent of the errors ζ_j. If the sampling error in $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ are random then this assumption requires the InSIDE assumption that γ_j is independent of ψ_j. We refer to this as model three (M_3). One small problem with the proposal in [15] is that they suggested fitting MR-Egger regression using weighted least squares with weights inversely proportional to V_j as in equation (4). It would perhaps be better to use weights that are inversely proportion to $V_j + \tau^2$ as in equation (8), but such an analysis would not have been as easy to implement with standard regression software. In the following section we develop a Bayesian analysis equivalent to the least squares analysis with the weights of equation (8).

3. Bayesian Analysis

In this section we develop a Bayesian formulation of the meta-analysis models from section 2 in which M_1, M_2 and M_3 are nested. Under a Bayesian model we are required to place distributions over the effect sizes of genetic variants on X and Y and over the size of the pleiotropy. We do this by assuming that the discrete distributions over the set of genetic variants in the MR can be approximated by continuous normal distributions, although it would be perfectly possible to adapt these models to allow other distributional forms. Under model M_3 the estimates $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ will be related by the regression equation,

$$M_3: \quad \hat{\Gamma}_j = \mu + \beta\hat{\gamma}_j + \zeta_j. \tag{12}$$

We will adopt the convention that each of the variants is coded so that the effect on the exposure, γ_j is positive. This convention is not necessary but does simplify the coding of the model as it will only require a single intercept μ rather than μ for positive γ_j and $-\mu$ for negative γ_j. The diagnostic plots mentioned in section 2 are also easier to interpret when all γ_j are positive. The pooled estimate of the exposure-outcome effect is given by the slope and the average pleiotropy is given by the intercept and the variance of the pleiotropy forms part of the variance of ζ_j since under independence of the error terms,

$$Var(\zeta_j) = Var(\psi_j) + S_j^2. \tag{13}$$

Provided that we can make the InSIDE assumption that ψ_j is independent of γ_j and assuming random estimation errors in $\hat{\gamma}_j$ and $\hat{\Gamma}_j$, then ζ_j will be independent of $\hat{\gamma}_j$ and we will be able to estimate β from the regression. In the Bayesian formulation we further assume that the pleiotropy terms, ψ_j, come from a common distribution with

$$Var(\psi_j) = \tau^2. \tag{14}$$

Under M_2 the pleiotropy cancels so that the intercept μ is zero and the regression line passes through the origin.

$$M_2: \quad \hat{\Gamma}_j = \beta\hat{\gamma}_j + \zeta_j \tag{15}$$

and here we still require the InSIDE assumption to hold.

Under M_1 there is no pleiotropy, i.e. $\psi_j = 0$ for all j, so,

$$M_1: \quad \hat{\Gamma}_j = \beta\hat{\gamma}_j + \zeta_j^* \tag{16}$$

where

$$Var(\zeta_j^*) = S_j^2. \tag{17}$$

In a Bayesian analysis it is a simple matter to allow for the uncertainty in the estimate of the variant-exposure effect, γ_j, that is ignored in most Mendelian randomization analyses of a two-sample design; see Burgess et al. and Bowden et al. for examples of papers that adjust for this uncertainty in a non-Bayesian analysis [36, 37]. We simply replace model M_3 with,

$$M_3: \quad \hat{\Gamma}_j = \mu + \beta\gamma_j + \zeta_j \tag{18}$$

and create a hierarchical structure in which we assume a normal distribution for the sampling error,

$$\hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \omega_j^2). \tag{19}$$

We could then approximate ω_j by the standard error s_j Applying de Finetti's theorem, those variant-exposure effects about which our prior opinions are exchangeable, can be treated as being independently drawn from a common distribution. In our examples we assume a normal distribution but in real data sets the distribution could be adapted to fit with our beliefs about the variant-exposure effects, see Gelman et al. Chapter 5 for a discussion [38].

$$\gamma_j \sim \mathrm{N}(m, \upsilon) \tag{20}$$

with m and υ estimated from the data. A similar hierarchical structure can be used to allow for variant-exposure sampling error in models M_2 and M_1. In all analyses we ignore the uncertainty in the standard errors s_j and S_j.

3.1. Bayesian model averaging

Let θ represent the full set of parameters, (β, μ, τ^2). Under M_2, $\theta = (\beta, 0, \tau^2)$ and under M_1, $\theta = (\beta, 0, 0)$. The dataset for the analysis, D, will contain the estimates $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ and their standard errors S_j and s_j. The joint posterior will be given by

$$P(\theta, M_k|D) \propto P(D|\theta, M_k)P(\theta|M_k)P(M_k) \qquad k = 1, 2, 3.$$

The researcher must specify their prior beliefs in the three models, $P(M_k)$, and their prior distributions for the relevant parameters given each model, $P(\theta|M_k)$.

The OpenBUGS code given in the supplement will fit this mixture of models using the method of Carlin and Chib [39] in which we introduce an indicator variable, T, to represent the current model choice. From the output, we can use the chain of values of T to estimate the posterior model probabilities, $P(M_k|D)$, and we can extract the parts of the chain for which $T = M_k$ in order to estimate the posterior distributions of the parameters, $P(\theta|M_k, D)$. The model averaged posterior $P(\theta|D)$ is obtained from the entire chain irrespective of T. In particular, we will be interested in the marginal posterior $P(\beta|D)$, which will estimate the exposure-outcome effect allowing for our model uncertainty.

The Bayes Factor is a commonly used measure of the support for one model over another as provided by the data. The Bayes Factor can be derived from the model averaging Markov chain Monte Carlo (MCMC) analysis by taking the posterior model probabilities, $P(M_k|D)$, and contrasting them with the priors, $P(M_k)$. The Bayes Factor BF_{21} that measures the support for M_2 over M_1 is calculated from

$$\mathrm{BF}_{21} = \frac{P(M_2|D)}{P(M_1|D)} / \frac{P(M_2)}{P(M_1)}.$$

A commonly used threshold for judging Bayes Factors is to take values over 20 as indicating that the data provide strong evidence in favour of one model over another and values over 150 as indicating very strong evidence [40]. In our analyses we start with equal priors on all models so strong evidence requires a posterior model probability over 0.952 and very strong evidence requires a model probability over 0.993. In cases where the data do provide strong evidence for choosing one model over another, the model averaged estimates will be heavily dominated by that choice of model and will naturally follow the model that the Bayes Factor would have chosen.

4. Choice of Priors

Applied researchers are usually happy to use Bayesian methods with vague or non-informative priors, but many of them are concerned when a prior has a noticeable influence on the posterior distribution. In contrast many theoretical researchers argue that truly non-informative priors do not exist [41]. In the context of Mendelian randomization it is much harder to limit oneself to minimally informative priors because our conclusions will depend on what we believe about pleiotropy and this information cannot come from the data alone. Causality cannot be deduced from observational data without external knowledge.

A sensible prior for β might be based on epidemiological studies that have attempted to adjust for known confounders between the exposure and the outcome. Under M_3, the likely size of μ might be influenced by whether we expect positive or negative pleiotropy. For instance, if pleiotropy is suspected because of an effect through an identified third factor, W, then we may have external information about both the effects of W on Y and of the genetic variants on W. Under M_2 and M_3, the prior for τ^2 will need to be scaled to fit with the anticipated size of the effect of the variants on the outcome and the anticipated severity of the pleiotropy. Sometimes it is helpful to think of the size of the pleiotropy relative to the size of the variant-exposure effect. It is not necessary to have the same prior for τ^2 under M_2 and M_3 but, in practice, it is unlikely that we will have the external information to distinguish the two conditions.

We fitted the models using OpenBUGS 3.2.2 (www.openbugs.net), in which the MCMC algorithm needs to update the full vector of parameters, θ, even under M_1 and M_2 when some of the values are assumed known, so the algorithm requires us to define pseudo priors, that is priors for updating μ and τ^2 when we are fitting models that assume them to be zero. The choice of the pseudo priors will not affect any of the posterior distributions but it can have a big impact on the speed of convergence of the MCMC algorithm. If informative priors are used for M_3 then those same priors are likely to be good choices for the pseudo priors under M_1 and M_2. While if non-informative priors are used, it will be necessary to define more restrictive pseudo priors in order to obtain convergence in a reasonable time.

5. Simulations

5.1. Design

In order to investigate the properties of Bayesian model averaging of MR analyses, we simulated data consisting of 50 independent genetic variants taken from a two-sample design in which there were either 50,000 subjects in each sample or 5,000 subjects in each sample. Allele frequencies were randomly selected between 0.1 and 0.9 and the variant-exposure effects were randomly generated from a normal distribution with mean 0.15 and standard deviation 0.04. The variance of the exposure was set to 2 so that on average each variant explained 0.47% of the variance and together the 50 variants are expected to explain 24% of the variance in the exposure. This is a larger proportion of the variance than is usually explained by the genetic variants in an MR but is chosen to emphasise the patterns and to avoid the extra complexity of weak instrument effects. Pleiotropy was assigned independently of the genetic effect on exposure by drawing it from a normal distribution with mean, m_p, and standard deviation s_p, under model M_1: $m_p = s_p = 0$, under M_2: $m_p = 0, \quad s_p \neq 0$ and under M_3: $m_p \neq 0, \quad s_p \neq 0$. Two exposure-outcome relationships were investigated, one in which there was no causal association ($\beta = 0$) and one in which the exposure was directly responsible for 25% of the variance in the outcome ($\beta = 0.5$). Confounding was introduced by allowing 50% of the unexplained variance in both the exposure and outcome to depend on a common random factor.

In order to summarise the size of the pleiotropy we calculated the statistic,

$$\text{MAPR} = 100 \times \text{Median}\left\{ \left| \frac{\psi_j/\sigma_Y}{\gamma_j/\sigma_X} \right| \right\}.$$

This is the Median Absolute Pleiotropy Ratio, that is the percentage ratio of the variant's pleiotropic effect size on a standardized outcome divided by the variant's effect on a standardized exposure. To create the standardized scales we divide by the standard deviations of the outcome and exposure, σ_Y and σ_X. When the pleiotropy, ψ_j, is not zero-centred, MAPR approaches the average standardized pleiotropy as a percentage of the average standardized variant-exposure effect size, and when the average pleiotropy is zero, MAPR is about 0.7 times the standard deviation in the standardized pleiotropy as a percentage of the typical standardized variant-exposure effect size.

All analyses were performed with equal prior probabilities for the three models. Realistically vague priors were placed on the other parameters, that is to say priors that had a negligible effect on the posterior but which ruled out such extreme values that the performance of the algorithm might be affected. β was given a normal prior with mean 0 and standard deviation 10. When analysing by MR-Egger, a vague normal prior with mean 0 and standard deviation 10 was placed on the regression intercept. The precision of the heterogeneity was given a gamma prior with parameters 2 and 0.00005; this makes the expected heterogeneity due to pleiotropy have a precision of 40,000, equivalent to a standard deviation of 0.005. These priors are plotted in the supplement as Figures S.1 to S.3 in which they are contrasted with the more informative priors used in the sensitivity analysis.

5.2. Illustrative simulation

To illustrate the model averaging and to help investigate the mixing, we simulated a single dataset of 50 genetic variants with random allele frequencies in the range 0.1 to 0.9 in two GWAS of size 50,000. The true variant-exposure effects were drawn from a normal distribution with mean 0.15 and standard deviation 0.04 and the true pleiotropy was drawn from an independent normal distribution with mean 0.06 and standard deviation 0.025. In this example our measure of the impact of pleiotropy, MAPR, is approximately 40%. The true causal exposure-outcome effect size, β, was set to 0.5. These data follow model M_3 and the InSIDE assumption holds because of the independence between the exposure effect size and the pleiotropy.

Figure 1 shows the two diagnostic plots for the simulated data. Estimated lines from an analysis of the illustrative data have been added to these diagnostic plots. The lines corresponding to M_1 are not shown for this particular example as M_1 is not supported by these data. In the left hand type of plot, models M_1 and M_2 would produce lines through the origin with different amounts of variation about the line, while M_3 would produce a line that does not pass through the origin. The slope of the lines corresponds to the exposure-outcome effect size, β. In the right hand type of plot, M_1 and M_2 would produce horizontal lines again with different variances and M_3 would create a sloping line. In the right hand plot, the intercept with the y-axis corresponds to the estimate of β.

Figure 2 shows trace plots of a run of 10,000 following a burn-in of 1,000. The data were simulated under M_3 but only 34% of the chain was spent in the MR-Egger model (M_3) while 66% of the time the analysis favoured the random effects model (M_2). So the non-zero intercept in the regression is not sufficiently obvious that model M_2 can be ruled out, but the excess variance about zero is sufficient to rule out the fixed effects model, M_1. The estimates of the intercept, μ, are small and positive when in M_3, while μ is fixed at zero under M_2 so those values in the trace plot are merely random values from the pseudo prior and should be ignored. The estimates of the exposure-outcome effect are very different under models M_2 and M_3 and the precision also differs; under M_2 the MR estimates of β are higher (mean 0.91 vs 0.63) and more precise. The standard deviation of the heterogeneity is larger under M_2 because there is no average pleiotropic shift from zero to help explain the heterogeneity.

A smoothed estimate of the posterior distribution for β is shown by the solid line in Figure 3. The model averaged posterior mean is 0.82 but we must be careful how this is interpreted because values of β close to 0.8 are very poorly supported by the evidence. The analysis suggests that either β is around 0.63 or it is around 0.91. In a model averaged

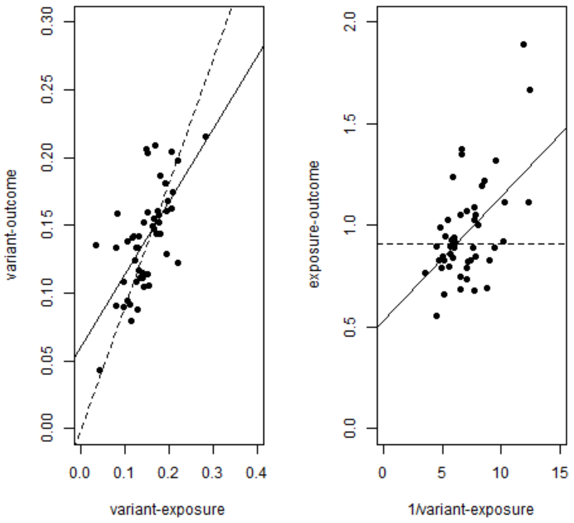

Figure 1. Diagnostic plots for a single simulated dataset. The solid line shows the model fit under M_3 and the dashed line shows the fit under M_2.

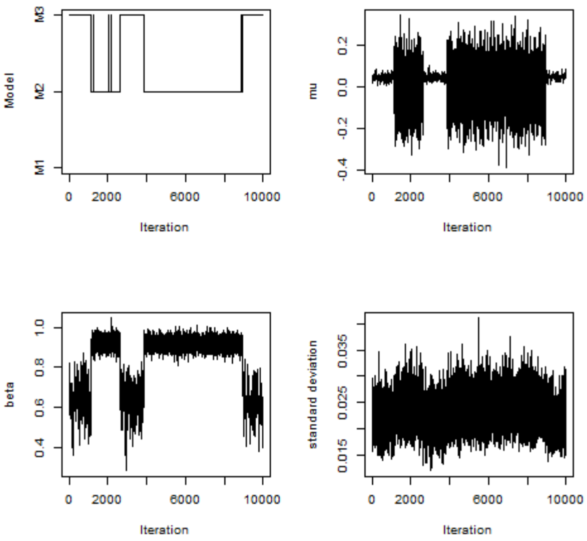

Figure 2. Illustrative trace plots for the analysis of a single simulated dataset

analysis, the posterior distributions are likely to be multi-modal, so it is important to report the whole posterior and to interpret the usual summary statistics with caution.

5.3. Length of chain and number of simulations

Figure 2 was generated from a chain of length 10,000 after a burn-in of 1,000. Re-running the analysis with random starting values showed that a chain of this length is insufficient to give reproducible estimates of the posterior model probabilities. By trial and error, we found that 10 chains for analysing the dataset shown in Figure 1, each with a burn-in of 10,000 and a length 500,000 thinned by 10 to a length 50,000, gave a posterior probability for M_3 that averaged 37.0%

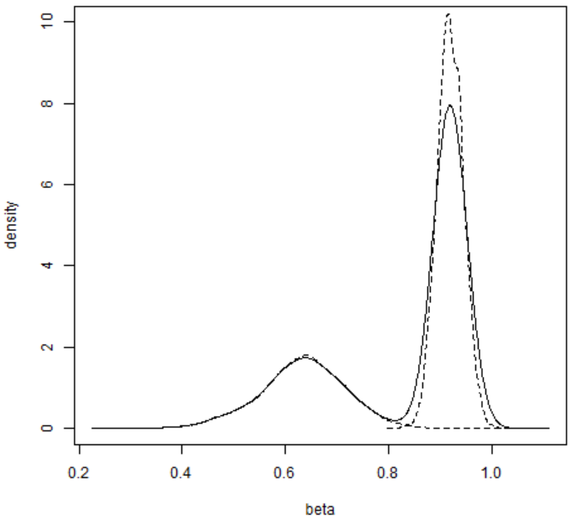

Figure 3. Smoothed estimate of the model averaged posterior distribution for the causal effect, β (solid line). Model specific posterior distribution estimates scaled to have area equal to the posterior model probability are shown as dashed lines.

with a between chain standard deviation of 2.6% and for the model averaged posterior mean of the causal effect, the average over the 10 chains was 0.812 and the standard deviation was 0.008. The necessary run length can be reduced by centring the explanatory variables in a Bayesian regression [42], but centring changes the meaning of the intercept. In this problem the intercept is a meaningful parameter about which we may have some prior knowledge and so we decided not to centre.

If we were to analyse 100 random datasets simulated under exactly the same scenario as was used in the illustrative example, then the average posterior model probabilities would have two sources of variability; one from the MCMC analysis of each dataset and the other from the variation between datasets generated under the same scenario. The standard error of the average performance due to MCMC error will be 0.26% and the corresponding standard error for the model averaged posterior means of β will be 0.0008. The standard error of the posterior model probabilities in the simulations of section 5.5 was up to 2.5%, so this run length is such that MCMC error makes a very small contribution to the results. The standard error of the posterior mean for β across datasets was negligible in comparison at about 0.001.

5.4. Mixed Simulation

The first simulation consists of 100 random datasets generated with equal probability of having (i) no pleiotropy (M_1), (ii) zero-centred pleiotropy (M_2) or (iii) non-zero mean pleiotropy (M_3). Details of the specific parameters used are given in the supplement. The datasets were analysed in four ways, (i) always using a fixed effect model (M_1), (ii) always using a random-effects model (M_2), (iii) always using MR-Egger regression (M_3) and (iv) always using model averaging. The results are shown in Figure 4. Within each graph the 100 datasets are sorted so that those with no pleiotropy are on the left, those with zero-centred pleiotropy are in the middle and those with non-zero-centred pleiotropy are on the right. The corresponding root mean square errors (rmse) of the point estimates of β (posterior means) about the true value of 0.5 are shown in Table 1.

The top left plot of Figure 4 shows the results when the data were analysed using a Bayesian Mendelian randomization based on a fixed effects model. This analysis assumes no pleiotropy. The analysis performed very well when the data were generated without pleiotropy (rmse=0.0099), but performance declined when the data were generated with zero-centred pleiotropy (rmse=0.0814) and performance is visibly poor to the right of the plot when the data were generated with directional pleiotropy (rmse=0.5551).

Table 1 confirms the pattern that is evident from the plot. Performance is optimal if we match the analysis with the model used to simulate the data, however in practice, we do not know what model should be used, so it is important to see what happens if the wrong analysis is used. Fixed and random effects analyses perform very poorly when directional pleiotropy is present. MR-Egger is best for data with directional pleiotropy but performance is noticeablly sub-optimally when either there is no pleiotropy (rmse 0.0311 vs 0.0099) or the pleiotropy is zero-centred (rmse 0.1977 vs 0.0814). Model averaging offers a method of analysis that stays close to optimal whatever the true pleiotropy model because it learns which analysis model is appropriate from the data.

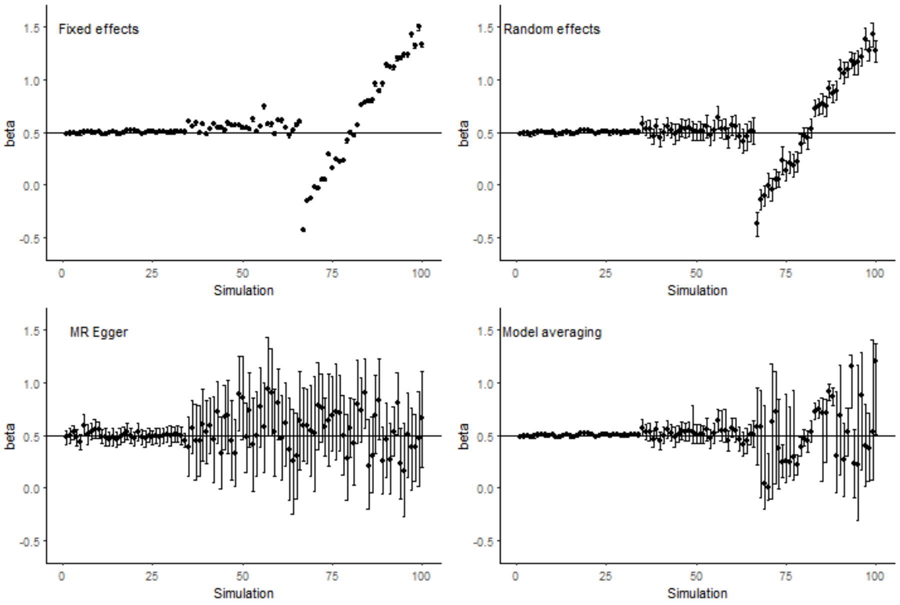

Figure 4. Posterior means and 95% credible intervals for 100 random datasets. The datasets are sorted within each graph so that those without pleiotropy are on the left, those with zero-centred pleiotropy are in the middle and those with non-zero mean pleiotropy are on the right. The four graphs show analysis by a fixed effects model (M_1), a random effects model (M_2), MR-Egger regression (M_3) and model averaging.

Table 1. Root mean square errors (rmse) of the point estimates (posterior means) that are displayed in figure 4

	Simulated pleiotropy		
Analysis	none	zero-centred	directional
Fixed Effects	0.0099	0.0814	0.5551
Random Effects	0.0099	0.0472	0.5263
MR-Egger	0.0311	0.1977	0.2036
Model average	0.0099	0.0481	0.2853

5.5. Average performance

Figure 4 and Table 1 demonstrate the benefit of the model averaging when there is relatively strong information in the data to help determine which model should be used. In the following simulations we consider the situation in which the amount of zero-centred or directional pleiotropy is small, so that it is harder to determine which model should be used. The results in terms of rmse are given in the supplement as Tables S.1 to S.3. They show the same pattern as that seen in Table 1. MR-Egger has large rmse because of the added uncertainty in its estimates and only becomes optimal in situations where the pleiotropy is strongly directional. Uncertainty in MR-Egger estimates will be greatest when there is little variation in the effects of the variants on the intermediate [14]. With weakly directional pleiotropy it is better to use a random effects meta-analysis in place of MR-Egger even though it would be biased. Model averaging performs as well as a random effects meta-analysis in these situations and stays close to optimal when the directional pleiotropy is stronger, hence it is the best method to use when there is any uncertainty about the true pattern of pleiotropy.

Table 2 shows the results when the data were simulated under M_1, that is with no pleiotropy and elaborates on the left hand thirds of the graphs in Figure 4. Each dataset was analysed four times, once assuming M_1 (fixed effects assuming no pleiotropy), once under M_2 (random effects assuming zero-centred pleiotropy), once under M_3 (MR-Egger assuming directional pleiotropy) and once using model averaging. We analysed the data either allowing for uncertainty in the estimate of the variant-exposure effect or treating the estimated variant-exposure value as exact (s_j=0). The model averaged analyses spend most of their time in M_1 with occasional visits to M_2 and almost never prefer M_3. Small sample sizes make model determination more difficult so they make it more likely that M_2 will be used even though the data were simulated under M_1. The penalty for using M_2 is seen in a small increase in the posterior standard deviation. When the studies have large samples, allowance for uncertainty in the variant-expsoure effect size makes little difference but the impact is greater when the sample sizes are smaller and acts such that we are less likely to conclude that a random effect analysis is needed because the uncertainty in the variant-exposure effect accounts for some of the observed heterogeneity. Allowing for possible non-zero pleiotropy by using the Bayesian equivalent of MR-Egger would inflate the posterior standard deviation by a factor of about 4.

Table 3 shows the results when the data were simulated under M_2, that is with pleiotropy that cancels so that the average pleiotropy, m_p, is zero. All of these simulations were carried out using the model that adjusts for uncertainty in the variant-exposure effect sizes. With small pleiotropic variance, s_p=0.005, MAPR=2%, the model averaging has difficulty distinguishing between M_1 and M_2 particularly with small sample sizes but the posterior distributions of β under M_1 and M_2 are very similar. Medium pleiotropy was defined to have s_p=0.02 and this makes MAPR equal to about 10% and this was large enough to lead the analysis to favour M_2. Large pleiotropy was defined to have s_p=0.04 so that MAPR was about 19%. Now the analysis detects M_2 with almost complete certainty but is rarely drawn into thinking that non-zero pleiotropy (M_3) might be appropriate. Although assuming no pleiotropy will not lead to bias in these scenarios, it will lead to an under-estimate of the uncertainty.

Table 4 shows the results when the data are simulated under M_3, that is with pleiotropy that does not cancel so that the average pleiotropy is non-zero. A small pleiotropic mean and variance was defined as m_p=0.01 and s_p=0.005; this meant that the MAPR was 7%. Large samples favoured M_2 but occasionally visited M_1 and M_3, while with small sample sizes the analyses were drawn toward M_1 because they could not be sure of either the non-zero effect or the heterogeneity. The posterior standard deviation under MR-Egger regression (M_3) was around 4 times that under the fixed or random effects models. A small pleiotropic mean and medium variance was defined as m_p=0.01 and s_p=0.02, which increased MAPR to 11%. The preference for M_2 over M_1 increased but once again the analysis rarely considered MR-Egger. A large pleiotropic mean and small variance was defined as m_p=0.04 and s_p=0.005, which meant a MAPR of 27%. In large samples the analysis was able to detect the need for M_3 but in small samples the non-zero average pleiotropy was not detected and the small variance led the analysis towards M_1. A large pleiotropic mean and medium variance was defined as m_p=0.04 and s_p=0.02, which meant a MAPR of 27% just as is was with large mean and small variance. The increased variance has the effect of making it harder for the analysis to determine that the intercept is non-zero and thus that M_3 is the underlying model.

Table 2. Results for 100 datasets simulated without pleiotropy (under M_1) and analysed with different assumptions about the pleiotropy. mean=average posterior mean, sd=average posterior standard deviation, the model averaged results include the posterior belief in each model.

	Analysis Model										
	Fixed effects		Random effects		MR-Egger		Model Average				
β	mean	sd	mean	sd	mean	sd	M_1	M_2	M_3	mean	sd
Without adjusting for variant-exposure uncertainty											
Small samples (n=5,000)											
0.0	0.00	0.029	0.00	0.029	0.00	0.095	90%	10%	0%	0.00	0.029
0.5	0.48	0.028	0.48	0.029	0.31	0.092	70%	27%	3%	0.47	0.045
Large samples (n=50,000)											
0.0	0.00	0.009	0.00	0.010	0.01	0.040	98%	2%	0%	0.00	0.009
0.5	0.50	0.009	0.50	0.010	0.47	0.040	89%	11%	0%	0.50	0.010
Adjusting for variant-exposure uncertainty											
Small samples (n=5,000)											
0.0	0.00	0.030	0.00	0.030	0.02	0.140	90%	10%	0%	0.00	0.031
0.5	0.50	0.032	0.50	0.033	0.46	0.141	89%	11%	0%	0.50	0.033
Large samples (n=50,000)											
0.0	0.00	0.009	0.00	0.010	0.01	0.039	99%	1%	0%	0.00	0.009
0.5	0.50	0.010	0.50	0.011	0.49	0.041	98%	2%	0%	0.50	0.010

5.6. *Impact of the priors*

The effect of changing the prior model probabilities can be calculated without further simulations. For instance, if the analysis with equal prior model probabilities 0.33:0.33:0.33 were to lead to posterior model probabilities 0.95:0.05:0.00 then the Bayes factor, BF$_{12}$, would be 19. Changing the prior probabilities to 0.1:0.5:0.4 would not alter the Bayes Factor so the ratio of posterior model probabilities would be 19x0.1/0.5=3.8, this would correspond to posterior model probabilities of $P(M_1|D)$=0.79 and $P(M_2|D)$=0.21. The model averaged posterior mean estimate could also be calculated from the original analysis as it is a weighted average of the estimates under the different models and these will not change,

$$E\{\beta\} = P(M_1|D)E\{\beta|M_1, D\} + P(M_2|D)E\{\beta|M_2, D\} + P(M_3|D)E\{\beta|M_3, D\}.$$

More difficult to assess is the impact of changes to the priors on the model parameters α, β and τ^2. Changing these priors will impact on the Bayes Factors and hence change the posterior model probabilities as well as the posterior parameter distributions. To investigate this effect we took the scenario described in the highlighted 12^{th} row of Table 4 where the sample size was 50,000, $\beta = 0.5$ and the pleiotropy was generated from N(0.045,0.01) so that MAPR=25%. With vague priors model M_3 is preferred but neither of the other models can be ruled out completely.

In Table 5 the baseline analysis used non-informative priors, that is to say priors that conveyed little information about the exact parameter values but which gave little support to extreme values of the parameters. We used N(0,sd=10) priors for α and β and a G(2,0.00005) prior for the precision, which implies a mean precision of 40,000, corresponding to a typical standard deviation for the pleiotropy of 0.005. We contrasted these results with those obtained when more informative priors were used. Taking one parameter at a time we used an informative prior that was either correctly

Table 3. Results for data simulated with zero-centred pleiotropy (M_2) and analysed with different assumptions about the pleiotropy. mean=average posterior mean, sd=average posterior standard deviation, the model averaged results include the posterior belief in each model.

	Analysis Model										
	Fixed effects		Random effects		MR-Egger		Model Average				
β	mean	sd	mean	sd	mean	sd	M_1	M_2	M_3	mean	sd
Small pleiotropic variance: MAPR=2%											
Small samples (n=5,000)											
0.0	0.00	0.029	0.00	0.030	0.01	0.143	50%	49%	0%	0.00	0.032
0.5	0.50	0.032	0.50	0.033	0.46	0.145	51%	49%	0%	0.49	0.036
Large samples (n=50,000)											
0.0	0.00	0.009	0.00	0.011	0.00	0.038	44%	56%	0%	0.00	0.010
0.5	0.50	0.010	0.50	0.011	0.51	0.041	47%	53%	0%	0.50	0.012
Medium pleiotropic variance: MAPR=10%											
Small samples (n=5,000)											
0.0	0.00	0.029	0.00	0.031	0.03	0.161	38%	60%	1%	0.00	0.040
0.5	0.50	0.032	0.50	0.034	0.56	0.162	42%	57%	1%	0.49	0.044
Large samples (n=50,000)											
0.0	0.00	0.009	0.00	0.020	0.00	0.048	0%	99%	1%	0.00	0.023
0.5	0.51	0.010	0.51	0.020	0.63	0.049	0%	99%	1%	0.50	0.024
Large pleiotropic variance: MAPR=19%											
Small samples (n=5,000)											
0.0	0.00	0.030	0.00	0.044	-0.03	0.239	3%	93%	3%	0.00	0.083
0.5	0.51	0.033	0.49	0.037	1.01	0.238	6%	88%	6%	0.43	0.096
Large samples (n=50,000)											
0.0	0.00	0.010	0.00	0.037	0.02	0.115	0%	99%	1%	0.00	0.043
0.5	0.52	0.011	0.50	0.037	0.52	0.115	0%	99%	1%	0.50	0.046

centred, centred so as to under-estimate the true parameter value, or to over-estimate it. The informative priors are shown in the supplement as Figures S.1 to S.3.

The informative priors only operate under some of the models, for instance the prior on μ is only relevent to model M_3. As a result choosing an informative prior can make one of the models seem less appropriate and hence affect the model selection as well as the parameter estimates.

6. Age at menarche and lung function

This example concerns the effect of age at menarche, X, measured in years on lung function in particular on forced vital capacity (FVC) Y, measured in ml, a spirometric measure of lung restriction. Sex hormones are known to affect lung function, and there has been interest in assessing whether the timing of sexual development can influence lung function later on in life. In women, observational evidence has suggested an association between earlier menarche and lower FVC in adulthood, but confounding could not be ruled out [43]. Based on the findings of a GWA meta-analysis on 182,416 women from 57 studies published by Perry et al. [44], 122 SNPs were identified that have a genome-wide significant effect on age at menarche. We then examined the effect of those SNPs on FVC using lung function GWA data on 2,829 adolescent women aged 14-17 years from two studies, the Avon Longitudinal Study of Parents and Children (ALSPAC) [45] and Northern Finland Birth cohort of 1986 (NFBC 1986) [46], and separately analysed the effect of those SNPs on adult lung function using data on 46,924 adult women aged 27-69 years from three studies, the European Community

Table 4. Results for data simulated under a directional pleiotropy model (M_3) and analysed with different assumptions about the pleiotropy. mean=average posterior mean, sd=average posterior standard deviation, the model averaged results include the posterior belief in each model. The highlighted simulation in row 10 is investigated further in Table 4.

	Analysis Model										
	Fixed effects		Random effects		MR-Egger			Model Average			
β	mean	sd	mean	sd	mean	sd	M_1	M_2	M_3	mean	sd
Small pleiotropic mean, small variance: MAPR=7%											
Small samples (n=5,000)											
0.0	0.07	0.030	0.07	0.030	0.01	0.138	50%	49%	0%	0.06	0.032
0.5	0.56	0.033	0.56	0.034	0.45	0.140	51%	49%	1%	0.55	0.040
Large samples (n=50,000)											
0.0	0.06	0.009	0.06	0.011	0.00	0.039	37%	61%	2%	0.06	0.014
0.5	0.56	0.010	0.56	0.012	0.51	0.041	47%	52%	1%	0.56	0.015
Small pleiotropic mean, medium variance: MAPR=11%											
Small samples (n=5,000)											
0.0	0.07	0.030	0.06	0.031	0.05	0.160	38%	61%	0%	0.06	0.040
0.5	0.57	0.033	0.57	0.034	0.54	0.158	43%	56%	1%	0.55	0.044
Large samples (n=50,000)											
0.0	0.06	0.009	0.06	0.020	-0.02	0.046	0%	99%	1%	0.06	0.026
0.5	0.57	0.011	0.56	0.020	0.61	0.050	0%	98%	2%	0.56	0.027
Large pleiotropic mean, small variance: MAPR=27%											
Small samples (n=5,000)											
0.0	0.25	0.031	0.25	0.031	-0.02	0.138	48%	49%	3%	0.23	0.056
0.5	0.75	0.036	0.75	0.037	0.46	0.144	48%	44%	7%	0.71	0.077
Large samples (n=50,000)											
0.0	0.25	0.010	0.25	0.013	0.00	0.039	0%	5%	95%	0.01	0.047
0.5	0.75	0.011	0.75	0.014	0.50	0.042	**6%**	**11%**	**83%**	**0.53**	**0.061**
Large pleiotropic mean, medium variance: MAPR=27%											
Small samples (n=5,000)											
0.0	0.26	0.031	0.26	0.033	-0.01	0.163	36%	57%	7%	0.22	0.075
0.5	0.76	0.037	0.76	0.038	0.53	0.152	44%	51%	5%	0.72	0.068
Large samples (n=50,000)											
0.0	0.25	0.010	0.25	0.022	-0.01	0.047	0%	70%	30%	0.15	0.076
0.5	0.76	0.012	0.75	0.023	0.61	0.050	0%	74%	26%	0.66	0.072

Respiratory Health Survey (ECRHS) [47], Northern Finland Birth cohort of 1966 (NFBC 1966) [48], and UK Biobank [49]. Mendelian randomization analyses were performed separately in adolescents and adults. Details of this Mendelian randomization study are reported elsewhere [50].

6.1. Adolescent women

Figure 5 shows the diagnostic plots for the data on adolescent women. The trace plots are shown in Figure S.4 of the supplement. The chain spends 86% of its time in model M_1 suggesting that there is little indication of pleiotropy. It does however spend 10% of the iterations in M_2 and 3% in M_3 so some pleiotropy cannot be ruled out. Under M_1 the posterior mean and standard deviation for the causal age at menarche-lung function effect are -55.2 and 25.3 and under M_2 the corresponding figures are very similar at -55.2 and 25.5. However under M_3 the estimated effect size drops to -21.5 and the posterior standard deviation increases considerably to 64.2. The model averaged result has a posterior mean of -54.1 with a standard deviation of 28.2.

Table 5. Impact of the priors. Averages across the 100 datasets corresponding to row 10 of Table 4 when an informative prior is placed on one of the parameters. mean=average posterior mean, sd=average posterior standard deviation.

parameter	M_1	M_2	M_3	mean	sd
Non-informative priors					
	6%	11%	83%	0.53	0.061
Correctly centred informative prior					
β	2%	3%	94%	0.51	0.050
μ	19%	77%	4%	0.74	0.018
τ^{-2}	30%	6%	64%	0.58	0.079
Informative prior that underestimates the parameter					
β	0%	0%	100%	0.41	0.044
μ	4%	11%	85%	0.70	0.032
τ^{-2}	100%	0%	0%	0.75	0.012
Informative prior that overestimates the parameter					
β	17%	55%	28%	0.70	0.048
μ	20%	80%	0%	0.75	0.014
τ^{-2}	11%	12%	76%	0.54	0.069

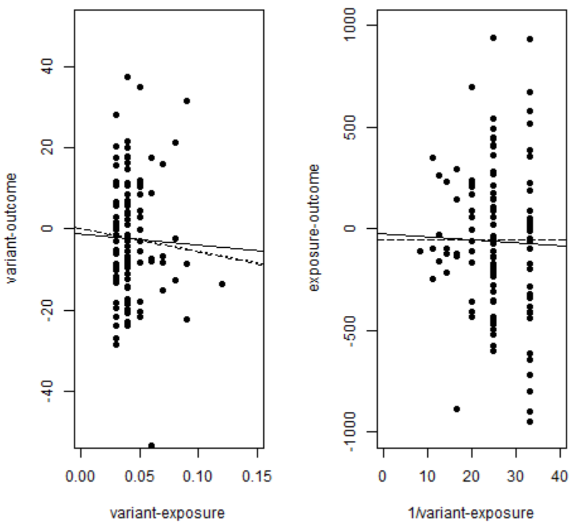

Figure 5. Diagnostic plots for adolescent women. The solid line shows the model fit under M_3 and the dashed line shows the fit under M_2 or M_1.

6.2. Adult women

Figure 6 shows the diagnostic plots for the data on adult women. The trace plots are shown in Figure S.5 in the supplement. The chain never visits M_1 suggesting strongly that there is pleiotropy. It spends 81% of the iterations in M_2 and 19% in M_3 so it favours pleiotropy that cancels but cannot rule out a directional effect. Under M_2 the posterior mean and standard deviation for the causal age at menarche-lung function effect are 19.9 and 14.4, while under M_3 the corresponding figures are -3.6 and 27.1. The model averaged result has a posterior mean of 15.5 with a standard deviation of 19.8.

6.3. Conclusion

Taken together these data suggest a negative causal effect of later age at menarche on lung function in adolescents and a positive effect in adults, although uncertainty over pleiotropy creates large uncertainty in the latter estimate. These

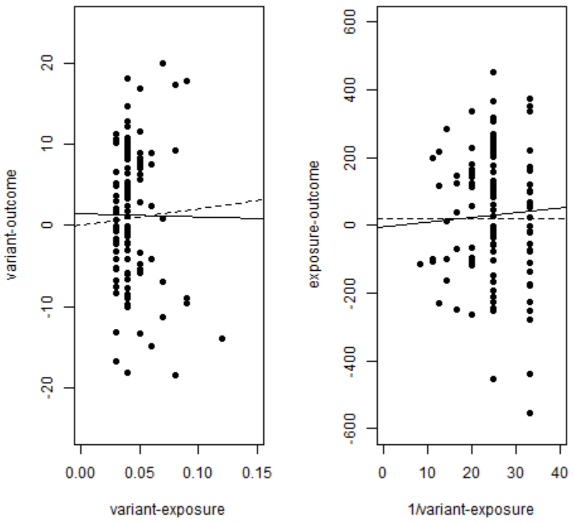

Figure 6. Diagnostic plots for adult women. The solid line shows the model fit under M_3 and the dashed line shows the fit under M_2.

findings confirm the patterns found in the same data by Gill and colleagues [50] who performed the MR analysis after excluding possibly pleiotropic SNPs (SNPs previously associated with secondary phenotypes that affect lung function) and showed strong evidence of a positive effect (44.6ml; 95%CI: 17.2 to 69.9; p=0.001).

This example illustrates the importance of time in an MR. Both the variant-age at menarche (GX) and the variant-lung function (GY) estimates relate to life-time effects of the genes, but X is observed in the early teenage years while in our two analyses the outcome Y is observed later in adolsecence or in adulthood. Not only might the effect of X on Y differ depending on the ages of the subject in the study, but it is perfectly possible that the pleiotropic effects will also vary with age, which could explain why we see more pleiotropy in the analysis of adult lung function. It is always important to consider the ages of the subjects when interpreting a two-sample MR.

7. Discussion

Mendelian randomization (MR) gives us the potential to draw causal conclusions from observational data. Unfortunately MR replaces traditional epidemiological assumptions, such as no unmeasured confounding and no reverse causation, with other assumptions, most importantly the assumption of no pleiotropy, that is to say, the genetic instruments must not affect the outcome through any pathway that does not pass through the exposure of interest. Like the assumption of no unmeasured confounding, the assumption of no pleiotropy cannot be justified based on the observational data alone but requires external knowledge of the problem.

Two-sample MR requires its own extra assumptions. First it assumes that the variant-exposure effects measured in the first sample will be valid estimates of the unmeasured variant-exposure effects in the second sample. Any gene-environment interaction would have the potential to cause this assumption to be violated. Second, although the design of an MR offers some protection against reverse causation, external knowledge of the time ordering between the exposure measured in one sample and the outcome measured in the other is usually required. This time ordering may not be straightforward, especially when the exposures or outcomes are the result of life-long influences. In our example, menarche clearly precedes adult lung function but lung function can be seen as a life-long process with early lung development in the womb, growth during childhood and adolescence, a peak in lung performance in early adulthood and then a gradual decline. Genes that affect early lung development will almost inevitably have some impact on adult lung function; were it the case that early lung development affects age at menarche then early lung development would be a confounder between exposure and outcome and it is a key assumption of any MR that the chosen genetic variants do not affect confounders. This assumption would be violated and the MR would be invalid. It is tempting to obtain genome-wide association estimates from two-samples and to automatically apply MR, but this is dangerous and potentially very misleading unless one has a deep understanding of the structure of the relationships between genetic variants, exposure and outcome.

A two-sample MR can be viewed as a meta-analysis over the set of genetic instruments and depending on what is believed about pleiotropy, this meta-analysis can be undertaken in different ways. A fixed effects meta-analysis is appropriate when there is no pleiotropy, a random effects meta-analysis is appropriate when the pleiotropy cancels in the sense that the average over all instruments is zero and MR-Egger is best when the pleiotropy is directional. We have placed these three methods within a Bayesian framework and used model averaging to provide estimates that allow for our uncertainty about the nature of the pleiotropy.

In creating a Bayesian analysis that encompasses the three models for pleiotropy, we have pointed out a slight difference between the Bayesian model with non-directional pleiotropy (M_2) and a random effects meta-analysis (M_{2a}) in that the former assumes that the expected pleiotropy is zero and requires the InSIDE assumption, while the latter assumes that the expected ratio between the pleiotropy and the effect on the exposure is zero. It is dificult to imagine circumstances in which $E\left\{\psi_j/\gamma_j\right\}$ would be zero unless it were the case that the two assumptions of M_2 held, so in practice, the difference is unlikely to be important.

The root mean square errors (rmse) is a measure that combines bias and variance and the benefits of model averaging in terms of rmse were evident in all of our simulations, see Tables 2, S.1 to S.3. Clearly we could out-perform model averaging if we knew the correct model, but we never do, and analyses that assume a single model will underestimate the full uncertainty in the results. It is an attractive feature of model averaging that when the evidence points strongly to one of the models, then the model average results will move close to that solution, but when there is no strong evidence to favour one particular model then the analysis will correctly capture that uncertainty. However, despite its useful properties, it is well to remember that model averaging only averages over a specified list of alternative models. Were it the case that none of the models in the list adequately captures the relationships in the data, then performance could be poor.

One of the aspects of Bayesian methods that many people find difficult is the specification of the priors; the results will depend not just on the priors for the models but also on the priors on the parameters of those models. It is common practice to adopt non-informative priors so as to allow the data to dominate the solution. There are two arguments against this in the context of MR; first, the analysis is already subjective because of the choices that are made when selecting the variants and picking the models over which we will average, and second, the data alone are unlikely to be sufficient to provide useful estimates. Particularly in the case of MR-Egger regression, the estimates provided by MR are very imprecise and may be unstable due to impact of unusual estimates from a few variants. As Table 5 shows, well-informed prior distributions can help overcome these limitations although, of course, very poorly informed prior knowledge could make things worse. It is important that the researchers treat the specification of prior distributions as an integral part of the study and only use prior distributions that they really believe in. Anyone reading a report of an MR study, whether it is Bayesian or not, should pay particular attention to the assumptions made by the researchers and interpret the findings in the light of those assumptions. Often Bayesian analyses are clearer because they make those assumptions explicit.

The Bayesian approach offers two advantages beyond the quantification of model uncertainty. First we can incorporate prior beliefs about the model parameters and even when these are relatively vague they can still act to limit the range of the posterior and avoid some of the instability that is associated with weak instruments. Second, the Bayesian approach offers a simple way for allowing for the uncertainty associated with the variant-exposure estimates, although this is not an automatic process and some thought needs to go into modelling of the distribution of the variant-exposure effects. For instance, it may not be reasonable to assume normally distributed variant effects when the estimates come from a discovery sample that is subject to the Winner's curse [51].

The model averaging algorithm is required to move freely between the three models but if the models differ because the posterior of the causal estimate is very different under two of the models then it may be difficult to get the chain to jump from one model to the other. The result would be poor mixing and a very long chain would be required in order to achieve convergence, especially for the posterior model probabilities. The mixing of the algorithm used in this paper is controlled by the choice of the pseudo priors, that is the distributions from which we draw values for parameters not in the current model. These pseudo priors need to be chosen with care so that they generate values appropriate to the other model options and avoid values that would make some models a poor fit to the data. Other algorithms that could be used for model averaging have been reviewed by Han and Carlin [52] and might be worth investigating in the context of MR.

The approach adopted in this paper could be extended in various ways. Obviously we could incorporate other model choices, for example we could have a variation on MR-Egger that allows a non-zero average pleiotropy but no excess heterogeneity, this would correspond to a fixed amount of pleiotropy affecting all of the variants. We decided not to include this possibility in our analyses because it seems biologically implausible. Bowden et al. [14] have noted the advantages of using a multiplicative random effects model that assumes zero-centred pleiotropy but is more robust to directional pleiotropy and this could also be included as an option in the Bayesian model averaging. Yet another option would be to allow different subsets of the genetic variants to have different patterns of pleiotropy, so for example, some genetic variants might not exhibit any pleiotropy so that M_1 would be appropriate for them, while other genetic variants might be subject to zero-centred pleiotropy and require M_2.

Another interesting extension of model averaging would be for one-sample MR. In a one-sample study one typically has access to the individual participant data (IPD) and so it would not be necessary to model the summary statistics. A Bayesian IPD model of a large number genetic variants is perfectly possible but would be computationally very demanding, so that it would be tempting to model the summary statistics as an approximation. Unfortunately summary statistics for GX and GY calculated on the same subjects would not be independent and the model would need to be adjusted to account for the correlation. This explains why the meta-analysis methods and MR-Egger regression were developed for two-sample studies.

Since Mendelian randomization requires untestable assumptions it will never be a method that is completely defined by the data, but rather it will always require external, biological information about the genetic instruments. Unfortunately MR is not a black box analysis that can be conducted by downloading GWAS data on variant-exposure and variant-outcome and combining them in some automatic manner. MR without biology is dangerous and has the potential to be very misleading; given this, it is natural to adopt a Bayesian approach to the analysis that enables our biological knowledge to be integrated with the data.

References

1. Lawlor D, Harbord R, Sterne J, Timpson N, Smith GD. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**:1133–1163.
2. Didelez V, Sheehan N. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**:309–330.
3. Stearns F. One hundred years of pleiotropy: a retrospective. *Genetics* 2010; **186**:767–773.
4. Freeman G, Cowling B, Schooling C. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *International Journal of Epidemiology* 2013; **42**:1157–1163.
5. Brion M, Shakhbazov K, Visscher P. Calculating statistical power in Mendelian randomization studies. *International Journal of Epidemiology* 2013; **42**:1497–1501.
6. Johnson T. Efficient calculation for multi-SNP genetic risk scores. *Technical Report*, The Comprehensive R Archive Network 2013. Available at http://cran.r-project.org/web/packages/gtx/vignettes/ashg2012.pdf [last accessed 2014/11/19].
7. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* 2013; **42**(4):1134–1144.
8. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, *et al.*. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* 2014; **46**(11):1173–1186.
9. Inoue A, Solon G. Two-sample instrumental variables estimators. *The Review of Economics and Statistics* 2010; **92**:557–561.
10. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* 2013; **37**(7):658–665.
11. Thompson JR, Minelli C, Del Greco M F. Mendelian randomization using public data from genetic consortia. *The International Journal of Biostatistics* 2016; .
12. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology* 2016; **45**(3):908–915.
13. Del Greco M F, Minelli C, Sheehan N, Thompson J. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine* 2015; **34**:2926–2940.
14. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* 2017; **doi:10.1002/sim.7221**.
15. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* 2015; **44**:512–525.
16. Faraway JJ. On the cost of data analysis. *Journal of Computational and Graphical Statistics* 1992; **1**(3):213–229.
17. Shen X, Huang HC, Ye J. Inference after model selection. *Journal of the American Statistical Association* 2004; **99**(467):751–762.
18. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science* 1999; **14**:382–401.
19. Hansen BE. Least squares model averaging. *Econometrica* 2007; **75**(4):1175–1189.
20. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* 2009; **10**(10):681–690.
21. Valdar W, Sabourin J, Nobel A, Holmes CC. Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genetic Epidemiology* 2012; **36**(5):451–462.
22. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology* 2009; **169**(4):497–504.
23. Ferreira T, Marchini J. Modeling interactions with known risk locia Bayesian model averaging approach. *Annals of Human Genetics* 2011; **75**(1):1–9.
24. Sebastiani P, Xie H, Ramoni MF, *et al.*. Bayesian analysis of comparative microarray experiments by model averaging. *Bayesian Analysis* 2006; **1**(4):707–732.
25. Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* 2001; **20**(21):3215–3230.
26. Bayarri MJ, Berger JO. The interplay of Bayesian and frequentist analysis. *Statistical Science* 2004; **19**:58–80.
27. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association* 1994; **89**(428):1535–1546.
28. Nagar A. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica: Journal of the Econometric Society* 1959; :575–595.
29. Ginestet CE, Emsley R, Landau S. Dose–response modeling in mental health using Stein-like estimators with instrumental variables. *Statistics in Medicine* 2017; **36**(11):1696–1714.
30. Borenstein M, Hedges LV, Higgins J, Rothstein HR. Generality of the basic inverse-variance method. *Introduction to meta-analysis* 2009; :311–319.
31. Thomas DC, Lawlor DA, Thompson JR. re: Estimation of bias in nongenetic observational studies using Mendelian triangulation by bautista et al. *Annals of Epidemiology* 2007; **17**(7):511–513.

32. Burgess S, Bowden J, Dudbridge F, Thompson SG. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv preprint arXiv:1606.03729* 2016; .
33. Minelli C, Thompson JR, Tobin MD, Abrams KR. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology* 2004; **160**(5):445–452.
34. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomizationa multivariate approach. *Statistics in Medicine* 2005; **24**(14):2241–2254.
35. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**:629–634.
36. Burgess S, Butterworth A, Thompson S. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* 2013; **37**:658–665.
37. Bowden J, M FDG, Minelli C, Smith GD, Sheehan N, Thompson J. Assessing the suitability of summary data for Mendelian randomization analyses using mr-egger regression: the role of the i-squared statistic. *International Journal of Epidemiology* 2016; **45**:1961–1974.
38. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis, (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2003.
39. Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; **57**:473–484.
40. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**(430):773–795.
41. Irony TZ, Singpurwalla ND. Non-informative priors do not exist a dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference* 1997; **65**(1):159–177.
42. Thompson J. *Bayesian analysis with Stata*. Stata Press, 2014.
43. Macsali F, Real FG, Plana E, Sunyer J, Anto J, Dratva J, Janson C, Jarvis D, Omenaas ER, Zemp E, *et al.*. Early age at menarche, lung function, and adult asthma. *American Journal of Respiratory and Critical Care Medicine* 2011; **183**(1):8–14.
44. Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, He C, Chasman DI, Esko T, Thorleifsson G, *et al.*. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 2014; **514**(7520):92–97.
45. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Smith GD. Cohort profile: the children of the 90sthe index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 2012; **42**:111–124.
46. Jääskeläinen A, Schwab U, Kolehmainen M, Kaakinen M, Savolainen MJ, Froguel P, Cauchi S, Järvelin MR, Laitinen J. Meal frequencies modify the effect of common genetic variants on body mass index in adolescents of the northern Finland birth cohort 1986. *PLoS One* 2013; **8**(9):e73 802.
47. Burney P, Luczynska C, Chinn S, Jarvis D. The european community respiratory health survey. *European Respiratory Journal* 1994; **7**(5):954–960.
48. Rantakallio P. The longitudinal study of the northern finland birth cohort of 1966. *Paediatric and Perinatal Epidemiology* 1988; **2**(1):59–88.
49. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, *et al.*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* 2015; **12**(3):e1001 779.
50. Gill D, Sheehan N, Wielscher M, Shrine N, Amaral A, Thompson J, Granell R, Leynaert B, Gomez-Real F, Hall I, *et al.*. Age at menarche and lung function: a Mendelian randomization study. *European Journal of Epidemiology* 2017; **doi:20.2007/s10654-017-0272-9**.
51. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nature Genetics* 2001; **29**(3):306–309.
52. Han C, Carlin BP. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association* 2011; **96**:1122–1132.

8. Online Supplement

8.1. OpenBUGS code

The OpenBUGS code given below performs model averaging allowing for uncertainty in the variant-exposure effect sizes as in Tables 3 and 4 of the main paper. The code is written in terms of the precision so that t in the code represents the reciprocal of the variance and itau2=$1/\tau^2$. T is the model indicator that can take values 1,2,3. The data supplied to the program contains N=number of instruments, x[] and y[] , containing the estimates $\hat{\gamma}_j$, $\hat{\Gamma}_j$ respectively, pg[] containing $1/s_j^2$ and vG[] containing S_j^2, the prior model probabilities, p[], and the parameters of the priors.

Under all three models we assume a normal prior for β with mean MB and precision PB. Under M_3 the intercept mu has a normal prior with mean PA and precision PB. Under M_1 and M_2 the intercept is zero so the values of mu are generated from a normal pseudo prior with mean zero and precision 100; this helps mixing between models since these values correspond to the approximate scale on which the simulations were created. The pleiotropy terms under M_2 and M_3 have a precision itau2 that is given a G(GA,GB) prior.

```
model {
   T ~ dcat(p[])
   for( i in 1:N ) {
      z[i] ~ dnorm(m,t)
      x[i] ~ dnorm(z[i],pg[i])
      m[i] <- mu*equals(T,3) + beta*z[i]
      pr[i] <- 1/(vG[i]+step(T-1.5)/itau2)
      y[i] ~ dnorm(m[i],pr[i])
   }
   ta <- 100 - (100-PA)*equals(T,3)
   mu ~ dnorm(MA,ta)
   beta ~ dnorm(MB,PB)
   at <- 1*equals(T,1) + GA*step(T-1.5)
   bt <- 0.001*equals(T,1) + GB*step(T-1.5)
   itau2 ~ dgamma(at,bt)
   m ~ dnorm(0.2,0.1)
   t ~ dgamma(0.1,0.001)
}
```

The priors on m and t need to be chosen to be appropriate for the particular problem. To remove the adjustment for uncertainty in the measurement of the gene-intermediate estimates the priors on m and t should be dropped and the loop should be replaced by,

```
   for( i in 1:N ) {
      m[i] <- mu*equals(T,3) + beta*x[i]
      pr[i] <- 1/(vG[i]+step(T-1.5)/itau2)
      y[i] ~ dnorm(m[i],pr[i])
   }
```

8.2. *Mixed simulation*

Data for the mixed simulation shown in Figure 4 were generated so that there was an equal probability of no pleiotropy (M_1), zero-centred pleiotrop (M_2) or non zero-centred pleiotropy (M_3). The true value of the effect of X on Y, β, was fixed to be 0.5. The effects of the genetic variants on X were generated from a Normal(0.2,0.05) distribution. Under M_2 and M_3 the precision of the pleiotropy was generated from a Gamma(10,0.04) distribution and under M_3 the mean pleiotropy was generated from uniform(-0.2,0.2) distribution. Two-sample data were generated with both sample sizes equal to 50,000.

8.3. *Average performance*

Tables showing the root mean square errors corresponding to the simulations in Tables 2, 3 and 4.

Table S.1. Root mean square errors for 100 datasets simulated without pleiotropy (under M_1). Corresponding to Table 2

	Analysis Model			
β	Fixed effects	Random effects	MR-Egger	Model Average
Without adjusting for variant-exposure uncertainty				
Small samples (n=5,000)				
0.0	0.029	0.029	0.094	0.029
0.5	0.034	0.034	0.213	0.051
Large samples (n=50,000)				
0.0	0.010	0.010	0.039	0.010
0.5	0.011	0.010	0.051	0.010
Adjusting for variant-exposure uncertainty				
Small samples (n=5,000)				
0.0	0.030	0.030	0.134	0.030
0.5	0.028	0.028	0.134	0.038
Large samples (n=50,000)				
0.0	0.010	0.010	0.041	0.010
0.5	0.011	0.011	0.043	0.011

Table S.2. Root mean square errors for 100 datasets simulated with zero-centred pleiotropy (under M_2). Corresponding to Table 3

	Analysis Model			
β	Fixed effects	Random effects	MR-Egger	Model Average
Small pleiotropic variance: MAPR=2%				
Small samples (n=5,000)				
0.0	0.033	0.032	0.168	0.027
0.5	0.036	0.036	0.147	0.032
Large samples (n=50,000)				
0.0	0.012	0.012	0.040	0.011
0.5	0.010	0.010	0.046	0.011
Medium pleiotropic variance: MAPR=10%				
Small samples (n=5,000)				
0.0	0.035	0.035	0.216	0.040
0.5	0.040	0.040	0.194	0.040
Large samples (n=50,000)				
0.0	0.020	0.020	0.103	0.022
0.5	0.025	0.022	0.167	0.023
Large pleiotropic variance: MAPR=19%				
Small samples (n=5,000)				
0.0	0.051	0.049	0.761	0.118
0.5	0.050	0.048	0.736	0.176
Large samples (n=50,000)				
0.0	0.043	0.039	0.160	0.042
0.5	0.046	0.039	0.156	0.042

Table S.3. Root mean square errors for 100 datasets simulated with directional pleiotropy (under M_3). Corresponding to Table 4

	Analysis Model			
β	Fixed effects	Random effects	MR-Egger	Model Average
Small pleiotropic mean, small variance: MAPR=7%				
Small samples (n=5,000)				
0.0	0.072	0.071	0.150	0.064
0.5	0.073	0.071	0.137	0.061
Large samples (n=50,000)				
0.0	0.063	0.063	0.041	0.061
0.5	0.066	0.065	0.044	0.062
Small pleiotropic mean, medium variance: MAPR=11%				
Small samples (n=5,000)				
0.0	0.076	0.076	0.213	0.067
0.5	0.082	0.080	0.181	0.069
Large samples (n=50,000)				
0.0	0.066	0.065	0.120	0.062
0.5	0.073	0.066	0.156	0.062
Large pleiotropic mean, small variance: MAPR=27%				
Small samples (n=5,000)				
0.0	0.257	0.257	0.152	0.240
0.5	0.249	0.248	0.137	0.232
Large samples (n=50,000)				
0.0	0.253	0.252	0.043	0.064
0.5	0.254	0.252	0.045	0.097
Large pleiotropic mean, medium variance: MAPR=27%				
Small samples (n=5,000)				
0.0	0.262	0.261	0.215	0.235
0.5	0.268	0.266	0.166	0.242
Large samples (n=50,000)				
0.0	0.255	0.251	0.111	0.200
0.5	0.265	0.255	0.157	0.207

8.4. *Impact of the priors*

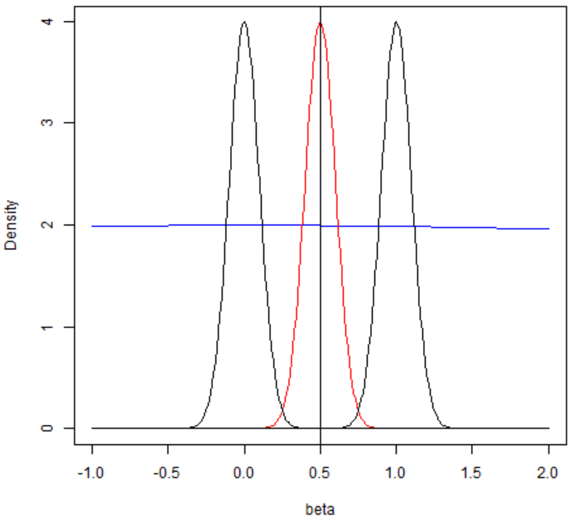

Figure S.1. Realistically vague prior for β times 50 (blue), correctly centred informative prior (red), over and under centred priors (black). True parameter value shown as a vertical line.

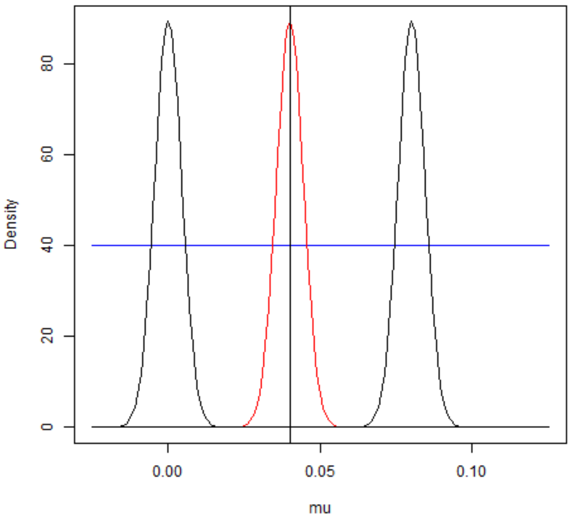

Figure S.2. Realistically vague prior for μ times 1000 (blue), correctly centred informative prior (red), over and under centred priors (black). True parameter value shown as a vertical line.

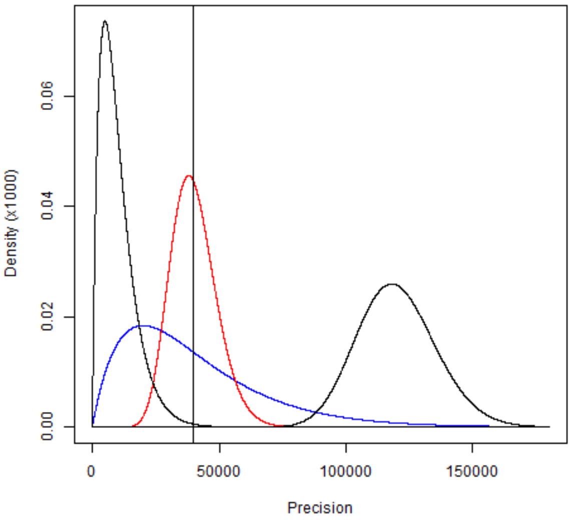

Figure S.3. Realistically vague prior for the precision (blue), correctly centred informative prior (red), over and under centred priors (black). True parameter value shown as a vertical line.

8.5. *Trace plots for age at menarche and lung function*

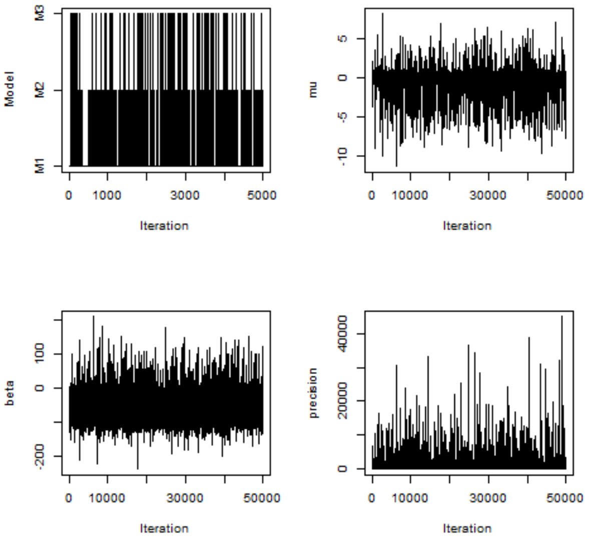

Figure S.4. Trace plots for adolescent women.

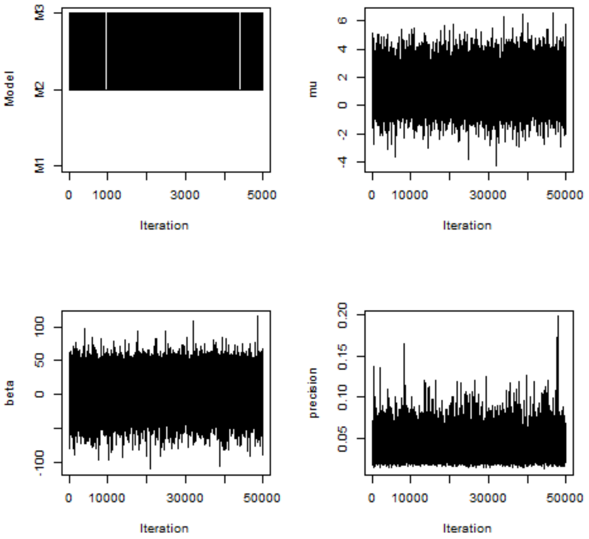

Figure S.5. Trace plots for adult women.