Research Article

Statistics in Medicine

Received 12 October 2011, Accepted 26 September 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5659

Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies

Roger M. Harbord,[a,b] Vanessa Didelez,[d] Tom M. Palmer,[a,b] Sha Meng,[c] Jonathan A. C. Sterne[a] and Nuala A. Sheehan[c*†]

Mendelian randomization studies estimate causal effects using genetic variants as instruments. Instrumental variable methods are straightforward for linear models, but epidemiologists often use odds ratios to quantify effects. Also, odds ratios are often the quantities reported in meta-analyses. Many applications of Mendelian randomization dichotomize genotype and estimate the population causal log odds ratio for unit increase in exposure by dividing the genotype-disease log odds ratio by the difference in mean exposure between genotypes. This 'Wald-type' estimator is biased even in large samples, but whether the magnitude of bias is of practical importance is unclear. We study the large-sample bias of this estimator in a simple model with a continuous normally distributed exposure, a single unobserved confounder that is not an effect modifier, and interpretable parameters. We focus on parameter values that reflect scenarios in which we apply Mendelian randomization, including realistic values for the degree of confounding and strength of the causal effect. We evaluate this estimator and the causal odds ratio using numerical integration and obtain approximate analytic expressions to check results and gain insight. A small simulation study examines finite sample bias and mild violations of the normality assumption. For our simple data-generating model, we find that the Wald estimator is asymptotically biased with a bias of around 10% in fairly typical Mendelian randomization scenarios but which can be larger in more extreme situations. Recently developed methods such as structural mean models require fewer untestable assumptions and we recommend their use when the individual-level data they require are available. The Wald-type estimator may retain a role as an approximate method for meta-analysis based on summary data.

Keywords: Mendelian randomization; instrumental variables; causal inference; binary outcomes; odds ratios

1. Introduction

Mendelian randomization studies estimate causal effects of exposures on outcomes using genetic variants associated with the exposure of interest. Mendelian randomization is an application of the method of instrumental variables [1–3], previously more familiar to econometricians than epidemiologists. However, several features of epidemiology in general and Mendelian randomization applications in particular make it less than straightforward to apply instrumental variable estimation techniques, particularly when the outcome is not continuous but binary, as is often the case for outcomes of clinical interest. The exposure of interest in Mendelian randomization is usually continuous (e.g. fibrinogen [4], plasma C-reactive protein (CRP) [5], body mass index [6]) rather than the binary 'treatment' commonly considered in econometrics and clinical trials, and epidemiologists often wish to quantify the effect of exposures on binary outcomes using odds ratios or risk ratios.

Before the connection with instrumental variables was pointed out, several authors [7–11] used a simple method to compare estimates from Mendelian randomization studies with those from observational studies without any formal derivation or discussion of its properties. Although the method

[a] *School of Social and Community Medicine, University of Bristol, Bristol, U.K.*
[b] *Medical Research Council Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, U.K.*
[c] *Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, U.K.*
[d] *School of Mathematics, University of Bristol, Bristol, U.K.*
**Correspondence to: Nuala A. Sheehan, Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, U.K.*
[†] *E-mail: nas11@le.ac.uk*

generally uses a dichotomized genotype, it is applicable when we specify a model for the genotype–exposure relationship that is based on a single parameter, such as an additive genetic model, or a model incorporating multiple genotypes or instruments. The method amounts to calculating the Mendelian randomization estimate of a causal effect (or the log transform of ratio measures of effect, e.g. log risk ratio or log odds ratio) of exposure on outcome per unit increase in exposure by dividing the effect of genotype on outcome by the difference in mean exposure between genotypes. This estimator arises intuitively by ignoring the population variability of exposure within each genotype, that is it would be correct if everyone with a given genotype had exactly the mean exposure for that genotype.

As a typical application of the aforementioned simple method, we consider circulating CRP, a marker of systemic inflammation that has repeatedly been associated with hypertension. However, this association may have arisen because of unmeasured confounding or reverse causation, rather than from a causal effect of CRP on hypertension. This issue was investigated in a Mendelian randomization study based on a sample of 3529 women aged 60–79 years from the British Women's Heart and Health Study (a long-term cohort study of women who were originally recruited in middle age). The authors used the 1059G/C polymorphism within exon 2 of the *CRP* gene, which is known to be associated with circulating CRP concentrations, as an instrumental variable [9]. A doubling of CRP was associated with an increase in hypertension risk in this study (odds ratio of 1.14; 95% CI 1.09, 1.19). However, the instrumental variable analysis gave an estimated odds ratio of 1.03 (0.61, 1.73) leading the authors to conclude that the association was not causal and that 'developing pharmaceutical agents to lower CRP levels will not be a productive strategy'.

When the instrument–exposure and exposure–outcome models are both linear, the aforementioned estimate is known as the Wald estimator [12–14], and (given certain model assumptions) its bias tends to zero as the sample size becomes larger, that is it is asymptotically unbiased. When applied to nonlinear models, we shall refer to such an estimate as a 'Wald-type' estimate [15]. In addition to its simplicity and intuitive appeal, the Wald-type estimate is particularly useful for reanalysis of published results, including meta-analysis, as it requires only commonly reported summary measures of the genotype–outcome and genotype–exposure associations. The Wald-type estimator of a risk ratio or rate ratio can also be shown to be asymptotically unbiased when the exposure–disease model is log linear if certain additional model assumptions hold [15, 16]. Wald-type estimates of the causal odds ratio or log odds ratio, however, are known to be asymptotically biased when the true effect is not null [15, 17]. It is unclear whether, and when, the extent of this asymptotic (large-sample) bias is large enough to be of practical importance.

The aim of this paper is to quantify the asymptotic bias of the Wald-type estimate of the causal odds ratio in scenarios typical of Mendelian randomization studies. We derive expressions for the bias involving integrals with no exact closed form, evaluate these integrals numerically for a realistic range of parameter values and also give approximate closed-form expressions to check the numerical results and provide further insight.

The outline of the paper is as follows: We first define the causal parameters we wish to estimate, then formally define the Wald-type estimator, as well as the associational estimator, of the causal odds ratio. We introduce a simple structural (causal) model for the relationships between instrument, exposure and outcome that incorporates unobserved variables to introduce confounding and heterogeneity. We show how the asymptotic bias can be evaluated for this model using both numerical evaluation and approximate closed-form expressions, giving results that are shown in the following section. The validity of these findings for finite samples and mild violations of the normal distribution for the exposure is addressed in a simulation study. We conclude by discussing the implications and limitations of our findings. We show key mathematical results in the main text but relegate derivations of results to an Appendix. We assume the reader is familiar with the epidemiological and biological basis of Mendelian randomization and the core assumptions of instrumental variables [2, 3, 8, 18, 19].

2. Estimating causal odds ratios

2.1. Causal parameters

To evaluate whether an estimate is biased, it is necessary to first clarify the target parameter we wish to estimate. In this section, we shall therefore review the definition of the causal parameters and distinguish them from model parameters [20].

Mendelian randomization is a method that uses genetic variants as instrumental variables (IVs) to assess the causal association of an exposure with a disease in the presence of unobserved confounding or

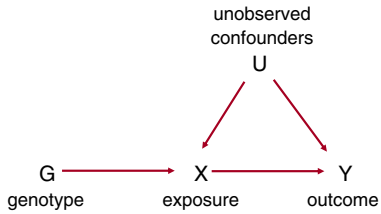

Figure 1. Directed acyclic graph relating genetic instrument G, exposure X, outcome Y and confounders U. Absent arrows encode conditional independence assumptions, but arrows do not necessarily imply causal relations.

reverse causation. Throughout this paper, we use X to denote the continuous exposure, Y to denote the binary outcome (coded 0 for event absent and 1 for event present), G to denote the genetic instrument and U to denote unobserved confounders. Figure 1 depicts the directed acyclic graph for which these four variables satisfy the core conditional independence assumptions that ensure G is a valid instrument [2]. Briefly, these state that G must be associated with X, must be independent of any confounding U and must be conditionally independent of Y given X and U. The latter would for instance be violated if G had a direct (not through X) causal effect on Y, or if G and Y were affected by further confounding. As explained in more detail elsewhere [2, 21], these core conditions are compatible with a range of causal structures, including structures in which the instrument G does not have a causal effect on X. For example, its association with X could be due to linkage disequilibrium.

The *population* causal odds ratio (which following Hernán [20] we shall refer to hereafter simply as the causal odds ratio) for exposure value $X = x_1$ relative to $X = x_0$ is

$$COR(x_1, x_0) = \frac{\Pr(Y = 1|\mathrm{do}(X = x_1))}{\Pr(Y = 0|\mathrm{do}(X = x_1))} \frac{\Pr(Y = 0|\mathrm{do}(X = x_0))}{\Pr(Y = 1|\mathrm{do}(X = x_0))},$$

where $\mathrm{do}(X = x)$ is Pearl's 'do-operator' notation for *setting* the variable X equal to the value x in an hypothetical intervention [21]. For the purposes of this paper, the 'do' notation is equivalent to the counterfactual notation used in many papers on causal inference in epidemiology and statistics [20–23]. The logarithm of $COR(x_0, x_1)$ is the causal log odds ratio

$$\log COR(x_1, x_0) = \mathrm{logit}\Pr(Y = 1|\mathrm{do}(X = x_1)) - \mathrm{logit}\Pr(Y = 1|\mathrm{do}(X = x_0)). \quad (1)$$

It is important to distinguish the causal log odds ratio from the coefficient β_S in a structural logistic equation such as

$$\mathrm{logit}(\Pr(Y = 1|\mathrm{do}(X = x), U = u) = \alpha + \beta_S x + \gamma u,$$

where U represents one or more measured or unmeasured additional covariates, which may or may not be confounders. When $\gamma \neq 0$ and there is no $X - U$ interaction, the coefficient β_S will in general be larger than $\log COR(x + 1, x)$ because of the non-collapsibility of the odds ratio [24, 25] and consequent 'shrinkage' of estimates that marginalize over a covariate compared with those that condition on it. Were U known or measured for each individual, β_s would be the causal log odds ratio conditional on U (i.e. the *individual* causal log odds ratio) but in general when U is unknown and unmeasured β_S cannot be identified [20, 26].

2.2. *Estimators*

We next define different estimators of the effect of continuous exposure X on binary outcome Y.

2.2.1. Associational estimator. We define the *associational log odds ratio* as that obtained from a logistic model that ignores confounding. This is defined as

$$\beta_A(x_1, x_0) = \mathrm{logit}\Pr(Y = 1|X = x_1) - \mathrm{logit}\Pr(Y = 1|X = x_0), \quad (2)$$

and conditions on the observed values of X rather than setting the values of X. The coefficient β_A is equal to the causal odds ratio when there is no confounding. Because X is continuous, we assume the log odds of disease to be a linear function of X, that is $\text{logit}\Pr(Y = 1|X = x) = a + \beta_A x$ to enable estimation.

2.3. *Wald-type estimator*

We now turn to Mendelian randomization using a genetic marker G as instrumental variable. A Wald-type estimator of the causal log odds ratio based on two genotypes g_1 and g_0 is the difference in the log odds of disease between the two genotypes (the logarithm of the odds ratio) divided by the difference in mean exposure comparing the two genotypes:

$$\beta_W(g_1, g_0) = \frac{\text{logit}\Pr(Y = 1|G = g_1) - \text{logit}\Pr(Y = 1|G = g_0)}{\text{E}(X|G = g_1) - \text{E}(X|G = g_0)}. \tag{3}$$

It is important to note that Wald-type estimators are not restricted to dichotomous instruments [15]. Equation (3) generalizes to ordinal or continuous G, if the numerator and denominator are extended to give a parametric description of the differences in terms of specified values of G. Furthermore, Wald-type estimators can be used for both individual-level data (in which the instrument–disease and instrument–exposure associations are estimated using the same individuals) and for summary-level data in which these quantities are estimated using different sets of individuals, for example from different studies. The latter property has made it an attractive option for meta-analyses.

The Wald-type estimate can also be obtained by fitting a linear regression of X on G using ordinary least squares and then 'plugging in' the fitted values for X in place of their actual values in a logistic regression of Y on X. This 'plug-in' estimator generalizes to more than one instrumental variable and therefore allows all three genotypes of a biallelic marker to be used without specifying the genetic model, as well as allowing more than one genetic marker, although it requires individual-level data. It also allows adjustment for covariates [27, 28]. This 'plug-in' estimator has been implemented in an add-on module for Stata statistical software [29] and has been used in applications of Mendelian randomization [9, 30, 31]. The general formulation of Wald-type IV estimators not restricted to binary instruments is given in [15].

2.4. *Model description*

Without additional model assumptions beyond those encoded in Figure 1, it is not possible even to bound the causal effect of interest when X is continuous [32, 33]. We will describe a simple model for the relationship between the genotype G, the continuous covariate X and the binary outcome Y, chosen to minimize the number of free parameters and to give as many as possible of the parameters a clear interpretation on an intuitive scale so that it is easier to judge what values for these parameters are realistic. The model requires two equations, one for X and one for Y. We choose to write the model with two separate unobserved variables (disturbances) in the equations for X and Y, respectively, and introduce confounding by assuming that they are correlated. An equivalent alternative model would have a single confounding variable U appearing in both equations but would add a second error term to the model for X to allow less than perfect correlation [34]. The form we choose leads to fewer, more interpretable parameters and is closer to that used in the econometric literature, but is not exactly the same [35]. Figure 2 shows a path diagram summarizing the structural model [36]. (Pearl has recently used a causal diagram of similar form [37, 38].)

We choose the origin and scale of X so that $\text{E}(X) = 0$ and $\text{Var}(X|G) = 1$ without loss of generality. The equation for X is

$$X = \delta(G - \text{E}(G)) + V, \tag{4}$$

where we assume that V has a standard normal distribution, that there is no interaction between G and V and that the distribution of V is independent of genotype G. Typical violations of these assumptions will show in sufficiently large samples by plotting the cumulative distribution function of X stratified by G, and a decision could be made to apply a suitable transform to X after checking such plots. For simplicity in the remainder of the paper, we assume that G is binary, that is a dominant or recessive genotype, in which case we can code G as 0 and 1 and the parameter δ is then the difference in the mean of X between genotypes. (For an additive genetic model, G would be coded as 0, 1 or 2 according to the number of minor alleles and δ would then be the difference in mean X per allele.)

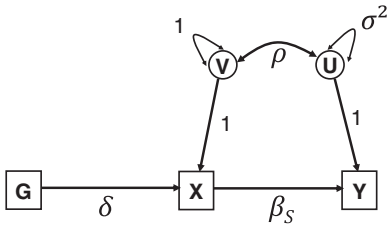

Figure 2. Path diagram for structural model. Showing separate but correlated error terms U and V. Squares indicate measured variables and circles indicate unmeasured variables (error terms). Each single-headed arrow is labelled with the corresponding model coefficient; the double-headed arrow between U and V represents a correlation (of magnitude ρ) whose source is unspecified; double-headed loops from a variable to itself are labelled with variances of error terms.

The structural equation for Y is

$$\text{logit}\Pr(Y = 1|X, U) = \alpha + \beta_S X + U, \tag{5}$$

where we assume U has a normal distribution with variance σ^2 and mean zero. Confounding of the X–Y association is introduced by assuming that U and V have correlation ρ (and therefore their covariance is given by $\text{Cov}(U, V) = \rho\sigma$). As for V in Equation (4), we assume the coefficient of U in Equation (5) is unity without loss of generality (because U is unobserved, we can always make the coefficient equal unity by a suitable re-scaling of U, or equivalently, we could include such a coefficient and instead scale U to have unit variance). In practice there would typically be many confounders, but we assume that their effect can be represented by this structure. We also assume that there is no interaction between X and U on the logit scale. As U is unobserved, all these assumptions cannot be empirically verified.

2.5. *Evaluation of estimators for this model*

2.5.1. Causal log odds ratio. Evaluation of the causal log odds ratio uses

$$\Pr(Y = 1|\text{do}(x)) = \int \Pr(Y = 1|x, u)p(u)\,\text{d}u = \int_{-\infty}^{\infty} \text{expit}(\alpha + \beta_S x + \sigma u)\phi(u)\,\text{d}u, \tag{6}$$

where $\phi(u)$ is the standard normal density function and $\text{expit}(x) = 1/(1 + e^{-x})$. This allows us to evaluate $\text{logit}COR(x_1, x_0)$ numerically for any chosen x_1 and x_0 and parameters α, β_S, and σ. Because logit $\Pr(Y = 1|$ do(x)) is nearly linear in x (unless σ is large and wide ranges of x are considered) [36], $\text{logit}COR(x_1, x_0)$ is nearly linear in x_1–x_0. For definiteness, we choose to evaluate the causal log odds ratio for a unit change in X (i.e. a change of one standard deviation) centred around the origin (the mean of X) and therefore define $\beta_C = \text{logit}COR(+1/2,\text{-}1/2)$.

As we show in the Appendix, an approximate closed form expression for β_C, which we will compare with numerical integration results, can be derived by approximating $\text{expit}(x)$ by a cumulative normal distribution function, giving

$$\beta_C \approx \beta_S/\sqrt{1 + c^2\sigma^2}, \tag{7}$$

where c is a constant, approximately 0.6.

2.5.2. Associational estimator. In practice, the associational and the Wald-type estimators will be functions of the sample. Here we derive their population versions, so that in the next section we can investigate their asymptotic biases.

To obtain an expression for the associational estimator, we need to calculate $\Pr(Y = 1|X)$. This involves integrating over G, U, and V and, as shown in the Appendix, is of the form

$$\Pr(Y = 1|X) = \sum_{g=0,1} \Pr(Y = 1|X, G = g)\Pr(G = g|X), \tag{8}$$

which can be calculated under our model assumptions for given frequencies $\Pr(G = g)$. We can then obtain the associational log odds ratio from $\beta_A(1/2, -1/2) = \text{logit}\Pr(Y = 1|X = 1/2) - \text{logit}\Pr(Y = 1|X = -1/2)$. An approximate closed form is (see Appendix)

$$\beta_A \approx (\beta_S + \rho\sigma)/\sqrt{1 + c^2\sigma^2(1 - \rho^2)}. \tag{9}$$

2.5.3. Wald-type estimator. Substituting Equation (4) for X into Equation (5) gives

$$\text{logit}\Pr(Y = 1|G, U, V) = \alpha + \delta\beta_S(G - \text{E}(G)) + U + \beta_S V,$$

which we can rewrite as

$$\text{logit}\Pr(Y = 1|G, W) = \alpha + \delta\beta_S(G - \text{E}(G)) + W, \tag{10}$$

where $W = U + \beta_S V$ has a normal distribution with zero mean and variance $\sigma^2 + \beta_S^2 + 2\sigma\rho\beta_S$.

We can proceed to evaluate $\Pr(Y = 1|G)$ by integrating over W and hence evaluate $\beta_W(g_1, g_0)$, the Wald-type estimate of the log odds ratio based on genotypes $G = 0$ and $G = 1$ defined in Equation (3).

We can once more derive (see Appendix) an approximate closed form:

$$\beta_W \approx \beta_S/\sqrt{1 + c^2\left(\sigma^2 + \beta_S^2 + 2\sigma\rho\beta_S\right)}, \tag{11}$$

which is independent of δ and the genotype frequencies.

2.5.4. Approximate closed form for asymptotic bias of Wald-type estimate. Using the aforementioned approximate results (7) and (9), we can eliminate the unknown parameters ρ and β_S from (11) in favour of β_C and β_A (assisted by the computer algebra package Maple 11 [39]). Surprisingly, the resulting expression does not depend on σ:

$$\beta_W \approx \beta_C \left(\frac{1 + c^2\beta_A^2}{1 + c^2\left[c^2\beta_C^2\beta_A^2 - \beta_C^2 + \beta_A^2 + 2\beta_C\beta_A\sqrt{1 + c^2\left(\beta_A^2 - \beta_C^2\right)}\right]} \right)^{1/2}, \tag{12}$$

As this expression is quite complex, we also used Maple to obtain a bivariate Taylor series expansion for $\beta_W - \beta_C$, the bias of the Wald-type estimate, around the point with zero causal effect ($\beta_C = 0$) and zero confounding ($\beta_A - \beta_C = 0$):

$$\beta_W - \beta_C \approx -1/2c^2\left[\beta_C^3 + 2\beta_C^2(\beta_A - \beta_C)\right].$$

It is useful to examine some limiting cases:

i $\beta_W = 0$ when $\beta_C = 0$, verifying that the Wald-type estimate is unbiased when there is no causal effect.
ii $\beta_W/\beta_C \to 1$ as $\beta_C \to 0$, so not only the absolute but also the relative bias diminishes with the causal effect.
iii $\beta_W/\beta_C < 1$ when $\beta_C \neq 0$ but there is no confounding (which occurs when $\rho = 0$ or $\sigma = 0$), so the Wald-type estimate is biased towards the null even in the absence of confounding, although (ii) shows this bias will be small for small causal effects.

Note also that β_W can be further from zero than β_C , that is the bias of the Wald-type estimate can be away from the null, for some other combinations of values of β_C and β_A.

2.6. Choice of parameter values for numerical evaluation of asymptotic bias of Wald-type estimate

We now turn to the choice of realistic ranges or small discrete sets of values for the model parameters at which to evaluate the bias for the graphs in the following section. There are a total of six model parameters, namely the parameters α, β_S, δ, σ and ρ from Section 2.4 earlier plus the genotype frequency $\Pr(G = 1)$. Neither α nor β_S is observable but they jointly determine the disease prevalence $\Pr(Y = 1)$ and the causal log odds ratio β_C, so we instead chose values for the latter pair of parameters of interest and numerically determined the required values of α and β_S. We chose disease prevalences of 2%, 10% and 50%.

We chose values of $COR(-1/2, +1/2) = \exp(\beta_C)$ of 1.2, 1.65 and 2.27. These correspond to odds ratios $\exp(2.18\ \beta_C) = 1.5$, 3 and 6, respectively, when comparing the top third versus the bottom third of the distribution of X (the factor of 2.18 is the difference between the means of the top and bottom third of a standard normal distribution) [40, 41]. We considered 1.5 to be around the typical odds ratio for top versus bottom third in a Mendelian randomization study, 3 to be around the largest credible odds ratio and 6 to probably be unrealistically large for a Mendelian randomization study to be necessary or appropriate, as strong associations are unlikely to be entirely or partially due to unmeasured confounders or other sources of modest bias [42].

The parameter σ, the standard deviation of the error term U in structural Equation (5), determines the variability in the probability of disease for individuals having the same value of exposure X. We chose values of 0.5, 1 and 5, giving a ratio of odds of disease for two individuals with the same X but values of U at its upper and lower quartiles of $\exp(1.349\ \sigma) = 1.96$, 3.85 and 850, respectively, corresponding to small, moderate and unrealistically large degrees of heterogeneity, the last because it is hard to determine a realistic upper bound (the factor of 1.349 is the difference between the upper and lower quartiles of a standard normal distribution).

The correlation ρ between the heterogeneity terms U and V determines (together with σ) the degree of confounding between exposure X and outcome Y. We used values of ρ of 0, ± 0.25, ± 0.5, ± 0.75, ± 0.9 and ± 1. The ratio of the associational odds ratio to the causal odds ratio, $\exp(\beta_A - \beta_C)$, is a more interpretable measure of the degree of confounding, so we plot results against this ratio.

Neither the genotype frequency nor the difference in mean exposure between genotypes, δ, enters into the approximate expressions, so we expected them to have little effect on the numerical evaluations. We fixed the genotype frequency at 0.5 and δ at 0.2 for the main results shown in the next section. In sensitivity analyses, we examined values for the genotype frequency of 0.1 and 0.02 and values for δ of 1 and 0.05.

3. Results

We calculated β_C, β_W and β_A for the range of parameter values described in the previous section, using Gauss–Hermite quadrature to evaluate the integrals involved. Figure 3 shows the degree of asymptotic bias of the Wald-type estimate relative to the causal odds ratio, $\exp(\beta_W - \beta_C)$. We plot this against the ratio of the associational odds ratio to the causal odds ratio, $\exp(\beta_A - \beta_C)$, for the chosen values of the causal odds ratio, the disease prevalence and σ, the standard deviation of U.

The degree of bias is usually small and exceeds 10% in either direction (ratio of odds ratios outside the range 0.9–1.1) only when the causal odds ratio is large (right-hand column) and there is considerable confounding (ratio of associational to causal odds ratios not close to unity). The bias can be in either direction depending on the nature of the confounding. Within each graph, the lines corresponding to σ values of 0.5, 1 and 5 virtually coincide except when the disease frequency is low and the causal effect reasonably large. This implies that the value of σ generally has little effect on the degree of bias of the Wald-type estimate, despite large values of σ leading to much stronger confounding and hence much greater bias in the associational estimator. In the absence of confounding, the Wald-type estimate has a small but noticeable bias towards the null when the causal odds ratio is large. The degree of bias decreases slowly with the marginal disease frequency unless the variance of the error term U is very large, that is unless the risk of disease varies greatly between individuals with the same value of exposure X. The approximate closed form expression (12) (shown by the dotted line on the graphs) provides a good approximation when the prevalence is high, but overestimates the bias when the prevalence is low.

Sensitivity analyses (not shown) found the results to be generally very insensitive to both the genotype frequency and the difference δ in mean exposure between genotypes, as we expected from the closed form approximate expressions. The only noticeable difference occurred for $\delta = 1$, $\sigma = 5$, large causal odds ratios and the smallest values of the associational odds ratio, when the bias of the Wald-type estimate was slightly higher than shown.

3.1. Simulation study

To evaluate the robustness of our asymptotic results, we conducted simulation studies that estimated the bias of the Wald-type estimator under normal as well as alternative distributions of V (and hence of X), in finite samples. These proceeded along the flow of causation depicted in Figure 2. First, we generated

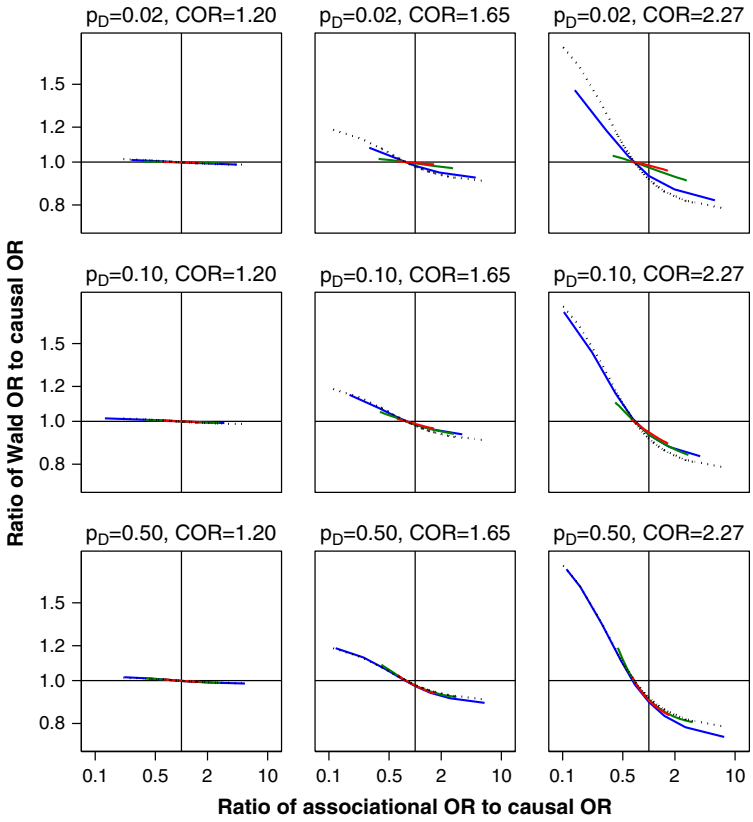

Figure 3. Asymptotic bias of the Wald-type estimate quantified by the ratio of the Wald-type odds ratio to the causal odds ratio (vertical axis) plotted against the ratio of the associational odds ratio to the causal odds ratio (horizontal axis; truncated at 0.1 and 10). Both ratios are plotted on log scales. Within each graph, red, green and blue lines show results of numerical evaluations for $\sigma = 0.5$, 1 and 5, respectively. The dotted line represents the closed form approximation. The disease frequency P(Y) is 2%, 10% and 50% in the top, middle and bottom rows of graphs, respectively. The causal odds ratio for a unit increase in X is 1.2, 1.65 and 2.27 in the left, middle and right columns of graphs, respectively, corresponding to causal odds ratios comparing the top with bottom thirds of the distribution of X of 1.5, 3 and 6, respectively.

a value of U as $N(0, 1)$. Second, we generated V as $V = \rho U + \varepsilon$, where we sampled ε from either a normal distribution, a t distribution with 4 d.f. or a χ^2 distribution with 1 d.f. recentred to have mean zero. We scaled each distribution to have unit variance, so that ρ corresponds to the correlation between U and V: we varied ρ between -0.6 and 0.6 in steps of 0.2. We generated values of X and Y according to Equations (4) and (5), with values of α and β chosen to give a causal odds ratio of 1.65 per unit increase in X (corresponding to the middle panel of Figure 3). We generated G according to a Bernoulli distribution with genotype frequency 0.5. We evaluated the Wald-type estimate from the sample analogue of Equation (3) and the associational log odds ratio from a logistic regression of Y on X. We ran 5000 replications for each combination of parameters.

We found median bias to be low, in each scenario [43]. Figure 4 compares the relative mean bias of the simple Wald-type estimate of the causal log odds ratio (plotted as geometric mean odds ratio, on a log scale) with the bias of the associational estimate, for sample sizes 500, 2000, 5000 and 20 000 and for the different exposure distributions. Results were little changed for sample sizes exceeding 20 000. Note that the bias of the associational estimate (extent of confounding) depends on the choice of distribution, as can be seen from the varying location of the points on the horizontal axes for the different distributions. The distribution of estimates became increasingly broad and skewed with decreasing sample size (plots not shown). For sample size 500, the mean bias was unstable and sometimes substantial. However, for sample sizes of 2000 or greater, mean bias was similar for the three distributions. Mean bias decreased with increasing sample size, and results for the normal distribution with 20 000 observations are similar to those in the middle panel of Figure 3, obtained by numerical integration.

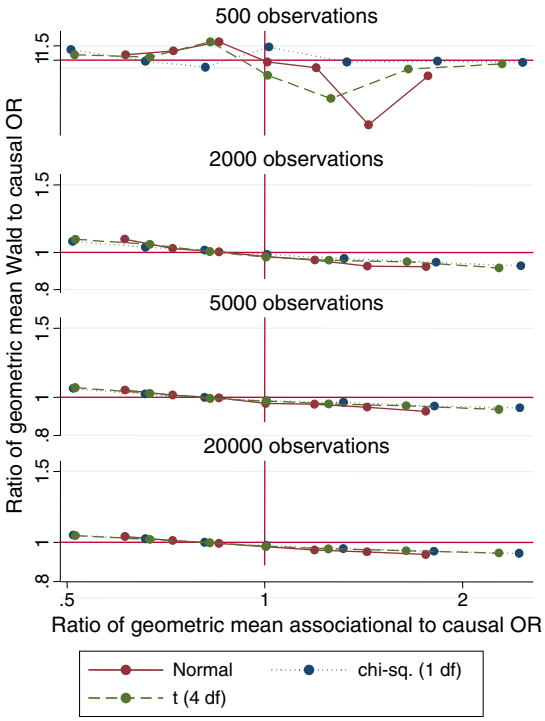

Figure 4. Results of simulations for the scenario shown as the centre panel of Figure 3 with $\sigma = 1$. Red continuous lines: normal distribution, blue dotted lines χ^2 distribution with 1 d.f., green dashed lines t distribution with 4 d.f. Note the expansion of vertical axis scale for sample size 500.

4. Discussion

We evaluated the asymptotic bias of the simple Wald-type estimate of the causal log odds ratio in a structural model in which distribution of exposure conditional on genotype is normal and homoscedastic, unobserved heterogeneity in the linear predictor for the outcome is modelled by a single variable with a normal distribution and confounding was described by a single correlation between these two error terms. The asymptotic bias was less than 10% providing that either the causal odds ratio was moderate (odds ratio for top vs. bottom third of exposure less than 3) or there was at most moderate confounding (ratio of associational to causal odds ratios within the range 0.5–2). We believe that such conditions correspond to those in which Mendelian randomization studies are most likely to be useful—when it is hard to rule out residual confounding and other sources of bias as explanations for observed associations. In our simulations, when sample sizes exceeded 2000 the mean value of the finite-sample bias was similar to its asymptotic value and appeared insensitive to mild violations of the normal exposure distributions. However substantial mean finite-sample bias was observed when sample sizes were smaller than this. All our results were restricted to the target parameter β_C defined as log $COR(+1/2, -1/2)$. We would expect larger bias if we moved the target range away from the mean exposure.

There are alternative IV estimators (such as those based on structural mean models) that make weaker assumptions than those under which we considered the bias of the Wald-type estimator, but these require individual-level data whereas the Wald-type estimator can be used for both individual-level and summary-level data. There are no IV methods for consistent estimation of a causal odds ratio so all estimators of odds ratios yield approximations. When we wish to estimate risk ratios rather than odds ratios, we can use a multiplicative generalized method of moments estimator [16, 44–46] or multiplicative structural mean models [21] to yield consistent estimates of causal risk ratios under certain assumptions. These can then be used as useful approximations of causal odds ratios in the case of a rare outcome. Logistic structural mean models [47] allow approximate estimation of odds ratios.

Bayesian approaches have been recommended as an attractive option for estimating causal odds ratios in Mendelian randomization analyses [48, 49]. They may also make strong assumptions: for instance a full likelihood model has to be specified and prior distributional assumptions made for all model parameters. Interpretation may be complicated by the influence of prior distributions and non-identifiability

issues [50, 51]. However, Bayesian approaches are worthy of further study, particularly in regard to the finite sample bias issues investigated in this paper and the situations in which they perform well.

Because it is the only estimator that does not specifically require individual level data, the Wald-type estimator is clearly of interest in a meta-analysis context where only summary data are available, and in particular when we are interested in constructing IV estimates from some studies that estimated the association between Y and G and others that estimated the association between X and G. In this situation, a sensible start point is to test for a causal effect of X on Y by testing for association between G and Y [2]. This was performed for the effect of circulating CRP levels on hypertension in the example mentioned earlier where it was noted that there was no evidence of association between CRP genotype and risk of hypertension and hence no evidence for a causal effect [9]. However, a reasonable point estimate of the causal effect will not be obtainable if the assumptions of Equations (4) and (5) cannot be justified. Properties of the Wald-type estimator under violations of the required conditions (e.g. independence of the studies and assumptions about correlation/confounding) need to be investigated and would be relevant to such applications [52–54].

Elsewhere, we have described situations in which Wald-type and structural mean model estimates differ markedly [45]. Such differences can arise in the presence of interaction between X and U in the structural equation for Y [55]. When the distribution of X given G is clearly non-normal (e.g. when X is binary), the Wald-type estimate can be badly biased [15, 56].

Recently developed methods such as structural mean models require fewer untestable assumptions than the Wald-type estimator. We recommend the use of structural mean models when the individual-level data that they require are available, because the assumptions about U and V under which we found the Wald-type estimator to have small bias (in particular the absence of interaction in Equations (4) and (5)) are untestable. The Wald-type estimator may retain a role as an approximate method for meta-analysis based on summary data.

Appendix: Derivation of approximate closed forms

Causal log odds ratio

We can obtain an approximate closed form expression for β_C by approximating expit(x) by $\Phi(cx)$,where Φ is the standard normal cumulative distribution function and c is a constant whose value is around 0.6. Various choices of c have been made: $c = 16\sqrt{3}/(15\pi) = 0.5881$ minimizes the maximum difference between expit(x) and $\Phi(cx)$, $c = 0.6071$ matches their values at logit(0.2) and logit(0.8) and $c = \sqrt{(2\pi)}/4 = 0.6267$ matches their derivatives at zero [57–59]. We chose the last of these when plotting the approximate closed form results.

Analogous to Zeger *et al.* [57] and appendix of Palmer *et al.* [34], from Equations (1) and (6), we find $\beta_C \approx \beta_S/\sqrt{1 + c^2\sigma^2}$. Figure 2 of Zeger *et al.* [57] shows that the bias is maximized when the fixed part of the linear predictor is near zero, that is when the prevalence of disease (the marginal probability of Y) is near a half. So the aforementioned equation gives an approximation to the maximum degree of attenuation, which should decrease to zero as the prevalence becomes very low, as then the logit function is very close to the log function and we know there is no bias if a log link was used in place of a logit link.

Wald-type estimator

Using the same approximation, from Equation (10) we obtain

$$\operatorname{logit}\Pr(Y = 1|G = g) \approx \alpha + \delta\beta_S g\sqrt{1 + c^2\left(\sigma^2 + \beta_S^2 + 2\sigma\rho\beta_S\right)},$$

and hence from the definition of the Wald-type estimator (3)

$$\beta_W \approx \beta_S/\sqrt{1 + c^2\left(\sigma^2 + \beta_S^2 + 2\sigma\rho\beta_S\right)}$$

(independently of the values of the genotypes g_0 and g_1).

Associational estimator

Write $U = \rho\sigma V + e$ where $e \sim N(0, \sigma^2(1-\rho^2))$ so that V and e are independent disturbance terms. Then from (4) and (5),

$$\Pr(Y = 1|X, G, e) = \text{expit}\left[\alpha + (\beta_S + \rho\sigma)X - \rho\sigma\delta(G - E(G)) + e\right].$$

We can then evaluate $\Pr(Y = 1|X)$ by first integrating out e and then summing over G using expression (8) in the main text. If we approximate the integral as earlier, we obtain

$$\text{logit}\Pr(Y = 1|X = x, G = g) \approx \alpha^* + \frac{(\beta_S + \rho\sigma)}{\sqrt{1 + c^2\sigma^2(1-\rho^2)}}x - \frac{\rho\sigma\delta}{\sqrt{1 + c^2\sigma^2(1-\rho^2)}}\left[g - E(G)\right],$$

where α^* is a constant that is not of interest. We can then use this in (8), together with an expression for $\Pr(G = g|X = x)$ derived from Equation (4) by a straightforward application of Bayes theorem, to give a closed form but lengthy expression for logit $\Pr(Y = 1| \quad X = x)$. If δ was zero, however, we could instead simply drop the last term in the aforementioned equation, and we have verified by plotting both expressions that this is also a reasonable approximation for small values of δ. We therefore choose to approximate β_A by the coefficient of x in the aforementioned equation.

Acknowledgements

The authors acknowledge research funding from the Medical Research Council (project grants G0601625 and G0600705) and the Leverhulme Trust (Research Fellowships RF/9/RFG/2009/0062 and RF-2011-320). We also thank Frank Windmeijer and Paul Clarke for helpful comments on an earlier version of this manuscript.

References

1. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. *International Journal of Epidemiology* 2004; **33**:21–25.
2. Didelez V, Sheehan NA. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**:309–330.
3. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* 2008; **27**:1133–1163.
4. Davey Smith G, Harbord R, Milton J, Ebrahim S, Sterne JAC. Does elevated plasma fibrinogen increase the risk of coronary heart disease? Evidence from a meta-analysis of genetic association studies. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2005; **25**:2228–2233.
5. Lawlor DA, Harbord RM, Timpson NJ, Lowe GDO, Rumley A, Gaunt TR, Baker I, Yarnell JWG, Kivim kM, Kumari M, Norman PE, Jamrozik K, Hankey GJ, Almeida OP, Flicker L, Warrington N, Marmot MG, Ben-Shlomo Y, Palmer LJ, Day INM, Ebrahim S, Davey Smith G. The Association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants. *PLoS ONE* 2008; **3**:e3011. DOI: 10.1371/journal.pone.0003011.
6. Timpson NJ, Harbord R, Davey Smith G, Zacho J, Tybjaerg-Hansen A, Nordestgaard BG. Does greater adiposity increase blood pressure and hypertension risk? Mendelian randomization using the FTO/MC4R genotype. *Hypertension* 2009; **54**:84–90.
7. Casas JP, Bautista LE, Smeeth L, Sharma P, Hingorani AD. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *Lancet* 2005; **365**:224–232.
8. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; **32**:22.
9. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Rumley A, Lowe GDO, Day INM, Ebrahim S. Association of C-reactive protein with blood pressure and hypertension—life course confounding and Mendelian randomization tests of causality. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2005; **25**:1051–1056.
10. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. *British Medical Journal* 2002. DOI: 10.1136/bmj.325.7374-1202.
11. Youngman LD, Keavney BD, Palmer A, Parish S, Clark S, Danesh J, Delepine M, Lathrop M, Peto R, Collins R. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: test of causality by 'Mendelian randomisation'. *Circulation* 2000; **102**:II-31–II-32.
12. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 1995; **90**:443–450.
13. Durbin J. Errors in variables. *Review of the International Statistical Institute* 1954; **22**:23–32.
14. Wald A. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics* 1940; **11**:284–300.

15. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Statistical Science* 2010; **25**:22–40.
16. Mullahy J. Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. *The Review of Economics and Statistics* 1997; **79**:586–593.
17. Thompson JR, Tobin MD, Minelli C. On the accuracy of estimates of the effect of phenotype on disease derived from Mendelian randomisation studies. *Technical Report 2003/GE1*, Centre for Biostatistics, Department of Health Sciences, University of Leicester.
18. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Internation Journal of Epidemiology* 2004; **33**:30–42.
19. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine* 2008. DOI: 10.1371/journal.pmed.0050177.
20. Hernan MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* 2004; **58**:265–271.
21. Hernán MA, Robins J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**:360–372.
22. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006; **60**:578–586.
23. Robins JM, Hern n MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
24. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**:29–46.
25. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **59**:227–240.
26. Didelez V, Kreiner S, Keiding N. Graphical models for inference under outcome dependent sampling. *Statistical Science* 2010; **25**:368–387.
27. Abadie A. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* 2003; **113**:231–263.
28. Angrist JD. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal of Business & Economic Statistics* 2001; **19**:2–28.
29. Hardin J, Schmiediche H, Carroll RJ. Instrumental variables, bootstrapping, and generalized linear models. *The Stata Journal* 2003; **3**:351–360.
30. Ding EL, Song Y, Manson JE, Hunter DJ, Lee CC, Rifai N, Buring JE, Gaziano JM, Liu S. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *New England Journal of Medicine* 2009; **361**:1152–1163.
31. Qi L, Rifai N, Hu FB. Interleukin-6 receptor gene variations, plasma interleukin-6 levels, and type 2 diabetes in US women. *Diabetes* 2007; **56**:3075–3081.
32. Bonet B. A calculus for causal relevance. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. Morgan Kaufmann: San Francisco, 2001; 40–47.
33. Pearl J. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann: San Francisco, 1995; 435–443.
34. Palmer TM, Thompson JR, Tobin MD, Sheehan NA, Burton PR. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *Internation Journal of Epidemiology* 2008; **37**:1161–1168.
35. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. MIT: Cambridge, Massachusetts, 2002.
36. McDonald RP, Ho MHR. Principles and practice in reporting structural equation analyses. *Psychological Methods* 2002; **7**:64–82.
37. Pearl J. Causal inference in statistics: an overview. *Statistics Surveys* 2009; **3**:96–146.
38. Pearl J. An introduction to causal inference. *International Journal of Biostatistics* 2010. DOI: 10.2202/1557-4679.1203.
39. Maple. Maplesoft, 2007. Waterloo, Ontario, Canada.
40. Butrum R, Cannon G, Heggie S, Kroke A, Miles L, Norman H, El S, N, James C, Stone E, Thompson R, Wiseman M. Systematic literature review specification manual version 15, World Cancer Research Fund/American Institute for Cancer Research, 2005. http:/www/dietandcancerreport.org/cancer_resource_center/downloads/SLR_Manual.pdf.
41. Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. *Journal of the American Medical Association* 1998; **279**:1477–1482.
42. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Lippincott Williams & Wilkins: Philadelphia, 2008.
43. Angrist JD. A note on bias in just identified IV with weak instruments, 2009.
44. IVPOIS. *Stata module to Estimate an Instrumental Variables Poisson Regression via GMM*. Boston College Department of Economics: Boston, MA, 2007.
45. Palmer TM, Sterne JAC, Harbord RM, Lawlor DA, Sheehan NA, Meng S, Granell R, Smith GD, Didelez V. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *American Journal of Epidemiology* 2011; **173**:1392–1403.
46. Windmeijer FAG. GMM for panel count data models. *Discussion Paper No. 06/591*, Department of Economics, University of Bristol, Bristol, UK, 2006.
47. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; **65**:817–835.
48. McKeigue PM, Campbell H, Wild S, Vitart V, Hayward C, Rudan I, Wright AF, Wilson JF. Bayesian methods for instrumental variable analysis with genetic instruments ('Mendelian randomization'): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome. *International Journal of Epidemiology* 2010; **39**:907–918.
49. Burgess S, Thompson SG. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine* 2012; **31**:1582–1600.

50. Goetghebeur E. Commentary: to cause or not to cause confusion vs transparency with Mendelian Randomization. *International Journal of Epidemiology* 2010; **39**:918–920.
51. Jones EM, Thompson JR, Didelez V, Sheehan NA. On the choice of parameterisation and priors for the Bayesian analyses of Mendelian randomisation studies. *Statistics in Medicine* 2012; **21**:231–239.
52. Burgess S, Thompson SG. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Statistics in Medicine* 2010; **29**:1298–1311.
53. Palmer TM, Thompson JR, Tobin MD. Meta-analysis of Mendelian randomization studies incorporating all three genotypes. *Statistics in Medicine* 2008; **27**:6570–6582.
54. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomization—a multivariate approach. *Statistics in Medicine* 2005; **24**:2241–2254.
55. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 1996; **91**:444–472.
56. Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variable estimation of causal odds ratios. *Statistical Science* 2001; **26**:403–422.
57. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
58. Johnson NL, Kotz S. *Distributions in Statistics, Continuous Univariate Distributions*. Houghton-Mifflin: Boston, 1970.
59. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: Cambridge UK, 1989.