Meta-analysis of Mendelian randomization studies

Tom Palmer, John Thompson and Martin Tobin

Department of Health Sciences,
University of Leicester

12 April 2007

Outline

Introduction

- Mendelian randomization is an active area of research in genetic-epidemiology.
- Aim: To extend existing meta-analysis models

Mendelian Randomization

- Dates back to [Katan, 1986]
- Recent interest due to the increasing use of genetic data in epidemiology [Katan, 2004]
- Bi-allelic polymorphism - receive one allele from each parent
- Mendel's 2nd law: genes segregate independently
- Therefore individuals randomized to a genotype at conception
- Randomization by genotype is independent of confounding factors

Introduction

- Mendelian randomization is an active area of research in genetic-epidemiology.
- Aim: To extend existing meta-analysis models

Mendelian Randomization

- Dates back to [Katan, 1986]
- Recent interest due to the increasing use of genetic data in epidemiology [Katan, 2004]
- Bi-allelic polymorphism - receive one allele from each parent
- Mendel's 2nd law: genes segregate independently
- Therefore individuals randomized to a genotype at conception
- Randomization by genotype is independent of confounding factors

Introduction

- Mendelian randomization is an active area of research in genetic-epidemiology.
- Aim: To extend existing meta-analysis models

Mendelian Randomization

- Dates back to [Katan, 1986]
- Recent interest due to the increasing use of genetic data in epidemiology [Katan, 2004]
- Bi-allelic polymorphism - receive one allele from each parent
- Mendel's 2^{nd} law: genes segregate independently
- Therefore individuals randomized to a genotype at conception
- Randomization by genotype is independent of confounding factors

Introduction

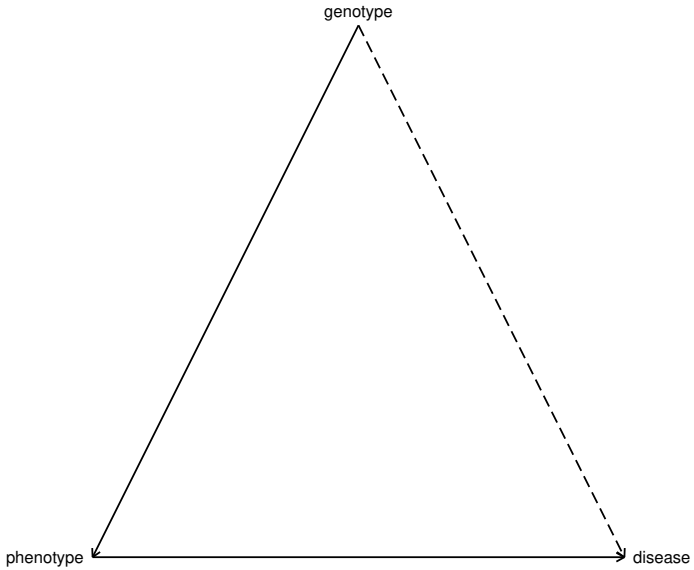
- Mendelian randomization is an active area of research in genetic-epidemiology.
- Aim: To extend existing meta-analysis models

Mendelian Randomization

- Dates back to [Katan, 1986]
- Recent interest due to the increasing use of genetic data in epidemiology [Katan, 2004]
- Bi-allelic polymorphism - receive one allele from each parent
- Mendel's 2^{nd} law: genes segregate independently
- Therefore individuals randomized to a genotype at conception
- Randomization by genotype is independent of confounding factors

- Estimate phenotype-disease effect
- Confounding
- Reverse causation
- [Davey Smith et al., 2005]; phenotype - C-Reactive Protein, disease - hypertension, genetic polymorphism - in the human CRP gene
- Statistically the genotype used as an instrumental variable
- Economics, IVs also applied to;
  - clinical trials [Angrist et al., 1996].
  - causal inference literature [Didelez and Sheehan, 2005]

- Estimate phenotype-disease effect
- Confounding
- Reverse causation
- [Davey Smith et al., 2005]; phenotype - C-Reactive Protein, disease - hypertension, genetic polymorphism - in the human CRP gene
- Statistically the genotype used as an instrumental variable
- Economics, IVs also applied to;
  - clinical trials [Angrist et al., 1996].
  - causal inference literature [Didelez and Sheehan, 2005]



- Use gene-disease & gene-phenotype effect estimates to estimate the phenotype-disease relationship
- Standard IV technique if they were all linear - TSLS
- gene-disease log odds-ratio: θ, difference in mean phenotypes: δ, phenotype-disease log odds-ratio: η
- Ratio of coefficients approach [Thomas and Conti, 2004], for a k-unit change in the mean phenotype difference,

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

Information from a case-control study

- A biallellic polymorphism (g,G)
  g: common allele G: risk allele
- 3 genotypes: gg, Gg, GG; $j = 1, 2, 3$
- Observed cases and controls y_{dj}, $d = 0,1$; control/case
- cell probabilities p_{dj}

	Genotype		
	gg	Gg	GG
Controls	y_{01}, p_{01}	y_{02}, p_{02}	y_{03}, p_{03}
Cases	y_{11}, p_{11}	y_{12}, p_{12}	y_{13}, p_{13}
Mean phenotype levels	μ_1	μ_2	μ_3

- Mean phenotype levels from controls

Information from a case-control study

- A biallellic polymorphism (g,G)
  g: common allele G: risk allele
- 3 genotypes: gg, Gg, GG; $j = 1, 2, 3$
- Observed cases and controls y_{dj}, $d = 0,1$; control/case
- cell probabilities p_{dj}

	Genotype		
	gg	Gg	GG
Controls	y_{01}, p_{01}	y_{02}, p_{02}	y_{03}, p_{03}
Cases	y_{11}, p_{11}	y_{12}, p_{12}	y_{13}, p_{13}
Mean phenotype levels	μ_1	μ_2	μ_3

- Mean phenotype levels from controls

Information from a case-control study

- A biallellic polymorphism (g,G)
  g: common allele G: risk allele
- 3 genotypes: gg, Gg, GG; $j = 1, 2, 3$
- Observed cases and controls y_{dj}, $d = 0,1$; control/case
- cell probabilities p_{dj}

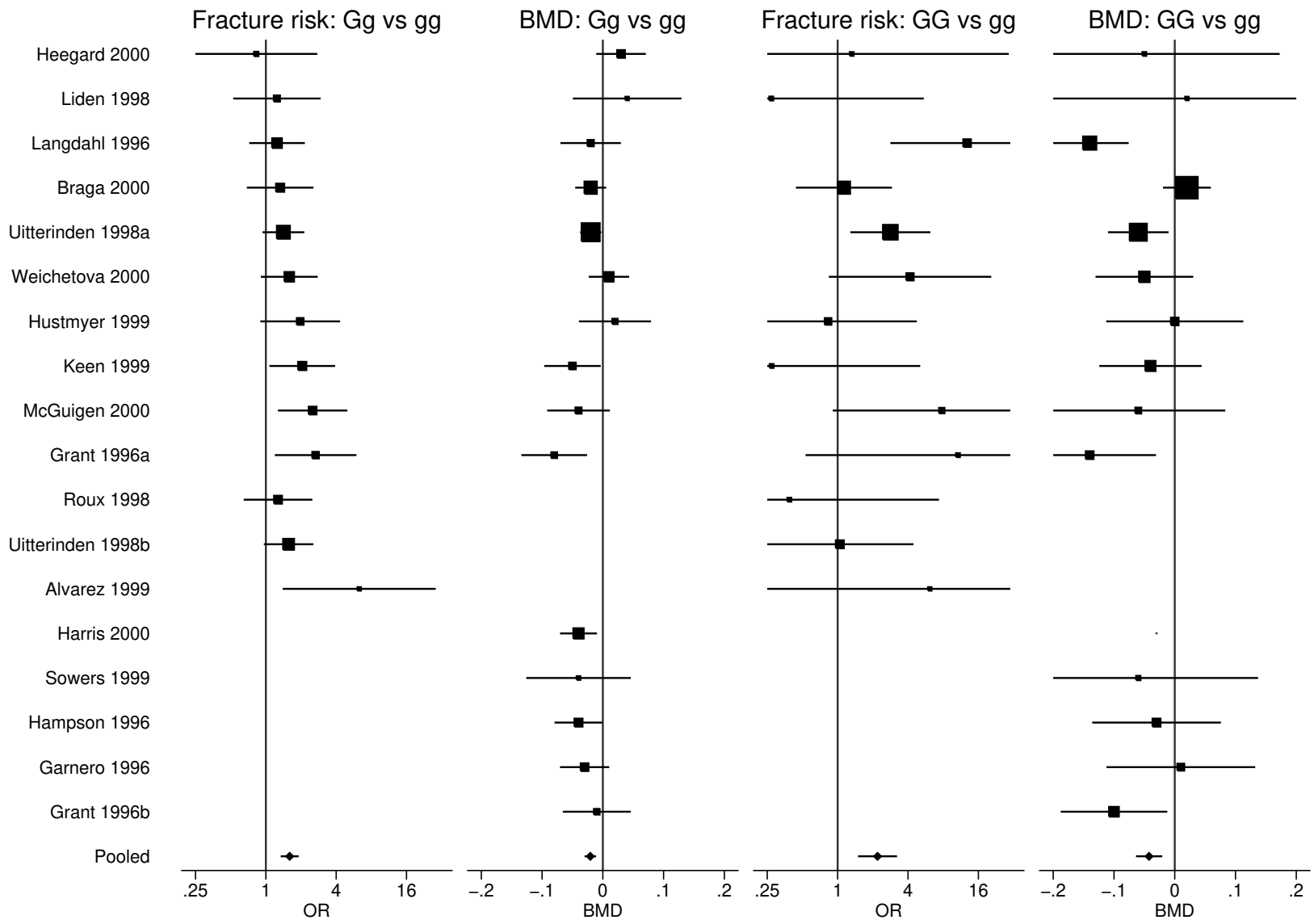
	Genotype		
	gg	Gg	GG
Controls	y_{01}, p_{01}	y_{02}, p_{02}	y_{03}, p_{03}
Cases	y_{11}, p_{11}	y_{12}, p_{12}	y_{13}, p_{13}
Mean phenotype levels	μ_1	μ_2	μ_3

- Mean phenotype levels from controls

Example meta-analysis

- Mann (2001): Bone mineral denisty (BMD) & risk of osteoporotic fracture
- *COL1A1* gene: codes for collagen
- Average BMD lower for GG versus gg
- Risk of fracture increased for GG versus gg

Meta-analysis results in a four column forest plot



Approach

- Existing meta-analysis models estimate η based on either the Gg versus gg genotype comparison or the GG versus gg comparison, [Thompson et al., 2005].
- Gg vs gg: Bigger sample size; smaller difference in disease risk
- GG vs gg: Smaller sample size; bigger difference in disease risk
- Proposed approach: Estimate η across both genotype comparisons

Modelling assumptions

- phenotype-disease relationship common across studies
- phenotype-disease relationship common across genotype comparisons

Approach

- Existing meta-analysis models estimate η based on either the Gg versus gg genotype comparison or the GG versus gg comparison, [Thompson et al., 2005].
- Gg vs gg: Bigger sample size; smaller difference in disease risk
- GG vs gg: Smaller sample size; bigger difference in disease risk
- Proposed approach: Estimate η across both genotype comparisons

Modelling assumptions

- phenotype-disease relationship common across studies
- phenotype-disease relationship common across genotype comparisons

Multivariate meta-analysis models

- Genotype comparison 2:(Gg,gg), 3:(GG,gg) for study i
  $(\theta_{2i}, \theta_{3i})$: gene-disease log odds-ratios
  $(\delta_{2i}, \delta_{3i})$: difference in mean phenotypes
- Inference at the population level
- Marginal distribution: combine within and between study distributions

Multivariate meta-analysis models

- Genotype comparison 2:(Gg,gg), 3:(GG,gg)
  for study i
  $(\theta_{2i}, \theta_{3i})$: gene-disease log odds-ratios
  $(\delta_{2i}, \delta_{3i})$: difference in mean phenotypes
- Inference at the population level
- Marginal distribution: combine within and between study distributions

$$\begin{bmatrix} \theta_{2i} \\ \delta_{2i} \\ \theta_{3i} \\ \delta_{3i} \end{bmatrix} \sim \text{MVN} \left(\underline{\psi} = \begin{bmatrix} \eta\delta_2 \\ \delta_2 \\ \eta\delta_3 \\ \delta_3 \end{bmatrix}, \mathbf{V}_i + \mathbf{B} \right).$$

$$\mathbf{V}_i = \begin{bmatrix} \text{v}(\theta_{2i}) & 0 & \text{v}(\theta_{2i}, \theta_{3i}) & 0 \\ 0 & \text{v}(\delta_{2i}) & 0 & \text{v}(\delta_{2i}, \delta_{3i}) \\ \text{v}(\theta_{3i}, \theta_{2i}) & 0 & \text{v}(\theta_{3i}) & 0 \\ 0 & \text{v}(\delta_{3i}, \delta_{2i}) & 0 & \text{v}(\delta_{3i}) \end{bmatrix}.$$

$$\mathbf{B} = \begin{bmatrix} \eta^2\tau_2^2 & \eta\tau_2^2 & \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho \\ \eta\tau_2^2 & \tau_2^2 & \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho \\ \eta^2\tau_2\tau_3\rho & \eta\tau_2\tau_3\rho & \eta^2\tau_3^2 & \eta\tau_3^2 \\ \eta\tau_2\tau_3\rho & \tau_2\tau_3\rho & \eta\tau_3^2 & \tau_3^2 \end{bmatrix}.$$

τ_2^2 between-study variance of the δ_{2i}'s
τ_3^2 between-study variance of the δ_{3i}'s
ρ between-study correlation between the δ_{2i}'s and the δ_{3i}'s

Maximum likelihood estimation

- Log-likelihood of the multivariate Normal distribution,

$$\log L \propto \sum_{i=1}^{n} -\frac{1}{2}\log(\det(\mathbf{V}_i+\mathbf{\Sigma}))-\frac{1}{2}(\underline{x_i}-\underline{\psi})^T(\mathbf{V}_i+\mathbf{\Sigma})^{-1}(\underline{x_i}-\underline{\psi})$$

- Maximisation using the Newton-Raphson algorithm
- Argument for using REML form of the likelihood for marginal models

Maximum likelihood estimation

- Log-likelihood of the multivariate Normal distribution,

$$\log L \propto \sum_{i=1}^{n} -\frac{1}{2}\log(\det(\mathbf{V}_i+\mathbf{\Sigma}))-\frac{1}{2}(\underline{x_i}-\underline{\psi})^T(\mathbf{V}_i+\mathbf{\Sigma})^{-1}(\underline{x_i}-\underline{\psi})$$

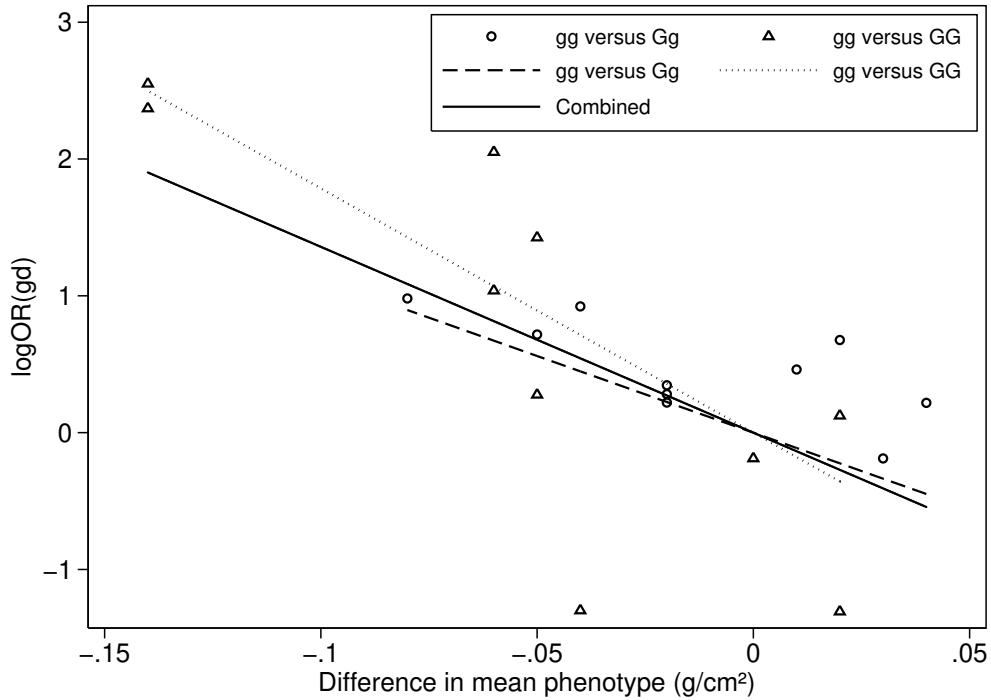
- Maximisation using the Newton-Raphson algorithm
- Argument for using REML form of the likelihood for marginal models

Results

Method of estimation	$OR_{pd,0.05}$	95% C.I./Cr.I.	
Gg vs gg	0.57	0.42	0.77
GG vs gg	0.40	0.28	0.57
Combined	0.50	0.39	0.62

- Gg vs gg expecting narrower CI - but wider
- GG vs gg bigger difference in disease risk - OR_{pd} further from 1
- combined model - weighted average of the separate estimates, with a narrower CI due to increased number of studies
- All results qualitatively the same
- 0.05 unit increase in BMD, implies typical patient at 40% risk of Osteoporotic fracture

Assessment of a common phenotype-disease odds-ratio



- MR assumptions fit straight line through the origin
- η gradient of the line

Incorporating the genetic model-free approach

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

- Interpretation of λ

λ	Genetic model
0	Recessive
0.5	Co-dominant
1	Dominant
> 1	Over-dominant, heteresis

- Meta-analysis models to estimate λ, [Minelli et al., 2005].

$$\begin{bmatrix} \theta_{2i} \\ \delta_{2i} \\ \theta_{3i} \\ \delta_{3i} \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} \eta\lambda\delta \\ \lambda\delta \\ \eta\delta \\ \delta \end{bmatrix}, \mathbf{V}_i + \mathbf{\Sigma} \right),$$

$$\mathbf{\Sigma} = \begin{bmatrix} \eta^2\lambda^2\tau^2 & \eta\lambda^2\tau^2 & \eta^2\lambda\tau^2 & \eta\lambda\tau^2 \\ \eta\lambda^2\tau^2 & \lambda^2\tau^2 & \lambda\eta\tau^2 & \lambda\tau^2 \\ \eta^2\lambda\tau^2 & \lambda\eta\tau^2 & \eta^2\tau^2 & \eta\tau^2 \\ \eta\lambda\tau^2 & \lambda\tau^2 & \eta\tau^2 & \tau^2 \end{bmatrix}$$

- τ^2 the between-study variance of the difference in mean phenotypes of the GG versus gg comparison

Bayesian estimation

- Product Normal Formulation [Spiegelhalter, 1998]
- 4 outcomes - univariate Normal distributions

$$\theta_{2i} \sim \text{N}(\eta\lambda\delta_i, v(\theta_{1i})), \qquad \delta_{2i} \sim \text{N}(\lambda\delta_i, v(\delta_{1i}))$$
$$\theta_{3i} \sim \text{N}(\eta\delta_i, v(\theta_{2i})), \qquad \delta_{3i} \sim \text{N}(\delta_i, v(\delta_{2i}))$$

- The correct covariances are induced in the model due to the relationships between the means and the sequential parameter updating under Gibbs sampling
- Prior distributions - vague

$$\delta_i \sim \text{N}(0, 1 \times 10^6), \quad \eta \sim \text{N}(0, 1 \times 10^6), \quad \lambda \sim \text{Beta}(0.5, 0.5)$$

Bayesian estimation

- Product Normal Formulation [Spiegelhalter, 1998]
- 4 outcomes - univariate Normal distributions

$$\theta_{2i} \sim \text{N}(\eta\lambda\delta_i, v(\theta_{1i})), \qquad \delta_{2i} \sim \text{N}(\lambda\delta_i, v(\delta_{1i}))$$
$$\theta_{3i} \sim \text{N}(\eta\delta_i, v(\theta_{2i})), \qquad \delta_{3i} \sim \text{N}(\delta_i, v(\delta_{2i}))$$

- The correct covariances are induced in the model due to the relationships between the means and the sequential parameter updating under Gibbs sampling
- Prior distributions - vague

$$\delta_i \sim \text{N}(0, 1 \times 10^6), \quad \eta \sim \text{N}(0, 1 \times 10^6), \quad \lambda \sim \text{Beta}(0.5, 0.5)$$

Results

Method of estimation	$OR_{pd,0.05}$	95% C.I./Cr.I.		λ	95% C.I./Cr.I.	
ML	0.42	0.28	0.61	0.33	0.19	0.47
Bayesian	0.46	0.32	0.61	0.30	0.17	0.45

- Genetic model between recessive and co-dominant

Summary

- Mendelian randomization - depends on random allocation of an individual's genotype
- Genotype used as an instrumental variable
- Meta-analysis model - joint analysis of two genotype comparisons
- Meta-analysis model - incorporating the genetic model-free approach


Angrist, J., Imbens, G., and Rubin, D. (1996).
Identification of causal effects using instrumental variables.
Journal of the American Statistical Association, 91(434):444–455.


Davey Smith, G., D.A.Lawlor, Harbord, R., Timpson, N., Rumley, A., Lowe, G., Day, I., and Ebrahim, S. (2005).
Association of C-Reactive Protein with Blood Pressure and Hypertension: Life Course Confounding and Mendelian Randomization Tests of Causality.
Arteriosclerosis, Thrombosis and Vascular Biology, 25:1051–1056.


Didelez, V. and Sheehan, N. (2005).
Mendelian randomisation and instrumental variables: What can and what can't be done.
University of Leicester, Department of Health Sciences Technical Report, 05-02.


Katan, M. (1986).
Apolipoprotein e isoforms, serum cholesterol, and cancer.
Lancet, 327:507–508.


Katan, M. (2004).
Commentary: Mendelian randomization, 18 years on.
International Journal of Epidemiology, 33(1):10–11.


Minelli, C., Thompson, J., Abrams, K., Thakkinstian, A., and Attia, J. (2005).
The choice of a genetic model in the meta-analysis of molecular association studies.
International Journal of Epidemiology, 34:1319–1328.


Spiegelhalter, D. (1998).
Bayesian graphical modelling: a case-study in monitoring health outcomes.
Applied Statistics, 47(1):115–133.


Thomas, D. and Conti, D. (2004).
Commentary: The concept of 'mendelian randomization'.
International Journal of Epidemiology, 33:21–25.


Thompson, J., Minelli, C., Abrams, K., Tobin, M., and Riley, R. (2005).
Meta-analysis of genetic studies using Mendelian randomization - a multivariate approach.
Statistics in Medicine, 24:2241–2254.