STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:6570–6582
Published online 3 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3423

Meta-analysis of Mendelian randomization studies incorporating all three genotypes

Tom M. Palmer[1,*,†], John R. Thompson[1] and Martin D. Tobin[1,2]

[1]*Department of Health Sciences, University of Leicester, Leicester, U.K.*
[2]*Department of Genetics, University of Leicester, Leicester, U.K.*

SUMMARY

In Mendelian randomization a carefully selected gene is used as an instrumental variable in the estimation of the association between a biological phenotype and a disease. A study using Mendelian randomization will have information on an individual's disease status, the genotype and the phenotype. The phenotype must be on the causal pathway between gene and disease for the instrumental-variable analysis to be valid. For a biallelic polymorphism there are three possible genotypes with which to compare disease risk. Existing methods select two of the three possible genotypes for use in a Mendelian randomization analysis. Multivariate meta-analysis models for Mendelian randomization case–control studies are proposed, which extend previous methods by estimating the pooled phenotype–disease association across both genotype comparisons by using the gene–disease log odds ratios and differences in mean phenotypes. The methods are illustrated using a meta-analysis of the effect of a gene related to collagen production on bone mineral density and osteoporotic fracture. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: Mendelian randomization; meta-analysis; instrumental-variable analysis

1. INTRODUCTION

Epidemiological studies investigating the relationship between biological risk factors and disease can be affected by confounding or reverse causation. The method known as Mendelian randomization has been proposed as a way of overcoming these difficulties [1, 2]. There has been a growing

*Correspondence to: Tom M. Palmer, Department of Health Sciences, University of Leicester, Leicester, U.K.
†E-mail: tmp8@le.ac.uk

Contract/grant sponsor: Medical Research Council capacity building Ph.D. Studentship in Genetic Epidemiology; contract/grant number: G0501386
Contract/grant sponsor: Medical Research Council Clinician Scientist Fellowship; contract/grant number: G0501942
Contract/grant sponsor: MRC Program Grant; contract/grant number: G0601625

Received 17 July 2008
Accepted 22 July 2008

interest in the application of Mendelian randomization because of the increased availability of genetic data.

Mendelian randomization analyses use an individual's genotype as an instrumental variable in order to estimate the association between a phenotype and the risk of disease. To fulfill the conditions for an instrumental variable the selected gene must be associated with the disease through the intermediate phenotype [3, 4]. The associations between the genotype and the phenotype and between the genotype and the disease should not be confounded by lifestyle or environmental factors because the genotype is assigned at conception before these exposures. As such an instrumental-variable estimate of the association between the phenotype and the disease derived from the gene–disease and gene–phenotype associations should also be free from confounding.

Statistical power can be low in individual Mendelian randomization studies, and large sample sizes are required to produce precise estimates of the phenotype–disease association [5, 6]. Therefore, it is an advantage if the genotype–disease and genotype–phenotype estimates are derived from meta-analyses.

2. METHODS

This section describes the information available from a case–control study and the estimation of the phenotype–disease association using Mendelian randomization. Methods are proposed for the meta-analysis of Mendelian randomization studies incorporating all three genotypes by using two genotype comparisons and an extension is given incorporating the genetic model-free approach [7, 8].

2.1. The ratio of coefficients approach for case–control studies

Suppose that the genotype and phenotype information are collected in the same study. For a genetic polymorphism with two alleles, the common and risk alleles denoted by g and G, there are three possible genotypes; the common or wild-type homozygote (gg), the heterozygote (Gg), and the mutant or uncommon homozygote (GG). Table I summarizes the genotype–disease and genotype–phenotype associations in a case–control study. In the table the counts of cases and controls are denoted by n_{dj}, subscript d indicates case or control status (1 or 0) and subscript j denotes the genotype (1, 2, or 3 corresponding to gg, Gg, and GG). The phenotype should be measured in the controls since reverse causation might affect the level of the phenotype in the cases. The observed mean phenotype levels in the controls are denoted by $\overline{x}_j$, which are estimates of the true mean phenotype levels denoted by μ_j. The observed standard deviations of the phenotype levels are denoted by sd_j. The observed mean phenotype differences between either the heterozygotes or the rare homozygotes versus the common homozygotes are given by $\widehat{\delta}_j = \overline{x}_j - \overline{x}_1$, the subscript indicates the genotype with which the common homozygotes are compared. The true genotype–phenotype mean differences are given by $\delta_j = \mu_j - \mu_1$ and the genotype–disease log odds ratios are denoted by θ_j.

In an individual study if the disease status variable were a continuous outcome measure, then the application of instrumental-variable methods would produce an unbiased estimate of the phenotype–disease association, assuming that the genotype met the core conditions to qualify as an instrumental variable [9, 10]. However, case–control studies typically rely on binary disease status variables

DOI: 10.1002/sim

Table I. Data available from a Mendelian randomization case–control study.

	Genotypes		
	gg	*Gg*	*GG*
Number of controls	n_{01}	n_{02}	n_{03}
Number of cases	n_{11}	n_{12}	n_{13}
Mean phenotypes in controls (s.d.)	$\overline{x}_1$ (sd_1)	$\overline{x}_2$ (sd_2)	$\overline{x}_3$ (sd_3)

that cause the instrumental-variable methods to produce biased estimates. The proposed approach uses gene–disease and gene–phenotype log odds ratios as continuous outcome measures in order to maintain linearity between studies [11]. The instrumental-variable method known as the ratio of coefficients approach is used to estimate the phenotype–disease log odds ratio, denoted by η, using equation (1) [12, 13]. Sometimes a unit increase in the phenotype will be biologically implausible and so an arbitrary constant k can be included in the ratio so that η represents the log odds ratio associated with a k-unit change in the phenotype [14]:

$$\eta_{[k]} \approx \frac{k\theta}{\delta} \tag{1}$$

From the data available from a Mendelian randomization case–control study reporting all three genotypes, two non-redundant estimates of the phenotype–disease log odds ratio are possible. One estimate of η is based on the comparison of the common homozygotes with the heterozygotes, using θ_2 and δ_2. The other is based on the rare homozygotes compared with the common homozygotes, using θ_3 and δ_3. In many situations it will be sensible to assume that the two estimates relate to a common underlying log odds ratio. In the meta-analysis model these two estimates of η can be combined into a single, more efficient, estimate.

2.2. *Meta-analysis incorporating two genotype comparisons*

The meta-analysis model incorporating two genotype comparisons builds on previous meta-analysis models for Mendelian randomization studies for a single genotype comparison [13, 15]. The model relates the pooled gene–disease log odds ratios and pooled gene–phenotype mean differences using the ratio of coefficients approach from equation (1) through the mean vector of a multivariate normal distribution. The model follows multivariate meta-analysis methodology, such as [16], through the specification of the marginal distribution of the study outcome measures by combining within- and between-study variance components. The approach is the multivariate analogue of the univariate random-effects meta-analysis model of DerSimonian and Laird [17].

In the following notation subscript i denotes a study. It is assumed that the observed mean phenotype differences are normally distributed such that $\widehat{\delta}_{ji} \sim \mathrm{N}(\delta_{ji},\ \mathrm{var}(\widehat{\delta}_{ji}))$ and that the true study-specific mean differences are normally distributed such that $\delta_{ji} \sim \mathrm{N}(\delta_j,\ \tau_j^2)$, where τ_j^2 is the between-study variance of the true study mean differences. Then the marginal distribution of the observed mean differences is given by $\widehat{\delta}_{ji} \sim \mathrm{N}(\delta_j,\ \mathrm{var}(\widehat{\delta}_{ji}) + \tau_j^2)$. Denoting the correlation between the pooled mean phenotype differences by ρ, the multivariate Mendelian randomization

DOI: 10.1002/sim

(MVMR) meta-analysis model then takes the following form:

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \eta\delta_2 \\ \delta_2 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix}, \mathbf{V}_i + \mathbf{B}_1 \right) \tag{2}$$

$$\mathbf{V}_i = \begin{bmatrix} \text{var}(\widehat{\theta}_{2i}) & 0 & \text{cov}(\widehat{\theta}_{2i}, \widehat{\theta}_{3i}) & 0 \\ 0 & \text{var}(\widehat{\delta}_{2i}) & 0 & \text{cov}(\widehat{\delta}_{2i}, \widehat{\delta}_{3i}) \\ \text{cov}(\widehat{\theta}_{3i}, \widehat{\theta}_{2i}) & 0 & \text{var}(\widehat{\theta}_{3i}) & 0 \\ 0 & \text{cov}(\widehat{\delta}_{3i}, \widehat{\delta}_{2i}) & 0 & \text{var}(\widehat{\delta}_{3i}) \end{bmatrix} \tag{3}$$

$$\mathbf{B}_1 = \begin{bmatrix} \tau_2^2 & \tau_2\tau_3\rho \\ \tau_2\tau_3\rho & \tau_3^2 \end{bmatrix} \otimes \begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix} = \begin{bmatrix} \eta^2\tau_2^2 & \eta\tau_2^2 & \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho \\ \eta\tau_2^2 & \tau_2^2 & \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho \\ \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho & \eta\tau_3^2 & \tau_3^2 \end{bmatrix} \tag{4}$$

The terms in the within-study covariance matrix, $\mathbf{V}_i$, are assumed to be known from the data reported by the studies and it is also assumed that there is no correlation between the gene–phenotype and gene–disease outcome measures as in [15]. From the use of the Kronecker product, it is apparent that $\mathbf{B}_1$ is singular; however, $\mathbf{V}_i + \mathbf{B}_1$ is not, which allows the calculation of the likelihood.

The parameters of this model can be estimated by maximizing the log-likelihood. For $i = 1 \ldots n$ studies Y_i represents the (4×1) vector of outcome measures, β represents the (4×1) mean vector of the multivariate normal distribution, and $\mathbf{\Sigma}_i = \mathbf{V}_i + \mathbf{B}_1$. The log-likelihood of the multivariate normal distribution up to a constant is given by

$$\sum_{i=1}^{n} -1/2\{\log(|\mathbf{\Sigma}_i|) + (Y_i - \beta)'\mathbf{\Sigma}_i^{-1}(Y_i - \beta)\} \tag{5}$$

To improve the quadratic properties of the log-likelihood the log of τ_2^2 and τ_3^2 and the Fisher's z-transform of ρ were used in the maximization that was performed using the optim function in R (version 2.7.0) [18].

2.3. Meta-analysis incorporating the genetic model-free approach

In the analysis of genetic association studies the mode of inheritance is usually unknown and so an assumption is made about the underlying genetic model. In contrast, the genetic model-free approach estimates this underlying genetic model from the available data through a parameter λ [7, 8]. When λ is equal to 0, 0.5, and 1, this represents recessive, additive, and dominant models for the risk allele, respectively.

The genetic model-free approach was devised in the context of a meta-analysis of two genotype comparisons for gene–disease outcome measures [7, 8]. A consequence of assuming that the phenotype–disease association is constant across the comparison of the heterozygotes with the

DOI: 10.1002/sim

common homozygotes and the comparison of the rare homozygotes with the common homozygotes in equation (2) is that the genetic model is assumed to be equal using either gene–disease or gene–phenotype outcomes such that

$$\lambda=\frac{\theta_2}{\theta_3}=\frac{\delta_2}{\delta_3} \tag{6}$$

The multivariate Mendelian randomization meta-analysis model incorporating the genetic model-free approach (MVMR-GMF) is given by

$$\begin{bmatrix} \widehat{\theta}_{2i} \\ \widehat{\delta}_{2i} \\ \widehat{\theta}_{3i} \\ \widehat{\delta}_{3i} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \eta\lambda\delta_3 \\ \lambda\delta_3 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix},\ \mathbf{V}_i+\mathbf{B}_2 \right) \tag{7}$$

$$\mathbf{B}_2=\begin{bmatrix} \lambda^2\tau_3^2 & \lambda\tau_3^2 \\ \lambda\tau_3^2 & \tau_3^2 \end{bmatrix}\otimes\begin{bmatrix} \eta^2 & \eta \\ \eta & 1 \end{bmatrix}=\begin{bmatrix} \eta^2\lambda^2\tau_3^2 & \eta\lambda^2\tau_3^2 & \eta^2\lambda\tau_3^2 & \eta\lambda\tau_3^2 \\ \eta\lambda^2\tau_3^2 & \lambda^2\tau_3^2 & \lambda\eta\tau_3^2 & \lambda\tau_3^2 \\ \eta^2\lambda\tau_3^2 & \lambda\eta\tau_3^2 & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\lambda\tau_3^2 & \lambda\tau_3^2 & \eta\tau_3^2 & \tau_3^2 \end{bmatrix} \tag{8}$$

Similar to the previous model $\mathbf{B}_2$ is singular but again $\mathbf{V}_i+\mathbf{B}_2$ is not. When prior knowledge about the gene suggests that $0<\lambda<1$, then the z-transform of λ can be used in the maximization along with the other transformations previously described to help improve the quadratic properties of the log-likelihood. This model was also fitted by maximizing the log-likelihood in equation (5).

It is also possible to estimate the parameters of this model using Bayesian methods. One Bayesian approach known as the product normal formulation (PNF) expresses the multivariate normal distribution for each study's outcome measures as a series of univariate normal distributions linked by the relationships between the means [19] such that

$$\begin{aligned} \widehat{\theta}_{2i} &\sim \text{N}(\eta\lambda\delta_{3i}, \text{var}(\widehat{\theta}_{2i})) \\ \widehat{\delta}_{2i} &\sim \text{N}(\lambda\delta_{3i}, \text{var}(\widehat{\delta}_{2i})) \\ \widehat{\theta}_{3i} &\sim \text{N}(\eta\delta_{3i}, \text{var}(\widehat{\theta}_{3i})) \\ \widehat{\delta}_{3i} &\sim \text{N}(\delta_{3i}, \text{var}(\widehat{\delta}_{3i})) \\ \delta_{3i} &\sim \text{N}(\delta_3, \tau_3^2) \end{aligned} \tag{9}$$

The following prior distributions were assumed for the parameters to be estimated:

$$\delta_3\sim\text{N}(0, 1\times10^6),\quad \tau_3^{-2}\sim\text{Gamma}(0.1, 0.1),\quad \eta\sim\text{N}(0, 1\times10^6),\quad \lambda\sim\text{Beta}(1, 1) \tag{10}$$

The prior distributions on δ_3, τ_3^{-2}, and η were chosen to be non-informative; for example, the normal prior distribution is approximately uniform over a broad range. The Beta prior distribution restricts λ to lie between 0 and 1.

DOI: 10.1002/sim

2.4. Missing outcomes

In a meta-analysis it is possible that some studies may not report all four outcomes. If studies are missing either gene–disease or gene–phenotype outcome measures these studies can be included in the model fitting using the appropriate bivariate log-likelihood derived by taking the appropriate rows and columns from equations (2)–(4) or equations (7), (3), and (8). This requires the assumption that the missing outcomes are missing at random and not missing for a systematic reason.

2.5. Diagnostic plots

The results of a bivariate Mendelian randomization meta-analysis have been presented using a two-column forest plot instead of two separate forest plots [13, 15]. For the models presented here

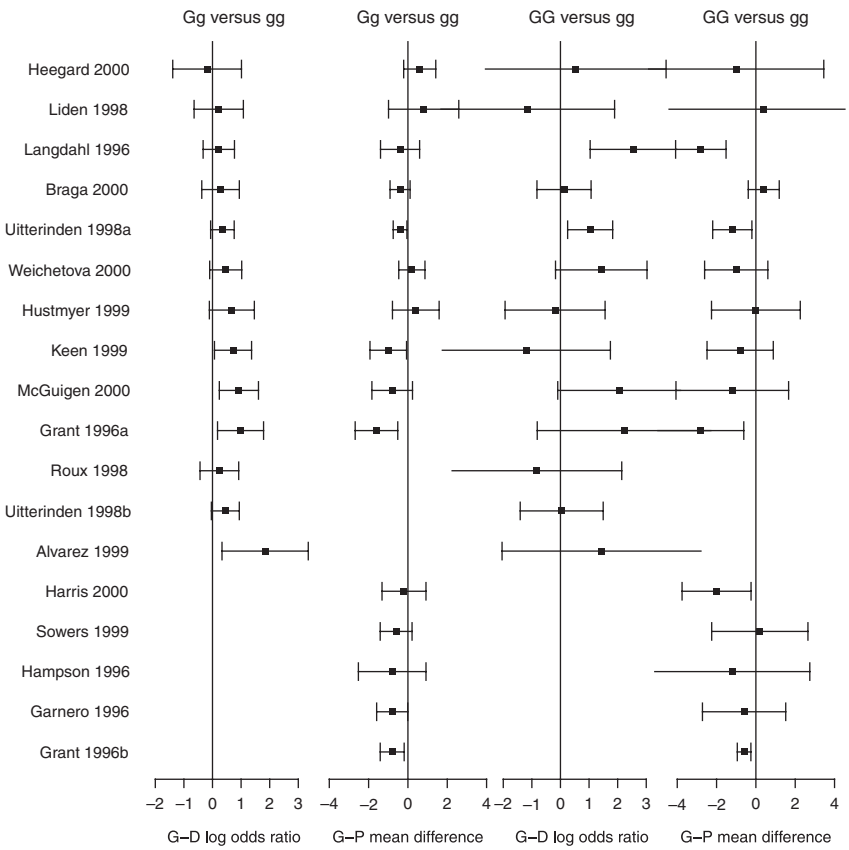

Figure 1. Four-column forest plot of the *COL1A1* multivariate meta-analysis. The genotype–phenotype (G-P) columns are on a per $0.05\,g/cm^2$ scale.

DOI: 10.1002/sim

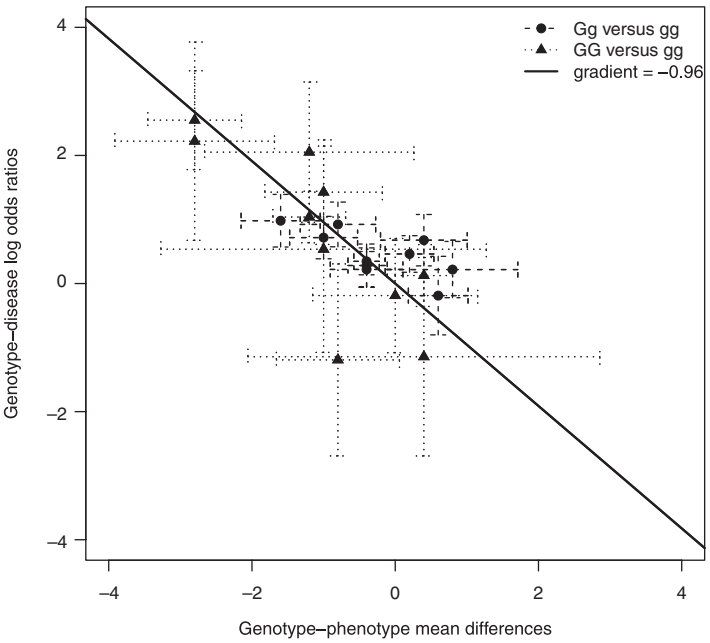

Figure 2. Gene–disease log odds ratios versus gene–phenotype mean differences (per $0.05\,g/cm^2$) plotted with one standard deviation error bars. The gradient of the line is given by $\widehat{\eta}$ from the MVMR meta-analysis model.

using four outcomes this can be extended to a four-column forest plot. To help compare the precision of the estimates, the two columns of gene–disease log odds ratios should use the same scale as should the two columns of gene–phenotype mean differences. This plot is shown in Figure 1.

In the meta-analysis models the assumption of the common phenotype–disease association in both genotype comparisons can be assessed by plotting the gene–disease outcome measures against the gene–phenotype measures [13]. From the ratio of coefficients approach, the phenotype–disease association can be expressed as the gradient of the line of best fit through the origin on this plot that is shown in Figure 2.

In the MVMR-GMF meta-analysis model the assumption that the genetic model is the same in the gene–disease and gene–phenotype outcomes can be assessed by plotting the Gg versus gg comparison against the GG versus gg comparison for each set of outcomes, respectively [7]. From the genetic model-free approach, λ is given by the gradient of the line of best fit through the origin on these plots that are shown in Figure 3.

3. APPLICATION TO BONE MINERAL DENSITY AND OSTEOPOROTIC FRACTURE

A meta-analysis that investigated the relationship between a polymorphism in the COL1A1 gene and bone mineral density (BMD) and the risk of osteoporotic fracture is used to illustrate the methodology [20].

DOI: 10.1002/sim

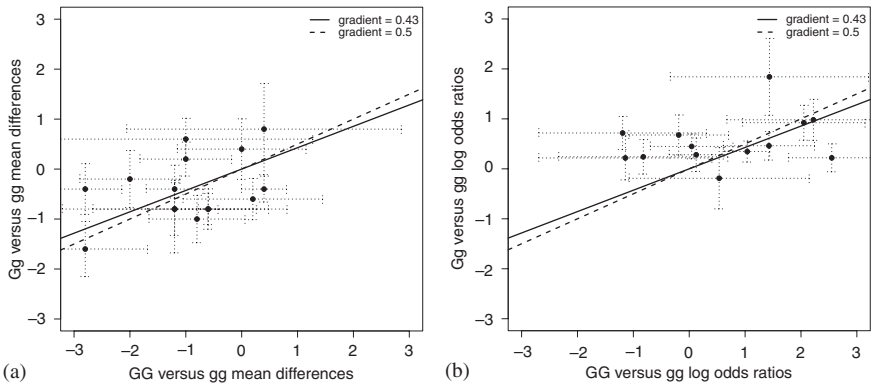

Figure 3. Graphical assessment of the estimated genetic model. The gradient of the bold lines is $\widehat{\lambda}$ from the MVMR-GMF model. A dashed line with gradient 0.5 representing the additive genetic model is also shown; lines with gradients 0 and 1 would represent the recessive and dominant genetic models, respectively: (a) genotype–phenotype information per $0.05\,g/cm^2$ and (b) genotype–disease information.

3.1. Description of the meta-analysis

The *COL1A1* gene codes for one of the main forms of collagen and the Sp1 polymorphism has been shown in epidemiological studies to be associated with both BMD and the risk of fracture [21, 22]. This polymorphism is therefore a candidate for use as an instrumental variable in the estimation of the association between BMD and fracture risk. The *COL1A1* study presented two meta-analyses based on a single nucleotide, *G* to *T*, polymorphism affecting a binding site for the transcription factor Sp1 in the *COL1A1* gene. One meta-analysis investigated studies into *COL1A1* and BMD and the other meta-analysis investigated studies of *COL1A1* and osteoporotic fracture risk. It is therefore possible to apply Mendelian randomization meta-analysis to this example. The studies of the gene–phenotype and gene–disease associations should be free from confounding, whereas studies of the association of BMD with fracture may be confounded by factors such as the subject's age or the amount of exercise they take, and there may also be unknown confounders that cannot be controlled for in the analysis.

The *G* and *T* alleles of the polymorphism in the *COL1A1* gene are sometimes labelled as *S* and *s* for the common and risk alleles, respectively, but for consistency with the Methods section they are labelled as *g* and *G*. In estimating the phenotype–disease association using Mendelian randomization, a one-unit change in the phenotype can have a large impact on disease risk. In the example the standard deviation of the mean difference in BMD was $0.05\,g/cm^2$ between the homozygote genotypes and $0.03\,g/cm^2$ for comparison of the heterozygotes versus the common homozygotes. Therefore, the scaling constant, *k*, was set to 0.05 in the analysis to ensure that the pooled phenotype–disease odds ratio was estimated on an appropriate scale.

3.2. Results of the meta-analysis

Figure 1 shows a four-column forest plot of the *COL1A1* meta-analyses. The first and second columns of the forest plot present the genotype–disease (G-D) and genotype–phenotype (G-P) outcomes for the *Gg* versus *gg* genotypes, while the third and fourth columns show the outcomes for the *GG* versus *gg* genotypes. The forest plot shows that there is an increased risk of fracture in

DOI: 10.1002/sim

the *Gg* over the *gg* genotype and an increased risk again in the *GG* genotype. The heterozygotes and the rare homozygotes had lower BMD than the common homozygotes. The forest plot shows that the comparison of the heterozygotes with the common homozygotes has more precise estimates because the confidence intervals around the point estimates are narrower and shows less between-study heterogeneity because the point estimates are more similar to one another.

The parameter estimates from the meta-analysis models incorporating all three genotypes are given in Table II. In the tables of parameter estimates, NA indicates a parameter that was not estimated in that particular model. The estimation of the PNF model was performed with a burn-in of 10 000 iterations followed by a chain of 50 000 iterations and MCMC convergence was assessed graphically. The estimates of η were similar across the three models with odds ratios of osteoporotic fracture of 0.38 and 0.39 per $0.05\,g/cm^2$ increase in BMD. All three pooled odds ratios were statistically significant at the 5 per cent level. The parameters in the PNF model had wider 95 per cent credible intervals than the 95 per cent confidence intervals in the MVMR-GMF model. The estimates of λ in the MVMR-GMF and PNF models were close to 0.5 with both 95 per cent intervals including 0.5 suggesting an additive model.

As a comparison parameter estimates from bivariate meta-analysis models similar to those considered by Thompson *et al.* [15] for the two genotype comparisons separately are given in Table III. The pooled odds ratio of fracture was 0.34 (95 per cent CI: 0.17, 0.68) per $0.05\,g/cm^2$ for the *Gg* versus *gg* comparison and 0.42 (95 per cent CI: 0.25, 0.72) for the *GG* versus *gg* comparison and the three estimates from the models in Table II are between the two values. The estimates in Table II are also more precise, as shown by the narrower confidence intervals, because of the inclusion of data for both genotype comparisons.

Parameter estimates from the bivariate meta-analysis models incorporating the genetic model-free approach using the gene–disease and gene–phenotype associations separately as in [7] are given in Table IV. The maximization of the gene–disease model failed to converge and so the between-study variance, $\tau^2_{\theta_3}$, was held constant. The fixed value of $\tau^2_{\theta_3}$ of 0.31 was taken from the univariate random-effects meta-analysis of the *GG* versus *gg* gene–disease log odds ratios. The estimate of λ was 0.44 (95 per cent CI: 0.19, 0.64) from the gene–disease log odds ratios and 0.42 (95 per cent CI: 0.08, 0.67) from the gene–phenotype mean differences and the estimate of λ from the MVMR-GMF model is between these two values with increased precision.

Table II. Parameter estimates for meta-analysis models using studies with complete and incomplete outcomes.

Parameter	MVMR Estimate (95 per cent CI) ($n=18$)	MVMR-GMF Estimate (95 per cent CI) ($n=18$)	PNF Estimate (95 per cent CrI) ($n=18$)
η	−0.96 (−1.39, −0.53)	−0.94 (−1.41, −0.47)	−0.97 (−1.53, −0.58)
$\exp(\eta)$	0.38 (0.25, 0.59)	0.39 (0.24, 0.63)	0.38 (0.22, 0.56)
λ	NA	0.43 (0.20, 0.61)	0.47 (0.28, 0.74)
δ_2	−0.47 (−0.63, −0.30)	NA	NA
δ_3	−0.85 (−1.35, −0.35)	−0.94 (−1.34, −0.55)	−0.92 (−1.44, −0.49)
τ^2_2	0.03 (0.001, 1.17)	NA	NA
τ^2_3	0.53 (0.15, 1.91)	0.35 (0.10, 1.28)	0.43 (0.07, 1.35)
ρ	0.05 (−0.89, 0.91)	NA	NA
Log-likelihood	−6.42	−11.24	NA

DOI: 10.1002/sim

Table III. Parameter estimates from bivariate Mendelian randomization meta-analysis models using studies with complete and incomplete outcomes.

Parameter	*Gg* versus *gg* Estimate (95 per cent CI) ($n=18$)	*GG* versus *gg* Estimate (95 per cent CI) ($n=18$)
η	-1.08 $(-1.76, -0.39)$	-0.86 $(-1.39, -0.33)$
$\exp(\eta)$	0.34 (0.17, 0.68)	0.42 (0.25, 0.72)
δ_2	-0.44 $(-0.59, -0.28)$	NA
δ_3	NA	-0.90 $(-1.42, -0.38)$
τ_2^2	0.02 (0.001, 2.27)	NA
τ_3^2	NA	0.56 (0.16, 1.96)

Table IV. Parameter estimates from bivariate genetic model-free meta-analysis models.

Parameter	Gene–disease Estimate (95 per cent CI) ($n=13$)	Gene–phenotype Estimate (95 per cent CI) ($n=15$)
λ	0.44 (0.19, 0.64)	0.42 (0.08, 0.67)
θ_3	0.96 (0.50, 1.43)	NA
$\exp(\theta_3)$	2.62 (1.65, 4.16)	NA
$\tau_{\theta_3}^2$	Fixed at 0.31	NA
δ_3	NA	-0.88 $(-1.40, -0.37)$
τ_3^2	NA	0.48 (0.10, 2.31)

Figure 2 shows the diagnostic plot to assess the pooled estimate of η with the gene–phenotype outcome measures on the x-axis and the gene–disease outcome measures on the y-axis. Given that two genotype comparisons are assessed, each study can contribute two points to the plot. A line with gradient equal to the pooled estimate of η is drawn on the plot to help assess the fit of the model. Only one point did not lie within one standard deviation of the fitted line. Figure 2 also shows that the point estimates from the *GG* versus *gg* comparison have greater between-study heterogeneity because the point estimates are spread over a wider range, and they are less precise than the point estimates from the *Gg* versus *gg* comparison.

Figure 3(a) and (b) assesses the estimated genetic model from the MVMR-GMF meta-analysis model. On both figures lines have been plotted with gradients equal to $\widehat{\lambda}$ from the MVMR-GMF model and 0.5 to represent the additive genetic model. For this meta-analysis these figures are sensitive to the fact that not all studies reported both sets of outcome measures and so not all studies are shown on each plot.

4. DISCUSSION AND CONCLUSIONS

In observational epidemiology estimates from a Mendelian randomization analysis can provide improved estimates of the association between a biological phenotype and a disease compared with direct estimates of this association. The proposed meta-analysis models extend previous literature

DOI: 10.1002/sim

by incorporating both genotype comparisons for a given genetic polymorphism into the same model. The MVMR-GMF and PNF meta-analysis models also incorporate the estimation of the underlying genetic model for the risk allele in a Mendelian randomization analysis.

The proposed meta-analysis models rely on two important assumptions, namely that the phenotype–disease association is the same in the *Gg* versus *gg* and the *GG* versus *gg* genotype comparisons and that the underlying genetic model is the same in the gene–phenotype and gene–disease associations. These assumptions are assessed in Figures 2 and 3. The modelling approach could be extended to allow the phenotype–disease log odds ratio, η, to vary across studies; this would most easily be implemented using Bayesian methodology. Figure 1 shows a four-column forest plot for a Mendelian randomization meta-analysis across two genotype comparisons. From the plot the relative precision of the estimates from the two genotype comparisons and the patterns in the estimates of individual studies can be assessed.

Incorporating multiple genotype comparisons into a Mendelian randomization analysis is advantageous because the comparison of the heterozygotes with the common homozygotes has the larger sample size, while the comparison of the rare homozygotes with the common homozygotes has the larger difference in disease risk. Therefore, the pooled estimates of the phenotype–disease association from the MVMR, MVMR-GMF, and PNF models in Table II were between the estimates for the two separate bivariate meta-analysis models using single genotype comparisons in Table III. The pooled estimate of the phenotype–disease association in the MVMR and MVMR-GMF models also showed increased precision over the single genotype comparison models because they included more information. Another advantage of incorporating all three genotypes is that if some of the studies omit to report either genotype–phenotype or genotype–disease outcome measures, then they can be accommodated in the meta-analysis model using the appropriate bivariate normal likelihood. This requires the additional assumption that the missing outcomes were missing at random and not missing for a systematic reason such as reporting bias.

The estimation of the underlying genetic model for the risk allele, known as the genetic model-free approach, can also be incorporated within this meta-analysis framework. The proposed approach extends previous literature through the joint synthesis of the genotype–disease and genotype–phenotype information to estimate the genetic model. This means that no strong assumptions about the genetic model are required prior to the analysis. In the example meta-analysis the genetic model was estimated close to the additive genetic model. The interpretation of estimates of λ not at one of the standard genetic models has been discussed elsewhere [7].

The estimation of bivariate meta-analysis models has been shown to be problematic when correlation parameters are near ± 1 [16, 23–25]. To overcome this problem an alternative form of the marginal distribution for a multivariate meta-analysis model has been proposed, which assumes a common correlation term both within and between studies; see model A in [15] or [25]. The advantage of this alternative covariance structure is that only study outcome measures and their respective variances are required to fit the multivariate meta-analysis model; the same information is required to perform the univariate meta-analyses for each outcome measure separately. A further discussion of how the relative magnitudes of the within- and between-study covariance matrices can affect parameter estimates in multivariate meta-analysis models is provided by Ishak *et al.* [26]. To fit multivariate meta-analysis models, the restricted log-likelihood could be used in the maximization as an alternative to the log-likelihood [25].

It would be possible to use these and the previously proposed bivariate meta-analysis models for Mendelian randomization studies reporting continuous disease outcome measures since the models assume that the log odds ratios are continuous and normally distributed. For case–control

DOI: 10.1002/sim

studies it would be possible to achieve similar pooled estimates of the phenotype–disease log odds ratio across two genotype comparisons using either a retrospective or a prospective likelihood for the genotype–disease outcome measures, which has previously been demonstrated for the genetic model-free approach [8]. Meta-analysis models have been used to estimate other parameters of interest from genetic data. For example, meta-regression has been used to investigate deviations from Hardy–Weinberg equilibrium [27] and merged genotype comparisons have been used to assess Hardy–Weinberg equilibrium and estimate the genetic model-free approach [28]. The work presented here also has parallels with modelling baseline risk in meta-analyses [29, 30].

The limitations that apply to the analysis of a single study using Mendelian randomization also apply to each of the studies in the meta-analysis. Therefore, it is important to assess that the selected genotype fulfills the conditions of an instrumental variable [10] and whether any of the factors that could potentially affect Mendelian randomization analyses such as pleiotropy or canalization are present [31]. Some further issues relating to the causal interpretation of meta-analyses of Mendelian randomization studies have been discussed by Nitsch *et al*. [32].

In conclusion, estimating the phenotype–disease association using separate genotype comparisons is often limited in that the comparison of the homozygote genotypes has a smaller sample size, whereas the comparison of the heterozygotes with the common homozygotes involves a smaller difference in disease risk. Pooling the phenotype–disease association across these comparisons produces an estimate that is a weighted average of the two but with increased precision. This meta-analysis framework can incorporate the estimation of the genetic model-free approach so that no strong prior assumptions about the underlying genetic model are required.

ACKNOWLEDGEMENTS

Tom Palmer is funded by a Medical Research Council capacity building Ph.D. Studentship in Genetic Epidemiology (G0501386). Martin Tobin is funded by a Medical Research Council Clinician Scientist Fellowship (G0501942). John Thompson receives support from an MRC Program Grant (G0601625) addressing causal inference in Mendelian randomization. Tom Palmer would like to thank the ISCB subcommittee on Student Conference Awards for the receipt of a Student Conference Award to attend ISCB28. The authors would like to thank two anonymous referees whose comments greatly improved the paper.

REFERENCES

1. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1986; **327**:507–508.
2. Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology* 2003; **32**:1–22. DOI: 10.1093/ije/dyg070.
3. Bowden RJ, Turkington DA. *Instrumental Variables*. Cambridge University Press: Cambridge, 1984.
4. Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000; **29**:722–729.
5. Davey Smith G, Harbord R, Fibrinogen ES. C-reactive protein and coronary heart disease: does Mendelian randomization suggest the associations are non-causal? *The Quarterly Journal of Medicine* 2004; **97**:163–166.
6. 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**(8):1133–1163. DOI: 10.1002/sim.3034.
7. Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. The choice of a genetic model in the meta-analysis of molecular association studies. *International Journal of Epidemiology* 2005; **34**:1319–1328.
8. Minelli C, Thompson JR, Abrams KR, Lambert PC. Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. *Statistics in Medicine* 2005; **24**:3845–3861.

9. Greene WH. *Econometric Analysis* (4th edn). Prentice-Hall: New York, 1999.
10. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 2007; **16**:309–330.
11. Thomas DC, Lawlor DA, Thompson JR. Re: estimation of bias in nongenetic observational studies using 'Mendelian triangulation' by Bautista *et al*. *Annals of Epidemiology* 2007; **17**(7):511–513.
12. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian randomization'. *International Journal of Epidemiology* 2004; **33**:21–25.
13. Minelli C, Thompson JR, Tobin MD, Abrams KR. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology* 2004; **160**(5):445–452.
14. Thompson JR, Tobin MD, Minelli C. GE1: on the accuracy of estimates of the effect of phenotype on disease derived from Mendelian randomisation studies. *Technical Report*, University of Leicester, Leicester, 2003. Available from: http://www2.le.ac.uk/departments/health-sciences/extranet/BGE/genetic-epidemiology/genepi_tech_reports.
15. 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.
16. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
17. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
18. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. Available from: http://www.R-project.org. ISBN 3-900051-07-0.
19. Spiegelhalter DJ. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics* 1998; **47**(1):115–133.
20. Mann V, Hobson EE, Li B, Stewart TL, Grant SFA, Robins SP, Aspden RM, Ralston SH. A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *The Journal of Clinical Investigation* 2001; **107**(7):899–907.
21. Grant SFA, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH. Reduced bone density and osteoporosis associated with a polymorphic Sp 1 binding site in the collagen type I α 1 gene. *Nature Genetics* 1996; **14**(2):203–205.
22. Uitterlinden A, Burger H, Huang Q, Yue F, McGuigan F, Grant S, Hofman A, van Leeuwen J, Pols H, Ralston S. Relation of alleles of the collagen type I alpha 1 gene to bone density and the risk of osteoporotic fractures in postmenopausal women. *New England Journal of Medicine* 1998; **338**(15):1016–1021.
23. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97.
24. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
25. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; **9**(1):172–186.
26. Ishak KJ, Platt RW, Joseph L, Hanley JA. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine* 2008; **27**:670–686.
27. Salanti G, Higgins JPT, Trikalinos TA, Ioannidis JPA. Bayesian meta-analysis and meta-regression for gene–disease associations and deviations from Hardy–Weinberg equilibrium. *Statistics in Medicine* 2007; **26**:553–567.
28. Salanti G, Higgins JPT. Meta-analysis of genetic association studies under different inheritance models using data reported as merged genotypes. *Statistics in Medicine* 2008; **27**(5):764–777. DOI: 10.1002/sim.2919.
29. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741–2758.
30. van Houwelingen HC, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(1):110–115.
31. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* 2004; **33**(1):30–42.
32. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *American Journal of Epidemiology* 2006; **163**:397–403.