On the Choice of Parameterisation and Priors for the Bayesian Analyses of Mendelian Randomisation Studies.

E. M. Jones[1], J. R. Thompson[1], V. Didelez[2], and N. A. Sheehan[1]

[1]Department of Health Sciences, University of Leicester, Leicester, U.K.
[2]Department of Mathematics, University of Bristol, Bristol, U.K.

Novemeber 2011

Abstract

Mendelian randomisation is a form of instrumental variable analysis that estimates the causal effect of an intermediate phenotype or exposure on an outcome or disease in the presence of unobserved confounding, using a genetic variant as the instrument. A Bayesian approach allows current knowledge to be incorporated into the analysis in the form of informative prior distributions, and the unobserved confounder can be modelled explicitly. We consider Bayesian methods for Mendelian randomisation in the case where all relationships are linear, and there are no interactions.

A 'full' model in which the unobserved confounder is included explicitly is not completely identifiable, although the causal parameter can be estimated. We compare inferences from this general but non-identified model with a reduced parameter model that is identifiable. We show that, theoretically, additional information about the causal parameter can be obtained by using the non-identifiable full model, rather than the identifiable reduced model, but that this is advantageous only when realistically informative priors are used and when the instrument is weak or the sample size is small. Furthermore, we consider the impact of using 'vague' versus 'informative' priors.

Keywords: Mendelian randomisation; Bayesian analyses; identifiability.

1 Introduction

A randomised controlled trial is the most effective way of eliciting the causal effect of an exposure on a clinical outcome, because the randomisation mechanism will act to ensure that all unmeasured variables are evenly distributed across all strata, at least in large samples. However, it is not always possible or ethical to use randomisation and researchers are often forced to rely on observational data. It can then be difficult to tell whether an observed association, or lack of association, between a possible causal factor, X, and the outcome, Y, is genuine or is the result of unobserved confounders, U. In epidemiology, an increasingly popular way of calculating consistent estimates of causal parameters, in the presence of unobserved confounding, is to use an instrumental variable (IV) analysis in which a third factor, G, is sought such that there is no direct link between G and Y, and G is independent of the confounders. The three conditions that G must satisfy in order to be classed as an IV for the effect of X on Y are:

1. G is independent of the set of all confounders U: $G \perp\!\!\!\perp U$.

2. G is associated with the exposure X: $G \not\!\perp\!\!\!\perp X$

3. G and Y are conditionally independent given X and U: $G \perp\!\!\!\perp Y|(X, U)$.

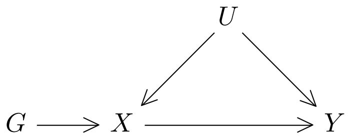

Figure 1: DAG depicting conditional (in)dependencies implied by the core IV conditions where G is an instrument for the causal effect of X on Y.

We use the notation $A \perp\!\!\!\perp B|C$ to denote conditional independence of A and B, given C [1]. The conditional independencies implied by these 'core' conditions are encoded in the unique directed acyclic graph (DAG) represented in Figure 1.

In epidemiological applications, it is sometimes possible to use a genetic variant as the instrument. The term 'Mendelian randomisation' has become widely used for an IV analysis that uses a genetic instrument [2, 3, 4, 5]. It is crucial, however, that the chosen variant satisfies each of the three 'core' conditions [5]. In recent years, numerous genetic studies using single-nucleotide polymorphisms (SNPs) have provided information on the associations between a range of variants and exposures (core condition 2), and since an individual's genes are randomly allocated before birth, they cannot be influenced by the life-style and environmental factors that often confound observational associations (core condition 1). While we can often be reasonably confident that the first two hold, the third condition is difficult to verify, as many genes act on more than one pathway and detailed functional knowledge is not always available. A further important consideration is that for most genetic instruments, the magnitude of the $G-X$ association will be comparatively small; in such a case, the instrument is loosely referred to as 'weak' and estimates are prone to 'weak instrument bias' [6].

We will focus on linear no-interaction models for the exposure and outcome, for which frequentist methods such as two-stage least squares give consistent estimates of the average causal effect of X on Y [7]. In the case of a single instrument, this is equivalent to the 'ratio' or Wald estimator [8], which is based on the regressions of Y on G, and of X on G. However, it is difficult to measure either association with precision when the instrument is weak, and very large sample sizes will be required for any causal inference. An advantage of using a Bayesian approach over frequentist methods, however, is that informative priors, when available, can be used to increase the precision of such estimates.

As the confounder in Figure 1 is unobserved, a key consideration when modelling a Mendelian randomisation study is the identifiability of the parameters. A model that can be parameterised by a k-dimensional vector $\boldsymbol{\theta} \in \mathbb{R}^k$, is non-identifiable if there exist multiple (distinct) vectors $\boldsymbol{\theta}$, each of which corresponds to the same distribution for the data, $\mathbb{D}$. This is the same for frequentist as it is for Bayesian approaches. However, whilst the former apply model restrictions in order to achieve identifiability, the latter also have the option of using prior information [9].

Gustafson [10] discusses the idea of reduced parameter models (model contractions) and model expansion as ways for creating an identifiable structure from a non-identifiable Bayesian model. Model expansion seeks to enlarge the model so that extra data can be included in the analysis, while the more obvious approach of model contraction seeks to simplify the model by removing parameters. The danger with such a simplification is that it can produce precise estimates under the wrong model and hence be misleading. However, the idea of model contraction leads naturally to the notion of a *transparent re-parameterisation* [11]: a bijective transformation $\boldsymbol{\theta} \mapsto (\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$, such that $\boldsymbol{\phi_N} \perp\!\!\!\perp \mathbb{D}|\boldsymbol{\phi_I}$, the model described by the parameter vector $\boldsymbol{\phi_I}$ is identifiable, and no other vector $\boldsymbol{\phi_A}$ can be found such that $\boldsymbol{\phi_N} \perp\!\!\!\perp \mathbb{D}|\boldsymbol{\phi_A}$ with $\dim(\boldsymbol{\phi_A}) < \dim(\boldsymbol{\phi_I})$.

A transparent reparameterisation does not exist for every non-identified model, but when it does, it has the following properties. Under very mild additional conditions [11], the posterior distribution of $\boldsymbol{\phi_I}$ must tend to a point mass at its true value ${\boldsymbol{\phi_I}}^*$ as the sample size grows indefinitely. The posterior conditional distribution of $(\boldsymbol{\phi_N}|\boldsymbol{\phi_I}, \mathbb{D})$ on the other hand is equal to the prior conditional distribution of $(\boldsymbol{\phi_N}|\boldsymbol{\phi_I})$, or in other words $\boldsymbol{\phi_N} \perp\!\!\!\perp \mathbb{D}|\boldsymbol{\phi_I}$. That is, conditional on the information contained in the model parameterised by

$\boldsymbol{\phi_I}$, $\boldsymbol{\phi_N}$ is not identifiable from the data alone [12], and the priors for these parameters will have a heavy influence on any resulting inference [13]. The limiting posterior distribution for $\boldsymbol{\theta}$ as the sample size grows indefinitely, considering that $\boldsymbol{\theta} \mapsto (\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$ is a bijective transformation, therefore consists of a point mass distribution for $\boldsymbol{\phi_I}$ at $\boldsymbol{\phi_I}^*$, and the prior conditional distribution of $(\boldsymbol{\phi_N}|\boldsymbol{\phi_I}^*)$. An obvious corollary of this is that for any quantity of interest in $\boldsymbol{\phi_I}$, the same quantity is also identifiable in the model parameterised by $\boldsymbol{\theta}$.

However in many situations, the fact that $\boldsymbol{\phi_N} \perp\!\!\!\perp \mathbb{D}|\boldsymbol{\phi_I}$ does not preclude some learning about $\boldsymbol{\phi_N}$, unless $\boldsymbol{\phi_N} \perp\!\!\!\perp \mathbb{D}$ [9, 12]. Often, there is some dependence between $\boldsymbol{\phi_N}$ and $\boldsymbol{\phi_I}$ which leads to a posterior marginal for $\boldsymbol{\phi_N}$ that is quite different from its prior. In such situations, $\boldsymbol{\phi_N}$ is *partially identified* [14]. Though the prior and posterior distribution for $\boldsymbol{\phi_N}$ are distinct, it is worth noting that collecting an infinite amount of data would not result in a posterior marginal for $\boldsymbol{\phi_N}$ concentrated entirely at a point.

Structure of Paper

Three approaches to parameterising the likelihood representing the structure in Figure 1 will be considered, each constructed to estimate the average causal effect of X on Y. These are loosely referred to as 'models' in sections 2.1, 2.2 and 2.5, although each is really a 'class' of models consisting of a particular parameterisation and a class of priors. We will be interested in the effects of different priors on these parameterisations. All models will be fitted using Markov chain Monte Carlo (MCMC), in particular using single site Gibbs sampling as implemented in WinBUGS 1.4.3.

In section 2.1, we introduce a model with likelihood built to represent exactly the structure implied by Figure 1. It explicitly models the unobserved confounder U, which renders the model non-identifiable. The likelihood of the second model is a transparent re-parameterisation of the first, and is introduced in section 2.2. We explore the relationship between these in sections 2.3 and 2.4. In section 2.5, a third model is considered, though its likelihood is not a re-parameterisation of the other two. This was used recently to estimate causal relationships in a meta-analysis using Bayesian methods [15]. We explain why, theoretically, all three parameterisations, subject to allocated priors, can yield very different estimates of the causal parameter under particular conditions.

The performance of the three parameterisations is investigated in section 3, where two examples are provided. In the first, simulated data are based on a real study (the Avon Longitudinal Study of Parents And Children) and is designed to show that in certain situations all three parameterisations yield comparable estimates of the causal parameter, and to demonstrate the effect of applying informative versus 'vague' priors. The second example is designed to highlight that all three parameterisations can perform very differently under particular conditions. Finally, we discuss the implications of our findings for the analyses of Mendelian randomisation studies.

2 Methods

Throughout this paper, we assume that the intermediate exposure X and the outcome Y are continuous random variables, and that the genetic instrument G is a discrete random variable taking values in $\{0, 1, 2\}$. The unobserved confounders will be denoted by U. We will let the target parameter of interest be the *average causal effect* (ACE), defined as

$$\mathrm{ACE}(x_1, x_2) = \mathbf{E}[Y|\mathrm{do}(X = x_2)] - \mathbf{E}[Y|\mathrm{do}(X = x_1)],$$

where the *do*-notation represents intervention [16], in this case intervention in X. This is the quantity that would be targeted in a study which randomised X. If we are willing to assume that U is a sufficient set of confounders, that is $\mathbf{E}[Y|\mathrm{do}(X), U] = \mathbf{E}[Y|X, U]$, and we are willing to assume that $\mathbf{E}[Y|X = x, U = u] = b_1x + h(U)$, then $\mathrm{ACE}(x_1, x_2) = b_1(x_1 - x_2)$ [3], so that b_1 becomes the causal parameter of interest.

The parameterisations that we consider assume a bivariate normal likelihood for (X, Y) given G. Indeed, this is one of the drawbacks to taking a Bayesian approach: the likelihood must be fully specified, unlike the classical two-stage-least-squares approach which can be derived from a method of moments argument or from a semiparamtric structural mean model [17]. We will introduce the various ways of parameterising the likelihood in sections 2.1, 2.2 and 2.5, and discuss the choice of priors in section 2.6.

2.1 The 'Full' Model

The likelihood of the 'full' model is built to represent the structure in Figure 1 exactly. We make the additional assumption that $h(U) = b_0 + b_2U$ and assume that X is also linear in G and U, with

$$X = a_0 + a_1G + a_2U + \epsilon_x \tag{1}$$

$$Y = b_0 + b_1X + b_2U + \epsilon_y, \tag{2}$$

where the errors ϵ_x and ϵ_y are zero mean normal random variables with variances τ_x^2 and τ_y^2 respectively. The variables $\{G, U, \epsilon_x, \epsilon_y\}$ are independent of each other. Note that the identity in (2) is structural in that it states how Y reacts to intervention in X.

This structure requires the modelling of the unobserved confounders, U, and here we make the assumption that U is a standard normal random variable. Conditional distributions of U, X and Y are now given by

$$U|G \sim N(0, 1),$$

$$X|(G, U) \sim N(\mu_x, \tau_x^2),$$

$$Y|(X, U) \sim N(\mu_y, \tau_y^2),$$

where $\mu_x = a_0 + a_1G + a_2U$ from equation (1), and $\mu_y = b_0 + b_1X + b_2U$ from the structural equation (2). Recall that our target causal parameter is b_1.

Throughout the paper, a 'full' model will be any Bayesian model with this likelihood parameterisation, and indepenedent priors placed on the eight parameters $\{a_0, a_1, a_2, b_0, b_1, b_2, \tau_x, \tau_y\}$. However, we defer discussion of prior choice until section 2.6.

2.2 The Correlated Errors Model

This particular parameterisation is an alternative to directly modelling the unobserved confounder U, which is also considered in [18, 19]. It is assumed that the error associated with X and Y can instead be modelled by the random variables V and W respectively, where (V, W) is a bivariate normal zero mean random vector. Specifically, let

$$X = a_0 + a_1G + V \tag{3}$$

$$Y = b_0 + b_1X + W, \tag{4}$$

where $V \overset{d}{=} a_2U + \epsilon_x$ and $W \overset{d}{=} b_2U + \epsilon_y$, and $\overset{d}{=}$ denotes equality in distribution. This structure is an elaboration of that in Figure 1 with $U = (V, W)$. The average causal effect is still b_1, since (4) has the same structural meaning as in equation (2). By substituting (3) into (4),

$$Y = b_0 + b_1a_0 + b_1a_1G + W', \tag{5}$$

where $W' = b_1V + W$ and Cov(W', V) =Cov(W, V) + b_1Var$[V]$. The equation in (5) is referred to as the reduced form in the Econometrics literature. Note that while (2) and (4) are structural equations in that they state how Y reacts to intervention in X, (5) is not a structural, but is observationally (distributionally) equivalent to (2) and (4) when we assume (3). The distribution of (X, Y) given G is bivariate normal with Var$[X|G] = \sigma_x^2$, Var$[Y|G] = \sigma_y^2$, and Cov$(X, Y|G) \equiv$ Cov$(V, W') =: \lambda$,

$$[X, Y|G] \sim \text{MVN}\left(\left[\begin{array}{c} a_0 + a_1G \\ b_0 + b_1a_0 + b_1a_1G \end{array}\right], \left[\begin{array}{cc} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{array}\right]\right). \tag{6}$$

We define the correlated errors model by the parameterisation in (6) with independent priors on $\{a_0, a_1, b_0, b_1\}$, and a prior on the covariance matrix which reflects the dependence between $\{\sigma_x, \sigma_y, \lambda\}$, but which is independent of the parameters of the mean structure. Again, we defer the discussion of prior choice until section 2.6.

2.3 The Relationship Between the Full Model and the Correlated Errors Model.

In both the 'full' and correlated errors models, the likelihood of the observables (X, Y) given G is bivariate normal. In particular, the correlated errors model must also make the implicit assumption that the set of confounders U is normally distributed, just as in the full model. Note that the likelihood structure assumed by the correlated errors model contains a variance-covariance matrix that is not a function of G, implying that at most three variance parameters can be identified, along with four parameters for the mean. The correlated errors model is therefore identifiable. However, since there are four variance parameters in the full model to estimate, it must be non-identifiable. This is clear by noting that a_2 and b_2 are the *unknown* coefficients of the *unobserved* confounders, U, and so the model cannot be identifiable.

The idea of a transparent re-parameterisation can be used to link the two models, from which we can make further useful observations relating to identifiability. Let $\boldsymbol{\theta} = (a_0, a_1, a_2, \tau_x, b_0, b_1, b_2, \tau_y)$ denote the relevant parameter vector corresponding to the full model, and $\boldsymbol{\phi_I} := (a_0, a_1, b_0, b_1, \sigma_x, \sigma_y, \lambda)$ denote the parameter vector of the correlated errors model. Tables 1 and 2 show that the mapping $\boldsymbol{\theta} \mapsto (\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$, with $\boldsymbol{\phi_N} = (a_2)$, is a bijective transformation. This transparent reparameterisation is not unique: we could replace $\boldsymbol{\phi_N} = (a_2)$ by $\boldsymbol{\phi_N} = (b_2), \boldsymbol{\phi_N} = (\tau_x)$ or $\boldsymbol{\phi_N} = (\tau_y)$, or let $\boldsymbol{\phi_N}$ be some function of $\boldsymbol{\theta}$, as long as the bijective relation between $\boldsymbol{\theta}$ and $(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$ still holds. For example, $\boldsymbol{\phi_N} = (a_2^2(a_2^2 + \tau_x^2)^{-1})$ will do.

$(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$		$\boldsymbol{\theta}$
$\boldsymbol{\phi_I}$	a_0	a_0
$\boldsymbol{\phi_I}$	a_1	a_1
$\boldsymbol{\phi_I}$	b_0	b_0
$\boldsymbol{\phi_I}$	b_1	b_1
$\boldsymbol{\phi_I}$	σ_x^2	$a_2^2 + \tau_x^2$
$\boldsymbol{\phi_I}$	σ_y^2	$(b_1 a_2 + b_2)^2 + b_1^2 \tau_x^2 + \tau_y^2$
$\boldsymbol{\phi_I}$	λ	$a_2(b_1 a_2 + b_2) + b_1 \tau_x^2$
$\boldsymbol{\phi_N}$	a_2	a_2

Table 1: Mapping $\boldsymbol{\theta}$ to $(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$.

$\boldsymbol{\theta}$	$(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$
a_0	a_0
a_1	a_1
b_0	b_0
b_1	b_1
τ_x^2	$\sigma_x^2 - a_2^2$
τ_y^2	$\sigma_y^2 - (\lambda - b_1\sigma_x^2)^2 a_2^{-2} - 2b_1\lambda + b_1^2\sigma_x^2$
b_2	$(\lambda - b_1\sigma_x^2)a_2^{-1}$
a_2	a_2

Table 2: Mapping $(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$ to $\boldsymbol{\theta}$.

Under the definition of a transparent reparameterisation with $\boldsymbol{\phi_N} = (a_2)$, it must be the case that $\mathbb{D} \perp\!\!\!\perp a_2|\{a_0, a_1, b_0, b_1, \sigma_x^2, \sigma_y^2, \lambda\}$. However, the parameters within $\boldsymbol{\phi_I}$ impose certain logical restrictions on a_2 (for example, $|a_2| \leq \sigma_x$) so that a_2 may be partially identifiable, depending on the choice of prior distribution. However, Table 2 shows that $\{\tau_x, b_2, \tau_y\}$ also depend on a_2, so that in practice we can only expect partial learning about these parameters too.

Furthermore, the correlated errors model could be defined by placing a point prior on some function of parameters in the full model, hence reducing the number of variance parameters to be estimated from four to three. Noting that we can parameterise the full model likelihood by $(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$ as in Table 2, it can be shown that the only allowable restrictions are those on a valid parameter vector $\boldsymbol{\phi}_N$. For example, since $\boldsymbol{\phi_N} = a_2$ is a valid choice of $\boldsymbol{\phi_N}$, we can specify the prior $a_2 = 1$ with probability 1. Alternatively, $\boldsymbol{\phi_N} = a_2^2(a_2^2 + \tau_x^2)^{-1}$ is also a valid choice for $\boldsymbol{\phi_N}$, which describes how the variance of X given G should be separated between U and ϵ_x, and so we can specify that this quantity be allocated some point prior. Such constraints on the full model will not be considered further.

However, the thrust of the transparent reparameterisation is that the identifiability of $\boldsymbol{\phi_I}$ implies that its posterior distribution must tend to a point mass at its true value as the sample size increases. As the causal parameter of interest, b_1, lies in $\boldsymbol{\phi_I}$, the same must be true for this parameter in the 'full' model [10]. The 'full' and correlated errors models should therefore both yield the correct mean value for b_1 in large samples. Furthermore, the one-to-one relationship between the parameters of $\boldsymbol{\theta}$ and the parameters of $(\boldsymbol{\phi_I}, \boldsymbol{\phi_N})$ implies that we can in theory find mathematically equivalent priors that yield exactly the same posterior distribution of b_1 in finite samples, subject to Monte Carlo error. In general, these equivalent priors are tedious to calculate, and we will not explore this further. Instead, the focus for the most part will be on the more interesting situation where the priors for the full and correlated errors models are not 'mathematically equivalent', and on how these models subsequently behave in practice.

2.4 Decomposing the Variance-Covariance Matrix.

A further comparison of the parameterisation implied by the full and correlated errors models reveals that the structure of the variance-covariance matrix in each parameterisation in some cases determines how accurately we can estimate b_1. Using the calculations from Table 1, we can rewrite the full model likelihood in the form

$$[X,Y|G] \sim N\left(\begin{pmatrix} a_0 + a_1 G \\ b_0 + b_1 a_0 + b_1 a_1 G \end{pmatrix}, \begin{pmatrix} a_2^2 + \tau_x^2 & a_2(b_1 a_2 + b_2) + b_1 \tau_x^2 \\ a_2(b_1 a_2 + b_2) + b_1 \tau_x^2 & (b_1 a_2 + b_2)^2 + b_1^2 \tau_x^2 + \tau_y^2 \end{pmatrix}\right). \tag{7}$$

In (7), it is clear that there are two potential sources of information about b_1: from the mean structure of (X, Y) given G, and, since b_1 also appears in the covariance structure, it too contains information on the causal parameter. It is thus possible that when combined, we can increase the precision with which we estimate b_1.

The mean structure of the correlated errors model in (6) is of course identical to the full model in (7). However, since a joint prior is applied to σ_x^2, σ_y^2 and λ in (6) rather than breaking down the covariance structure into its building blocks $\{b_1, a_2, \tau_x, b_2, \tau_y\}$ as in (7), this model cannot use the information about b_1 contained in its covariance structure. The correlated errors model therefore relies *only* on the mean structure to estimate b_1, and imposes no particular structure on the covariance matrix, as opposed to the full model.

Though the full model theoretically contains additional information about b_1 by exploiting the covariance structure, it is only useful if it contributes additional information to that gained from the mean structure alone, and if it is *accessible*. There are at least four features that influence these factors: the strength of the instrument, the sample size, the choice of priors on the parameters $\{a_2, b_2, \tau_x, \tau_y\}$, and the relative magnitudes of $\{b_1, a_2 b_2, a_2^2 + \tau_x^2, b_2^2 + \tau_y^2\}$. We treat each of these issues separately in the next section.

2.4.1 Effect of Weak Instruments, Sample Size and Prior Choice.

We investigate the role of the covariance matrix in determining the precision with which we can theoretically estimate b_1. Simple simulations are used to establish this principle. Data are generated based on the relations

$$X = a_1 G + 3U + \epsilon_x \tag{8}$$

$$Y = X + U + \epsilon_y, \tag{9}$$

where $U \sim N(0,1)$, $\epsilon_x \sim N(0,1)$ and $\epsilon_y \sim N(0,1)$, independent of each other. The minor allele frequency for the genetic variant G is taken to be 0.4. We consider the effect of a 'weak' versus stronger instrument, where weakness of instrument is assessed by the F value for the regression of X on G, with $F < 10$ indicating a 'weak' instrument [6, 20]. We compare setting $a_1 = 3$ with $a_1 = 0.3$, and generate one hundred datasets for each scenario, each with 1,000 observations. Average F and R^2 when $a_1 = 0.3$ were approximately 5.1 and 0.005 respectively, so that the instrument is classified as 'weak'. Average F and R^2 when $a_1 = 3$ were approximately 434 and 0.3 respectively, so that the instrument is not classified as 'weak'. All analyses were conducted in WinBUGS 1.4.3 using 10,000 iterations and the first 2,000 samples were dismissed. Two chains were run to inform convergence. Initial values for all chains were generated automatically in WinBUGS 1.4.3. We will report the posterior mean as the 'estimate' of the target parameter, b_1.

The parameter a_1 will be fixed at the correct instrument strength level. All likelihood parameters of the full model are given point priors at their true value, with the exception of b_1, to which we allocate a $N(0, 1)$ prior. Using (7), its structure is given below in (10):

$$[X,Y|G] \sim N\left(\begin{pmatrix} a_1 G \\ b_1 a_1 G \end{pmatrix}, \begin{pmatrix} 10 & 10b_1 + 3 \\ 10b_1 + 3 & 10b_1^2 + 6b_1 + 2 \end{pmatrix}\right) \tag{10}$$

For comparability, so that we can in effect ignore the influence of priors, we also allocate point priors to all parameters of the correlated errors model at their true value, with the exception of b_1, to which we allocate a $N(0,1)$ prior. So that the conditional distribution of the covariance matrix is the same in each model, we define an additional parameter for the correlated errors model, z_1, which is *independent* of b_1 and is also allocated a $N(0,1)$ prior. The resulting structure is given in expression (11):

$$[X,Y|G] \sim N\left(\begin{pmatrix} a_1 G \\ b_1 a_1 G \end{pmatrix}, \begin{pmatrix} 10 & 10z_1+3 \\ 10z_1+3 & 10z_1^2+6z_1+2 \end{pmatrix}\right). \tag{11}$$

By using this structure, we mimic the loss of information on b_1 in the correlated errors model, since this parameter is estimated using only the mean structure.We compare the estimates of b_1 from both the full and correlated errors models, to establish the magnitude of potential information in the covariance structure of the full model in this particular case. In addition, monitoring z_1 allows us to determine the magnitude of information on b_1 available in the covariance structure alone.

Results from each model are given in Table 3, with the plots of the posterior density of b_1 from both models, and z_1 from the correlated errors model, given in Figure 2 below. At least in principle, the results imply that valuable information about b_1 can be found in the covariance structure of the full model when the instrument is weak. By linking this information with that obtained from the mean structure, the precision of the estimates of b_1 is thus much higher than when we use the correlated errors model. With a strong instrument, however, there is no advantage in using the full over the correlated errors model, since there is an abundance of information on the causal parameter in the mean structure of (X,Y) given G alone, implying that the additional information from the covariance structure will make little difference to the estimates of b_1.

Statistic	Weak instrument			Strong instrument		
	Full	CE		Full	CE	
	b_1	b_1	z_1	b_1	b_1	z_1
Average mean	1.00	0.98	1.00	1.00	1.00	1.00
Average median	1.00	0.98	1.00	1.00	1.00	1.00
Average posterior SD	0.01	0.10	0.01	0.01	0.01	0.01
Minimum mean	0.97	0.69	0.97	0.98	0.97	0.98
Maximum mean	1.03	1.16	1.03	1.02	1.02	1.02
Average width of 95% CrI	0.04	0.41	0.04	0.03	0.04	0.04

Table 3: Summary statistics for the posterior distribution of b_1 and z_1. True value of b_1 (z_1) is 1.

A further factor to consider is sample size, since the posterior distribution of b_1 will tend to a point mass at its true value as the sample size grows, regardless of whether we link up the information in the covariance structure to that in the mean structure. We would therefore expect that both models yield comparable estimates when the sample size is sufficiently large. This is equivalent to stating that the additional information in the covariance structure of the full model contributes little or nothing to that available in the mean structure alone. Results from simulations comparing the inference from using sample size 200 versus a sample size of 2,000, assuming the same structure as in (8) and (9), fixing $a_1 = 0.3$, are given in Table 4. With a sample size of 200, average F and R^2 were approximately 2.1 and 0.01 respectively, indicating that the instrument is 'weak'. By increasing the sample size to 2,000, average F and R^2 were approximately 10.0 and 0.01 respectively. The simulations confirm that the full model outperforms the correlated errors model when small sample sizes are used, with far narrower 95% credible intervals for b_1. However, there is less difference in performance between the two models when the sample size is increased to 2,000, as expected.

Even in situations where the instrument is weak and the sample size small, the additional information in the covariance structure of the full model is not necessarily accessible, though it theoretically exists. Accessibility is dependent on the priors placed on the elements of the covariance matrix, namely $\{b_1, a_2, b_2, \tau_x, \tau_y\}$. In section 2.3, it was established that the parameters $\{a_2, b_2, \tau_x, \tau_y\}$ are non-identifiable, and consequently

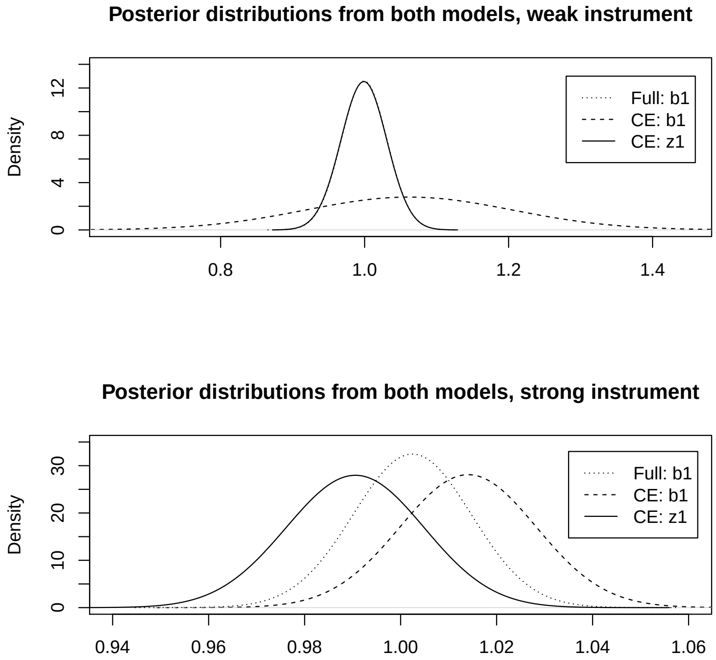

Figure 2: Smoothed posterior distributions of b_1 and z_1 from the correlated errors (CE) model, and b_1 from the full model, for one dataset with 1,000 observations. Note that the posterior of b_1 in the full model and that of z_1 in the CE model overlap exactly when the instrument is weak, implying that in this case, the information on b_1 comes predominantly from the covariance structure.

Statistic	Small Sample			Large Sample		
	Full	CE		Full	CE	
	b_1	b_1	z_1	b_1	b_1	z_1
Average mean	1.00	0.95	1.00	1.00	1.00	1.00
Average median	1.00	0.95	1.00	1.00	1.00	1.00
Average posterior SD	0.02	0.23	0.02	0.01	0.07	0.01
Minimum mean	0.94	0.55	0.94	0.98	0.81	0.98
Maximum mean	1.06	1.46	1.06	1.02	1.18	1.02
Average width of 95% CrI	0.09	0.9	0.09	0.03	0.29	0.03

Table 4: Summary statistics for the posterior distribution of b_1. True value of b_1 is 1. Once again, the results for z_1 indicate that the covariance structure drives the estimates of b_1 in the full model.

heavily reliant on their allocated priors. Only with good prior information on $\{a_2, b_2, \tau_x, \tau_y\}$ will the information on b_1 in the covariance structure therefore be accessible. The question is whether we can realistically specify sufficiently informative priors so that this additional source of information in the full model can be accessed. A more detailed discussion of prior choice is deferred until section 2.6.

A further subtle factor, which dictates whether information on b_1 exists in the covariance structure is the relative magnitude of its elements. From Table 1,

$$\sigma_y^2 = (b_1 a_2 + b_2)^2 + b_1^2 \tau_x^2 + \tau_y^2 = b_1^2 \sigma_x^2 + 2 b_1 a_2 b_2 + b_2^2 + \tau_y^2 \tag{12}$$

$$\lambda = a_2(b_1 a_2 + b_2) + b_1 \tau_x^2 = b_1 \sigma_x^2 + a_2 b_2. \tag{13}$$

Note that when the multipliers of b_1 in (12) are sufficiently small, or when $b_1^2\sigma_x^2 + 2b_1a_2b_2$ is very small relative to $b_2^2 + \tau_y^2$, a wide range of values for b_1 will give similar values of σ_y^2 (an identifiable quantity), even when very informative priors for $\{a_2, b_2, \tau_x, \tau_y\}$ are specified. Similar logic applies to (13). Under such circumstances, it is conceivable that even when information about b_1 is combined from σ_y^2 *and* λ that the possible range of b_1 will still be wide. In the most severe cases, this results in having no useful information in the covariance structure of the full model with which to estimate b_1, regardless of instrument strength, sample size or specified priors. Under such conditions, both models should yield comparable results, subject to allocated priors.

2.5 Independent Errors Model

For comparison with the full and correlated errors models, we introduce a third model which was recommended in [15] as a Bayesian method for estimating b_1. It uses equations (3) and (5), making the additional assumption that X and Y are independent given G. The likelihood of the observational model is defined as

$$X|G \sim N(\mu_x, \sigma_x^2); \tag{14}$$

$$Y|G \sim N(\mu_y, \sigma_y^2), \tag{15}$$

where $\mu_x = a_0 + a_1 G$ and $\mu_y = b_0 + b_1 a_0 + b_1 a_1 G$. A version of this approach was used in a meta-analysis context in [15], but using the predicted values for X from the first regression in the second regression, rather than regressing Y on G, in the spirit of a two stage least squares analysis. This is mathematically equivalent to the equations above.

This parameterisation assumes that the error terms V and W' in the correlated errors model in (6) are independent. Regardless of which parametric modelling assumptions are used, this cannot be true since it does not follow from either the core conditions or the DAG in Figure 1 that $X \perp\!\!\!\perp Y|G$, even when there is no confounding. By writing the above parameters of the independent errors model in terms of those of the 'full' model, it is evident that the former must assume

$$\rho := \mathrm{Corr}(X, Y|G) = \frac{a_2(b_1 a_2 + b_2) + b_1 \tau_x^2}{\sigma_x \sigma_y} = \frac{\lambda}{\sigma_x \sigma_y} = 0. \tag{16}$$

The independent errors model is hence any Bayesian model with this parameterisation and independent priors on all parameters. Note that this is equivalent to applying a point prior $\lambda = 0$ in the correlated errors model, with subsequent independent priors on σ_x and σ_y.

The causal parameter is still b_1, but an implicit restriction involving b_1 is placed on the parameters, since equation (16) implies

$$b_1 \sigma_x^2 = -a_2 b_2, \tag{17}$$

where again, the right hand side of (17) is written in terms of the 'full' model parameters. Equation (17) implies that the confounding effect and the causal effect cancel out, though this does not necessarily mean that the resulting estimate of b_1 will be poor. Taking into account that a_2 and b_2 are not fully identifiable given the data in the 'full' model, the independent errors model should adjust its estimates of other parameters to compensate, for example, by adjusting σ_x and σ_y which ultimately affects the width of the credible intervals of b_1. This holds even in the absence of confounding ($a_2 = 0$ and/ or $b_2 = 0$) where b_1 would not necessarily be estimated as zero as implied by equation (17). This explains why the model estimates the mean causal parameter well in some cases [15]. What is of interest, however, is how large a sample needs to be in order to overcome the misspecification, or how b_1 behaves in relatively small samples.

2.6 Choosing Priors and Modelling Principles.

All three models have a bivariate normal likelihood with identical mean structures but differ in the way that the covariance matrix is composed, so that the different models necessarily facilitate different prior specifications. The exact choice of priors therefore requires very careful consideration. We discuss some practical pointers, and in particular, explore how feasible it is to apply informative priors to each parameter. This, in general, is essential in order to obtain estimates of b_1 with sufficient precision to be realistically useful. For example, the non-identifiability of the full model, as discussed in section 2.3, suggests that 'vague priors' might yield parameter estimates with very low precision. Indeed, even though both the correlated errors and independent errors models are identifiable, the use of 'vague' priors should still be questioned: in many realistic scenarios, the instrument will be very weak, and so one can expect that models incorporating vague priors will also yield low-precision estimates of b_1. We will pay considerable attention to those parameters for which there is likely to be existing information, and to those parameters which can be bounded by considering their mathematical role within a model.

The identical mean structure of all three models consists of the independent parameters $\{a_0, a_1, b_0, b_1\}$. It is likely that information about $\{a_0, a_1\}$ is obtainable, since it is assumed that G is a valid instrument for X. Special consideration should always be given to the prior distribution allocated to a_1, as we will always have a priori information on this parameter given that an IV analysis makes the assumption that G and X are correlated: a variable G would not be used as an IV if there was reason to believe that $a_1 = 0$, since this violates the second core condition. Any prior should therefore give at most very small weight at and around zero, even if the instrument is known to be 'weak'. One might even consider a skewed distribution excluding zero for a_1, if the sign of a_1 is known a priori.

Prior information on b_0 may also be available since it represents the overall mean of the variable Y. It is reasonable to expect that researchers will have some insight into its magnitude. However, information on b_1 will likely be difficult to find as it is not directly observable owing to the unobserved confounders U. In most cases, however, the causal effect is unlikely to be very large, and so a somewhat informative prior is reasonable here.

The models differ in terms of their covariance structure. We look at two components: (i) the covariance of (X, Y) given G and the variance of Y given G, and (ii) the (hidden) confounding structure, and discuss how each model incorporates these factors.

In the full model, we specify priors on the (hidden) confounding structure dictated by the parameters $\{a_2, b_2, \tau_x, \tau_y\}$, independently of the causal structure. This makes implicit assumptions about the covariance of (X, Y) given G and the variance of Y given G. However, the parameters $\{a_2, b_2, \tau_x, \tau_y\}$ rely in some way on the unobserved quantity U which has no real-world counterpart. Specifying priors for these parameters will therefore inevitably be difficult, since any prior knowledge on these parameters must thus come from subject matter as they are not directly observable quantities. However, logical bounds can be deduced for one of a_2 or b_2. Table 1 reveals that the variance-covariance structure of (X, Y) given G in the full model depends on $\{a_2, b_2\}$ only via $\{a_2^2, b_2^2, a_2 b_2\}$, and so the sign of *both* a_2 and b_2 cannot be identified, though once the sign of one is fixed, it is easy to deduce the sign of the other. It is therefore logical to specify a prior distribution on a_2 *or* b_2 with a range restricted to either the positive *or* negative values, for example, either $a_2 \geq 0$ or $a_2 \leq 0$ will do. Furthermore, note that since $a_2^2 + \tau_x^2 = \sigma_x^2$, so that conservatively, $a_2 \leq \sigma_x$ and $\tau_x \leq \sigma_x$, the prior on a_2 (conversely τ_x) becomes influential only if its upper bound is less than the model's estimate of the variance of X given G, which is an identifiable quantity. Care should be taken in specifying the prior on τ_x and τ_y, however. For example, specifying a uniform distribution with lower bound too close to zero can cause WinBUGS to fail: since these parameters are only partially identifiable in that it may be possible to find an upper, but not lower, bound, the restriction on their lower bound is given by the lower bound of their respective priors. Sampling close to the lower bound, when this is not truncated appropriately, results in very low variance for ϵ_x or ϵ_y, and hence a very large precision, triggering numerical overflow. A further complication for the full model can arise when additional information is available, potentially creating a dependence between parameters. For example, as it is assumed that G is an established instrument for X, then it may be the case that information relating to σ_x^2 can be found. In light of this, it would not be appropriate to apply independent priors to a_2 and τ_x in the full model, since $\sigma_x^2 = a_2^2 + \tau_x^2$.

Other parameterisations that avoid having to specify priors for a_2 and b_2 are possible, while still preserving the information on b_1 in the covariance structure. For example, by integrating out the unobserved confounder U from the full model, we can specify

$$[X, Y|G] \sim \text{MVN}\left(\left[\begin{array}{c} a_0 + a_1 G \\ b_0 + b_1 a_0 + b_1 a_1 G \end{array}\right], \left[\begin{array}{cc} \sigma_x^2 & \rho' \sigma_x \sigma_{y|x} + b_1 \sigma_x^2 \\ \rho' \sigma_x \sigma_{y|x} + b_1 \sigma_x^2 & b_1^2 \sigma_x^2 + \sigma_{y|x}^2 + 2 b_1 \rho' \sigma_x \sigma_{y|x} \end{array}\right]\right) \quad (18)$$

where $\text{Var}[X|G] = \text{Var}[V] := \sigma_x^2$, $\text{Var}[Y|X] = \text{Var}[W] := \sigma_{y|x}^2$, and $\text{Corr}[V, W] := \rho'$, and the parameters $\{\sigma_x^2, \sigma_{y|x}^2, \rho'\}$ are independent of one another. We can therefore place independent priors directly on the parameters $\{\sigma_x^2, \sigma_{y|x}^2, \rho'\}$, for which we may have some scientific intuition about how they behave. However, from a practical perspective, implementing this model in a package such as WinBUGS results in slow parameter updates when compared with the alternative models, since it is not conditionally conjugate.

In contrast to the full model, the correlated errors model specifies priors on the covariance of (X, Y) given G and the variance of Y given G, which makes implicit assumptions about the confounding structure. Since both the covariance of (X, Y) given G and the variance of Y given G depend on the causal parameter b_1 (see Table 1), this implies that the causal and confounding structures are not distinct. Priors must therefore reflect the fact that the parameters of the covariance structure of this model are not independent.

There are several choices for priors on the covariance structure of this model. A popular choice is the inverse Wishart distribution, though by choosing this, we discard the information on b_1 in the covariance structure. The question is whether this, in reality, reduces the precision of resulting estimates of b_1 when compared to the full model, the latter of course hampered by its need for good prior knowledge on the non-identifiable parameters.

The independent errors model, on the other hand, assumes that the covariance of (X, Y) given G is zero. This is a strong version of component (i) above. Its likelihood is not therefore implied by any valid instrumental variable model, and prior knowledge on the direction or the strength of confounding cannot be integrated into the model. However, its simplifying assumption means that independent priors on all parameters are valid. This aside, as G is an established instrument for X, finding priors for its parameters will, for the most part, be a simpler process than choosing such quantities for the full or correlated errors model.

In section 3, simulated data based on a real study will provide the basis for further investigation into prior choice, and the difference in performance we can realistically expect to see between the models.

3 Simulation Study

We compare the performance of the full, correlated errors and independent errors models. In particular, we investigate whether the covariance matrix of the full model contains additional information with which to estimate the causal parameter of interest, and whether this is realistically accessible.

We initially compare the inference from the full and correlated errors models using simulations based on a real dataset. This allows us to use existing information on all parameters to build realistic prior distributions, and compare the estimates with those from models in which 'vague' priors are applied. Since the erroneous model specified in section 2.5 has also been used in practice, we compare its estimates with those from the full and correlated errors models in section 3.3.

In section 3.4, results from a second set of simulations are presented. The data are loosely based on the data structure in section 2.4. We compare the performance of all three models, and investigate whether the full model can still access the additional information on b_1 in the covariance structure.

All models are fitted using Markov chain Monte Carlo (MCMC) using single site Gibbs sampling as implemented in WinBUGS 1.4.3.

3.1 Comparing the Performance of the Models in a Realistic Scenario.

The CARDIA study [21] investigated the dependence of lung function on BMI, and found substantial lung function loss in people who were overweight as youngsters. They concluded that "the obesity epidemic threatens the lung heath of the general population". In order to compare the performance of each of the models discussed in Section 2, we simulated a series of datasets reflecting the association between lung function and BMI in the Avon Longitudinal Study of Parents And Children (ALSPAC) [22] and the CARDIA study.

The ALSPAC study provides longitudinal data on a large cohort of children but we only consider the association between the outcome Y, representing the standardised lung function measurement of FEF25-75 taken at age 8, and exposure X, which is $\log_{10}(\text{BMI})$ at age 7, in a sub-sample of 3,309 children with no history of asthma. FEF25-75 is a measure of small airway function, which is standardised so that the majority of observations fall in the interval $(-2, 2)$. There is a very weak positive association between X and Y in the ALSPAC cohort, but it is non-significant. This finding does not support the conclusions of the CARDIA study, but the possibility remains that the true causal association between X and Y is not apparent because of confounding, or perhaps because effects do not become apparent until later in life.

The FTO gene has been widely investigated for its effects on fat mass and BMI and hence seems an appropriate instrumental variable for this problem [23, 24]. The regression coefficient of X on G is 0.005 ($p < 0.001$) and of Y on G of -0.037 ($p = 0.14$), with the resulting ratio estimate of the causal effect being -8 ($p = 0.175$). Though the estimate is non-significant, the estimate is in the same direction as that found in the CARDIA study.

Datasets were simulated in R, informed by the typical values of FEF25-75, BMI and FTO, and the number of observations in the ALSPAC study. We used a small negative causal effect of X on Y, as hypothesised by the CARDIA study. One hundred datasets were generated, each containing 3,309 observations. The minor allele frequency for G was fixed at 0.4, which was the observed frequency of FTO, and the exposure (X) and outcome (Y) were generated as follows:

$$\begin{aligned} X &= 1.25 + 0.005G + 0.0025U + \epsilon_x \\ Y &= 0.12 - 0.2X + 0.7U + \epsilon_y, \end{aligned}$$

where $U \sim N(0, 1)$, $\epsilon_x \sim N(0, 0.125^2)$ and $\epsilon_y \sim N(0, 0.8^2)$, independently of one another. Weakness of instrument is often assessed by F values, where an instrument is generally classified as 'weak' if F$<$ 10 [6, 20]. Average F value over the 100 datasets was 3.7, so that the instrument appears to be 'weak', and average adjusted R^2 was 0.0008. We compare the estimates of b_1 from the full and correlated errors model.

Simple simulations using point priors on all parameters except b_1, like those in section 2.4 reveal that no *additional* information on b_1 exists in the covariance matrix of the full model, though the instrument is very weak (results not shown). This is most likely due to to the scale of the parameters in the regression of Y on X and U. In particular,

$$\sigma_y^2 := (b_1 a_2 + b_2)^2 + b_1^2 \tau_x^2 + \tau_y^2 = 0.0156 b_1^2 + 0.0035 b_1 + 1.13 \simeq 1.13 \tag{19}$$

$$\lambda := a_2(b_1 a_2 + b_2) + b_1 \tau_x^2 = 0.0156 b_1 + 0.00175 \simeq -0.01. \tag{20}$$

The information on b_1 contained in equation (19) is limited since the multipliers of b_1 terms are small so that a change in b_1 will not produce a large change in σ_y^2. Similarly for λ in equation (20). Combining these sources makes little difference in this case. These observations imply that the full and correlated errors model should yield comparable estimates of b_1. Indeed, since the sample size is relatively large, with $\text{Corr}[X, Y|G] = -0.01 \simeq 0$, then the independent errors model should also behave similarly.

We demonstrate the process of choosing realistically informative priors for each model as was outlined in section 2.6. All models include the parameters $\{a_0, a_1, b_0, b_1\}$, and so the same prior distributions are used for these parameters in each model. At age 7, most children will have a BMI that ranges between about 10 and 30, so on a $\log_{10}$ scale, this will range between approximately 1 and 1.5; this suggests that a normal prior on a_0 with mean 1.25 and standard deviation 0.1 is reasonable. The prior mean effect of FTO on $\log_{10}$BMI of 0.005 is chosen based on published literature, see for example [23, 25], and so we choose a half-normal prior $N_+(0.005, 0.005^2)$ for a_1 representing our prior knowledge that a_1 is likely to be small and positive. While this is straightforward for the full and independent errors model, this could not be done in

WinBUGS for the correlated errors model. This could be rectified by writing one's own code rather than resorting to a standard package like WinBUGS. Note, however, that choosing a $N(0.005, 0.005^2)$ prior on a_1 makes little difference to the final estimate of b_1 since a_1 is relatively easy to estimate from the data and the model should not need the additional information restricting it to positive values of a_1.

Since FEF25-75 has been standardised so that most values lie in the interval $(-2, 2)$, we place a $N(0, 0.5^2)$ prior on b_0. There is no information on the causal parameter of interest, but considering the range of FEF25-75 observations, a $N(0, 1)$ prior on b_1 seems appropriate.

The parameters $\{a_2, \tau_x, b_2, \tau_y\}$ in the full model are also given realistically informative priors. It has already been noted that without loss of generality, we can apply the constraint $a_2 \geq 0$. Since $\log_{10} \text{BMI} \in (1, 1.5)$, and we assume that $U \sim N(0, 1)$, it is reasonable to conclude that a_2 will be relatively small. A half-normal distribution, $N_+(0, 0.1)$ is placed on a_2 to reflect this. For the same reason, we expect the standard deviation τ_x to be small. A uniform prior, as is standard for prior specification of a standard deviation, bounded between 0.01 and 1 is placed on τ_x. Using the known range of FEF25-75 observations, a $N(0, 1)$ prior distribution is placed on b_2, and a $U[0.01, 5]$ prior on τ_y.

Apart from the parameters $\{a_0, a_1, b_0, b_1\}$, the correlated errors model requires the estimation of the variance-covariance matrix, Σ, for which we choose an inverse-Wishart prior. Considering the relationship between Σ and $\{a_2, b_2, \tau_x, \tau_y\}$, as displayed in Table 2, constructing somewhat informative prior information for Σ is theoretically possible, but this information is difficult to translate into a prior distribution. Therefore, we choose to place an inverse-Wishart prior on Σ with base matrix as in table 5 and 10 degrees of freedom. In WinBUGS, this will be equivalently defined as Σ^{-1} distributed as a Wishart random variable, with the same base matrix and degrees of freedom.

The impact of applying so-called 'vague priors' is also considered for each model. Note, however, that these are more informative than those typically used in practice, or those recommended in the WinBUGS manual. Chosen priors are summarised in Table 5.

Model	Parameter	True value	Prior (Vague)	Prior ('Informative')
All	a_0	1.25	$N(0, 10)$	$N(1.25, 0.01)$
Full	a_1	0.005	$N(0, 10)$	$N_+(0.005, 0.000025)$
CE	a_1	0.005	$N(0, 10)$	$N(0.005, 0.000025)$
All	b_0	0.12	$N(0, 10)$	$N(0, 0.25)$
All	b_1	-0.2	$N(0, 10)$	$N(0, 1)$
Full	a_2	0.0025	$N_+(0, 10)$	$N_+(0, 0.1)$
Full	τ_x	0.125	$U(0.01, 5)$	$U(0.01, 1)$
Full	b_2	0.7	$N(0, 10)$	$N(0, 1)$
Full	τ_y	0.8	$U(0.01, 5)$	$U(0.01, 5)$
CE	$\begin{bmatrix} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{bmatrix}$	$\begin{bmatrix} 0.016 & -0.001 \\ -0.001 & 1.13 \end{bmatrix}$	$W^{-1}\left(\begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}, 10\right)$	$W^{-1}\left(\begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}, 10\right)$

Table 5: Vague versus informative priors for each model. CE denotes 'correlated errors'. $W^{-1}(,)$ denotes the inverse-Wishart distribution.

Each model was run in WinBUGS 1.4.3 for 100,000 iterations and the first 20,000 samples were dismissed. A long burn-in was chosen on account of slow mixing. Two chains were run from different starting values to inform convergence. Initial values for all chains were generated automatically in WinBUGS 1.4.3. We will report the posterior mean as the 'estimate' of the target parameter, b_1.

Table 6 summarises the estimates of b_1, with 'true' value -0.2, for all models. Mixing was slow for the causal parameter b_1 for both models when using realistically informative priors, though it was considerably better than when 'vague' priors were used. Autocorrelation remained high for b_1, regardless of model choice.

Posterior mean estimates of b_1 from each model were consistently poor. Using 'vague' priors resulted in wildly varying posterior mean estimates of b_1 with credible intervals wide enough to render the inference of little use. With realistically informative priors, however, the models estimate b_1 with higher precision, though the estimates are still relatively poor. The credible intervals for b_1 produced by the full model are generally slightly narrower than those of the correlated errors model, even though this model produced slightly more variable mean estimates of b_1. We note that the models' failure to estimate b_1 well is due in large part to the use of a very weak instrument.

Note that we used a $N_+(0.005, 0.005^2)$ prior for a_1 in the full model, and a $N(0.005, 0.005^2)$ prior for a_1 in the correlated errors model, because of difficulties in running the latter model with the former prior on a_1. By running the full model with a $N(0.005, 0.005^2)$ prior on a_1 instead, we get virtually the same estimates of b_1 as in Table 6.

Results were also compared to a 'naïve' Bayesian analysis, where we regress the outcome Y on the exposure X only. The parameters of this model, $\{b_0, b_1, \sigma_y\}$ were allocated the priors $N(0, 0.25)$, $N(0, 1)$ and $U[0.01, 5]$ respectively. The naïve analysis thus ignores any confounding.

Statistic	'Vague' priors		Informative priors		Naïve
	Full	CE	Full	CE	
Overall mean	-0.12	-0.10	-0.09	-0.09	-0.09
Average median	-0.06	-0.09	-0.09	-0.09	-0.09
Average posterior SD	1.44	1.59	0.33	0.36	0.14
Minimum mean	-2.34	-2.35	-0.23	-0.20	-0.38
Maximum mean	2.92	1.85	0.06	0.04	0.16
Average width of 95% CrI	5.47	6.17	1.32	1.41	0.53

Table 6: Summary statistics for the posterior distribution of b_1. True value of b_1 is -0.2.

The results are fairly consistent, with the full and correlated errors models on average yielding similar estimates of b_1. This is in line with previous comments on the fact that in this case, no additional useful information on b_1 in the covariance structure of the full model existed.

3.2 Using the Prior Predictive Distribution to Reduce the Influence of Priors

In order to ensure some comparability between models, given that we do not use mathematically equivalent priors, we check that the implied prior-predictive distribution of (X, Y) given G is similar for both models. Generally this approach is used to ensure that the priors imply a sensible starting point for the distribution of (X, Y) given G, but we use this method to check that the priors allocated make the same initial assumptions about the distribution of (X, Y) given G.

The priors were chosen as follows. An initial set of priors for the full model was selected, and the implied prior predictive distribution of (X, Y) given G generated. Through trial and error, a set of priors for the correlated errors model were found, giving a similar prior predictive distribution of (X, Y) given G. These distributions were compared on the basis of the mean vector and the variance-covariance structure. They are given in Table 7, with corresponding results in Table 8.

Mixing was slow for both models, with consistently high autocorrelation for b_1 observed. Inference about b_1 is virtually identical to that using the informative priors of section 3.1. We cannot, however, rule out that the results are still influenced by our particular choice of priors, though we have to some extent standardised them.

3.3 Performance of an Independent Errors Model

Versions of the independent errors model of section 2.5 have appeared elsewhere [15], and so we also present estimates from this model. The independent errors model is disadvantaged by two faults: it does not makes use of the information on b_1 in the covariance matrix, and makes a potentially incorrect assumption about the covariance structure of the data. The question here is whether these factors sufficiently impair the model's

Model	Parameter	True value	Prior
All	a_0	1.25	$N(1.25, 0.01)$
All	a_1	0.005	$N(0.005, 0.000025)$
All	b_0	0.12	$N(0, 0.25)$
All	b_1	-0.2	$N(0, 1)$
Full	a_2	0.0025	$N_{+}(0, 0.1)$
Full	τ_x	0.125	$U(0.01, 1)$
Full	b_2	0.7	$N(0, 1)$
Full	τ_y	0.8	$U(0.01, 5)$
CE	$\begin{bmatrix} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{bmatrix}$	$\begin{bmatrix} 0.016 & -0.001 \\ -0.001 & 1.13 \end{bmatrix}$	$W^{-1}\left(\begin{bmatrix} 3 & 0 \\ 0 & 72 \end{bmatrix}, 10\right)$

Table 7: CE denotes 'correlated errors'. $W^{-1}(,)$ denotes the inverse-Wishart distribution. Priors are constructed with the prior predictive distribution of (X, Y) given G in mind.

Statistic	Full	CE
Overall mean	-0.09	-0.09
Average median	-0.09	-0.09
Average posterior SD	0.34	0.36
Minimum mean	-0.26	-0.19
Maximum mean	0.12	0.03
Average width of 95% CrI	1.33	1.41

Table 8: Summary statistics for the posterior distribution of b_1 using priors constructed so that the prior predictive distribution of (X, Y) given G for each model is roughly the same. True value of b_1 is -0.2.

ability to yield good estimates of b_1, and whether therefore the independent errors model, in practice, yields estimates which are at least comparable to those obtained via the full and correlated errors models.

Note that with data generated as in section 3.1, we have that $\text{Corr}[X, Y|G] = -0.01$. By definition, the independent errors model assumes this quantity to be zero, which in this particular example, is not a poor approximation to the true value. We would therefore expect that, in this situation, the independent errors model yield estimates of b_1 similar to those from a correctly specified model, subject to the allocation of priors. When $|\text{Corr}[X, Y|G]| >> 0$, then this incorrect assumption is likely to have a greater impact.

We run three versions of the model: one with so-called 'vague' priors, one with realistically informative priors, and finally with priors generated so that the prior predictive distribution of (X, Y) given G is similar to those generated in section 3.2, matching these distributions on the same quantities. Note that this is only possible since $\text{Corr}[X, Y|G]$ is close to zero. If this were not the case, we would never be able to generate a similar prior predictive distribution of (X, Y) given G for the independent errors model.

We place the same priors on $\{a_0, a_1, b_1, b_1\}$ as in Table 5 for each of these scenarios, together with a $U[0.01, 5]$ prior on σ_x, and a $U[0.01, 5]$ prior on σ_y as 'vague' priors. The realistically informative priors on these parameters are defined as $U[0.01, 3]$ for σ_x, and $U[0.01, 5]$ for σ_y, in recognition of the fact that $\log_{10}$BMI$\in (1, 1.5)$ and FEF25-75$\in (-2, 2)$, respectively. A $U[0.01, 1.2]$ prior for σ_x, and a $U[0.01, 5.6]$ prior for σ_y generates a similar prior predictive distribution to that in section 3.2. Results are displayed in Table 9.

With 'vague' priors, the results are typically poor, as was the case for the full and correlated errors model. However, this model yields results comparable with the full and correlated errors models when realistically informative priors were used. Results under the priors generated by considering the prior predictive distribution were very similar to those under the informative priors. The similarity of results with the full and correlated errors model is not surprising in this case, since the sample size was fairly large and

Statistic	‘Vague’ priors	Informative priors	P.P. priors
Overall mean	-0.27	-0.10	-0.09
Average median	-0.26	-0.10	-0.10
Average posterior SD	1.64	0.37	0.37
Minimum mean	-2.39	-0.18	-0.22
Maximum mean	2.08	-0.01	0.00
Average width of 95% CrI	6.33	1.44	1.44

Table 9: Summary statistics for the posterior distribution of b_1 under the independent errors model. ‘P.P. priors’ denote the priors generated by considering the prior predictive distribution. True value of b_1 is -0.2.

$\text{Corr}[X, Y|G] = -0.01$.

3.4 A further example

The simulated data in section 3.1 were designed to show the effect of informative versus vague priors, and it so happened that all models gave similar estimates of b_1. The generated data of section 2.4 however, gave a very different story, albeight under point priors for all parameters excluding b_1. We loosely base the data in this section on that in section 2.4, and investigate the impact of applying proper priors to see whether we can still access the additional information on b_1 from the covariance structure of the full model.

All three models are considered using a small versus larger a_1, and small versus larger samples. The data using a weak instrument are generated using

$$X = 0.3G + 3U + \epsilon_x \tag{21}$$
$$Y = X + U + \epsilon_y, \tag{22}$$

where $U \sim N(0,1)$, $\epsilon_x \sim N(0,1)$ and $\epsilon_y \sim N(0,1)$, independently of one another. The data with a stronger instrument are generated using

$$X = 1.2G + 3U + \epsilon_x \tag{23}$$
$$Y = X + U + \epsilon_y, \tag{24}$$

where U, ϵ_x and ϵ_y, are distributed as before. To study the impact of a small sample on each parameterisation under the data generated according to (21) and (22), and (23) and (24), we analysed 100 datasets each containing 200 observations, and a further 100 datasets containing 2,000 observations. Average F and adjusted R^2 values are given in Table 10, implying that the instrument is considered ‘weak’ when $a_1 = 0.3$. The same priors were used for each set of data, and these are presented in Table 11. Each model was run in WinBUGS 1.4.3 for 50,000 iterations and the first 10,000 samples were dismissed. A long burn-in was chosen on account of slow mixing. Two chains were run to inform convergence. Initial values for all chains were generated automatically in WinBUGS 1.4.3.

	200 obs. per dataset		2000 obs. per dataset	
Instrument	F	R^2	F	R^2
Weak	1.2	0.001	9.3	0.004
Strong	14.6	0.06	116.1	0.05

Table 10: F and adjusted R^2 values for the data.

In Table 12, we present the results of the analyses. It is clear that the results for the full and correlated errors models follow the same pattern as seen in section 2.4. The full model outperforms both the correlated

Model	Parameter	Prior
All	a_0	$N(0, 0.25)$
All	a_1	$N(0, 1)$
All	b_0	$N(0, 0.25)$
All	b_1	$N(0, 1)$
Full	a_2	$N_+(0, 4)$
Full	τ_x	$U(0.1, 5)$
Full	b_2	$N(0, 4)$
Full	τ_y	$U(0.1, 5)$
CE	$\begin{bmatrix} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{bmatrix}$	$W^{-1}\left(\begin{bmatrix} 84 & 0 \\ 0 & 168 \end{bmatrix}, 10\right)$
IE	σ_x	$U(0.1, 5.6)$
IE	σ_y	$U(0.1, 8.2)$

Table 11: Priors for each model. CE denotes 'correlated errors' and IE 'independent errors'.

errors and independent errors models, especially when the instrument is weak and the sample size is small. We therefore conclude that the priors placed on the parameters $\{a_2, \tau_x, b_2, \tau_y\}$ in this case were sufficiently informative to access additional information on b_1 from the covariance structure of the full model.

With a stronger instrument or a larger sample size, estimates from the correlated errors and independent errors models improve considerably, though they are not as precise as those from the full model. However, whilst inference from the correlated errors model and 'full' model appear almost identical when we use a large dataset *and* a stronger instrument, the independent errors model still yields credible intervals for b_1 which are far wider than those from either the correlated errors or the 'full' models, though its posterior mean estimates are comparable to those from the correctly specified models. Note that the independent errors model yields estimates with even less precision owing to its incorrect assumption that $\mathrm{Corr}[X, Y|G] = 0$. Unlike the example in section 3.1 where this assumption was in fact close to the true value, here we have that $\mathrm{Corr}[X, Y|G] = 0.97$. It is likely that vast sample sizes would be required here for the independent errors model to yield comparable results to the full model.

It is interesting to note that the naïve estimate for the association between X and Y is around 1.3 for both the models with weak and strong instrument, with $\mathrm{Corr}(X, Y|G) \simeq 0.97$. The 'full' model, when the instrument is weak, appears to be biased towards the naïve estimate. In this case, since the naïve estimate and the true underlying causal effect of X on Y, i.e. 1, are not too different, the mean estimate of b_1 produced is relatively close to the true underlying value. The reduced models are very sensitive to the prior placed on b_1. We placed a zero mean prior on b_1, as would be commonplace in practice, and resulting inference on b_1 from each of these models tends to be biased towards this value. By placing a different prior on b_1, for example centered at one (which happens to be the true value of b_1), a dramatic improvement in estimates of b_1 from the reduced models is seen (results not shown).

4 Discussion

We compared three possible approaches to the Bayesian analyses of Mendelian randomisation studies, under the assumption that all relationships are linear and that no interactions are present. Crucially, the models differ in terms of how prior knowledge on the causal and confounding structure is specified.

The 'full' representation explicitly models the unobserved confounder U, so that it makes the 'hidden' regression coefficients of the unobserved confounder and the residual standard deviations explicit, while it leaves the implied covariance matrix of (X, Y) given G implicit. This structure allows us to learn about the causal parameter b_1 via its mean *and* covariance structure, though it renders the model non-identifiable. In contrast, the correlated errors model is identifiable since it leaves the confounding structure implicit, but

		200 obs. per dataset			2000 obs. per dataset		
Instrument	Statistic	Full	CE	IE	Full	CE	IE
Weak	Average mean	1.13	0.62	0.39	1.02	0.76	0.76
	Average median	1.18	0.72	0.39	1.04	0.85	0.74
	Average posterior SD	0.28	0.72	0.76	0.17	0.40	0.51
	Minimum mean	0.51	-0.01	-0.02	0.62	0.12	0.08
	Maximum mean	1.31	1.21	1.14	1.24	1.17	1.12
	Average width of 95% CrI	1.17	2.88	3.03	0.66	1.55	2.04
Stronger	Average mean	0.98	0.85	0.92	0.99	0.99	1.00
	Average median	1.00	0.90	0.89	0.99	0.99	1.00
	Average posterior SD	0.16	0.27	0.41	0.04	0.05	0.15
	Minimum mean	0.66	0.21	0.10	0.87	0.86	0.89
	Maximum mean	1.27	1.19	1.27	1.08	1.08	1.09
	Average width of 95% CrI	0.63	1.08	1.63	0.17	0.18	0.59

Table 12: Summary statistics for the posterior distribution of b_1, true value 1.

the covariance matrix of (X, Y) given G explicit. However, it learns about b_1 via the mean structure only due to the choice of prior distribution on the variance-covariance matrix. The third model, the independent errors model, restricts the covariance by making the implicit assumption of independence between X and Y given G. It relies on the mean structure only to estimate b_1, under this assumption. It is not a reduction of the 'full' model, but is included on the basis that it was found to estimate the causal parameter well in [15].

The question is whether the full model, in practice, can access the information on b_1 in the covariance structure, and whether in reality this yields estimates with higher precision when compared to either of the reduced models. The discussion in section 2.4.1 reveals that there are at least four factors which determine this: the strength of the instrument, sample size, the priors that are placed on the non-identifiable parameters $\{a_2, b_2, \tau_x, \tau_y\}$, and the relative magnitudes of these parameters. Only when the instrument is weak or the sample is small is there potential to learn about b_1 from the covariance as well as the mean structure. However, even under these conditions, accessibility of this additional information is highly dependent on having good prior knowledge on the non-identifiable parameters of the full model. In situations where additional information in the covariance structure exists, but cannot be accessed, the full and correlated errors models should yield comparable results, subject to allocated priors for each model.

The strength of the correlated errors model, however, appears to be its robustness to misspecification of the prior on the covariance matrix, since the model does not directly rely on the covariance structure to estimate b_1. Note also that an inverse Wishart prior on the covariance matrix does not restrict its range, unlike some choices of priors for the comparable elements of the full model. This feature of the correlated errors model is particularly advantageous in situations where prior knowledge on the covariance structure is poor.

The independent errors model may in some situations perform as well as the full and correlated errors models as discussed in section 3.3. This is especially the case when the true underlying covariance is close to zero, owing to its implicit assumption that the confounding and causal effects cancel each other out. It should not be a viable alternative when this is not the case, or when the sample size is small, especially when the instrument is weak. However, prior specification for the parameters of the model may be conceptually easier, though we cannot specify any prior knowledge on the direction or the strength of confounding. Results from this model should therefore be treated with caution.

These observations have numerous practical implications, which we summarise as *choice of likelihood* and *prior choice*. Results in section 3 suggest that the non-identifiable nature of the full model should not be seen as a barrier, as it is capable of yielding estimates of b_1 which are at least as precise as those from the correlated errors or independent errors models. This finding supports conclusions elsewhere [26, 27], and reinforces the message that we should not see non-identifiability as "bad" and identifiability as "good"; the question of which model is appropriate for a particular data set must be judged on a case-by-case basis. Furthermore, in line with observations in [26, 27], we have shown that the non-identifiable full model can, under certain conditions, *outperform* both the reduced parameter models. This, however, is subject to applying realistically informative priors on the parameters of the full model, and dependent on the instrument being weak or the sample size being small, and on the relative magnitudes of particular parameters.

Despite the theoretical superiority of the full model under particular conditions, the nature of the prior information available may dictate that an alternative parameterisation is more appropriate, and so the choice of model should be at least partly based on which model utilises the available prior information most effectively. This is especially important when the instrument is weak, as is typical of a Mendelian randomisation study. Under such conditions, informative priors are crucial to the success of all models, regardless of whether they are identifiable, and so-called 'vague priors' should be avoided whenever possible. However, care must be taken when specifying informative priors, as substantial variation in parameter estimates can occur as the prescribed priors are modified [10, 28].

Identifying the causal parameter of interest in this simple setting by applying Bayesian methods is not as straightforward as one might expect. Some of the difficulties encoutered are WinBUGS-related and could be avoided by writing one's own code, while others, such as the issue of choosing suitable priors, are likely to remain problematic. In particular, given the difficulty in estimating the causal parameter in the linear, no interactions case, this caution must be extended to more complex models such as when the outcome is binary [29, 30], where models are known to be more sensitive to misspecifications and incorrect assumptions, and prone to more complex forms of bias [7].

Acknowledgements

The authors acknowledge support from the Medical Research Council (Project Grant G0601625) and the Leverhulme Trust (Research Fellowships RF/9/RFG/2009/0062 and RF-2011-320). We are grateful for access to the ALSPAC data which informed our simulation study. ALSPAC respiratory function measures were supported by the Medical Research Council (Project Grant G0401540). We are also grateful to two anonymous referees for very helpful comments and suggestions.

References

[1] Dawid, AP. Conditional independence in statistical models, *J. R. Statist. Soc. B* 1979; **41(1)**: 1-31.

[2] Davey-Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?, *Int. J. Epidemiol.* 2003; **32(1)**: 1-22.

[3] Didelez V, Sheehan NA. Mendelian Randomization as an Instrumental Variable Approach to Causal Inference, *Stat. Med.* 2007; **16**: 309-330.

[4] Didelez V, Sheehan NA. Mendelian Randomisation: Why Epidemiology needs a Formal Language for Causality. *Causality and Probability in the Sciences.* College Publications London, 2007; eds. Russo F, Williamson J.

[5] Lawlor D, Harbord R, Sterne JAC, Timpson N, Davey-Smith G. Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology, *Stat. Med.* 2008; **27**: 1133-1163.

[6] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* 1995; **90**: 443-450.

[7] Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology, *Statistical Science* 2010; **25**: 22-40.

[8] Wald, A. The fitting of straight lines if both variables are subject to error, *Annals of Mathematical Statistics* 1940; **11(3)**: 284-300.

[9] Poirer, DJ. Revising beliefs in nonidentified models, *Econometric Theory* 1998; **14**: 483-509.

[10] Gustafson, P. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science* 2005; **20(2)**: 111-140.

[11] Gustafson, P. What Are the Limits of Posterior Distributions Arising From Nonidentified Models, and Why Should We Care? *American Statistical Association* 2009; **104(488)**: 1682-1695.

[12] Gelfand AE, Sahu SK. Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models, *American Statistical Association* 1999; **94(445)**: 247-253.

[13] Greenland, S. Multiple-Bias Modelling for Analysis of Observational Data, *J. R. Statist. Soc. A* 2005; **168(2)**: 267-306.

[14] Gustafson, P. Bayesian Inference for Partially Identified Models, *The International Journal of Biostatistics* 2010; **6(2)**: Article 17.

[15] Burgess S, Thompson SG. Bayesian Methods for Meta-Analysis of Causal Relationships Estimated Using Genetic Instrumental Variables, *Stat. Med.* 2010; **29(12)**: 1298-1311.

[16] Pearl, J. *Causality.* Cambridge University Press, 2000.

[17] Hernán MA Robins JM. Instruments for Causal Inference: An Epidemiologists Dream? *Epidemiology* 2006; **17(4)**: 360-372.

[18] Lancaster, T. *Bayesian Econometrics Modelling.* Blackwell publishing, 2004.

[19] Pearl, J. An Introduction to Causal Inference, *The International Journal of Biostatistics* 2010; **6(2)**: Article 7.

[20] Didelez V, Sheehan NA. Commentary: Can 'many weak' instruments ever be 'strong'? *International Journal of Epidemiology* 2011; DOI: 10.1093/ije/dyr017.

[21] Thyagarajan B, Jacobs DR, Apostol GG, Smith LJ, Jensen RL, Crapo RO, Barr RG, Lewis CE, Williams OD. Longitudinal association of body mass index with lung function: The CARDIA Study, *Respir. Res.* 2008; **9**: 31.

[22] Golding, J. 'Children of the nineties. A longitudinal study of pregnancy and childhood based on the population of Avon (ALSPAC). *West Engl. Med. J.* 1990; **105(3)**: 80-2.

[23] Frayling TM, Timpson NJ, Weedon MN, Zeggini E et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity, *Science* 2007; **316**: 889-894.

[24] Timpson NJ, Harbord RM, Davey- Smith G, Zacho J et al. Does greater adiposity increase blood pressure and hypertension risk? Mendelian randomization using the FTO/MC4R genotype, *Hypertension* 2009; **54**: 84-90.

[25] Loos RJF, Lindgren CM, Li S, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics letters* 2008; **40(6)**: 768-775.

[26] Gustafson, P. Measurement error modelling with an approximate instrumental variable. *J. R. Statist. Soc. B* 2007; **69(5)**: 797-815.

[27] Gustafson P, Greenland S. Interval estimation for messy observational data. *Statistical Science* 2009; **24(3)**: 328-342.

[28] Greenland, S. Relaxation Penalties and Priors for Plausible Modeling of Nonidentified Bias Sources. *Statistical Science* 2009; **24(2)**: 195-210.

[29] McKeigue PM, et al. Bayesian methods for Instrumental Variable analysis with Genetic Instruments (‘Mendelian Randomisation’): Example with Urate Transporter SLC2A9 as an Instrumental Variable for Effect of Urate Levels on Metabolic Syndrome, *Int. J. Epidemiol.* 2010; **39(3)**: 907-918.

[30] Goetghebeur, E. Commentary: To Cause or Not to Cause Confusion vs Transparency with Mendelian Randomization, *Int. J. Epidemiol.* 2010; **39(3)**: 918-920.