

# Meta-analysis of causal relationships using genetic instrumental variables

Stephen Burgess and Simon G. Thompson

MRC Biostatistics Unit, University of Cambridge

ISCB 2009, 23rd to 27th August 2009



# Outline

- ▶ Introduction to Mendelian randomization
- ▶ Introduction of Bayesian method through simulated examples
- ▶ Applying method in one study
- ▶ Applying method in multiple studies
- ▶ Conclusion

# Introduction to Mendelian randomization

- ▶ Mendelian randomization is a technique for using genes (G) as instrumental variables (IV) to assess the true causal association where direct experiment is not possible.

# Introduction to Mendelian randomization

- ▶ Mendelian randomization is a technique for using genes (G) as instrumental variables (IV) to assess the true causal association where direct experiment is not possible.
- ▶ We use the random allocation of genes at conception in an analogous way to treatment assignment in a randomized control trial.

# Introduction to Mendelian randomization

- ▶ Mendelian randomization is a technique for using genes ( $G$ ) as instrumental variables ( $IV$ ) to assess the true causal association where direct experiment is not possible.
- ▶ We use the random allocation of genes at conception in an analogous way to treatment assignment in a randomized control trial.
- ▶ We seek to estimate the causal effect of change in outcome ( $Y$ ) for unit increase in phenotype ( $X$ ) keeping all other factors ( $U$ ) equal.

## Instrumental variables

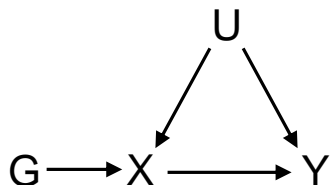


Figure: DAG of assumptions

G = gene  
X = phenotype  
Y = outcome  
U = confounders

## Instrumental variables

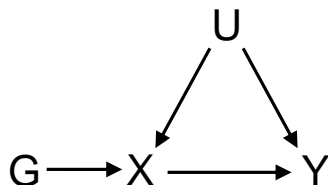


Figure: DAG of assumptions

G = gene  
X = phenotype  
Y = outcome  
U = confounders

Assumptions:

- i. the genotype is associated with the phenotype ( $G \dashv\vdash X$ ),

# Instrumental variables

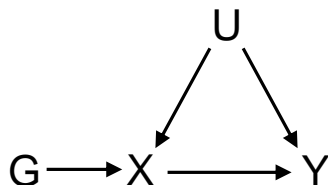


Figure: DAG of assumptions

G = gene  
X = phenotype  
Y = outcome  
U = confounders

Assumptions:

- i. the genotype is associated with the phenotype ( $G \not\perp X$ ),
- ii. the genotype is not associated with any confounders ( $G \perp U$ ),

# Instrumental variables

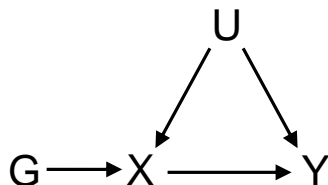


Figure: DAG of assumptions

G = gene  
X = phenotype  
Y = outcome  
U = confounders

Assumptions:

- i. the genotype is associated with the phenotype ( $G \not\perp X$ ),
- ii. the genotype is not associated with any confounders ( $G \perp U$ ),
- iii. the genotype is conditionally independent of the outcome given the phenotype ( $G \perp Y \mid X, U$ ).

# CRP CHD Genetic Collaboration

- ▶ C-reactive protein (CRP) is an acute-phase protein.

# CRP CHD Genetic Collaboration

- ▶ C-reactive protein (CRP) is an acute-phase protein.
- ▶ Associated with coronary heart disease (CHD), but this association attenuates on adjustment for confounders.

# CRP CHD Genetic Collaboration

- ▶ C-reactive protein (CRP) is an acute-phase protein.
- ▶ Associated with coronary heart disease (CHD), but this association attenuates on adjustment for confounders.
- ▶ CRP CHD genetic studies collaboration (CCGC) has 24+ studies:
  - ▶ measuring over 20 different SNPs, although different studies measure different subsets of these,
  - ▶ including cohort studies, case-cohort studies, prospective and retrospective case-control studies,
  - ▶ studies with and without individual CRP measurements . . .

# CRP CHD Genetic Collaboration

- ▶ C-reactive protein (CRP) is an acute-phase protein.
- ▶ Associated with coronary heart disease (CHD), but this association attenuates on adjustment for confounders.
- ▶ CRP CHD genetic studies collaboration (CCGC) has 24+ studies:
  - ▶ measuring over 20 different SNPs, although different studies measure different subsets of these,
  - ▶ including cohort studies, case-cohort studies, prospective and retrospective case-control studies,
  - ▶ studies with and without individual CRP measurements . . .
  - ▶ 100 000 subjects, 14 000 cases (so far!).

# CRP CHD Genetic Collaboration

- ▶ C-reactive protein (CRP) is an acute-phase protein.
- ▶ Associated with coronary heart disease (CHD), but this association attenuates on adjustment for confounders.
- ▶ CRP CHD genetic studies collaboration (CCGC) has 24+ studies:
  - ▶ measuring over 20 different SNPs, although different studies measure different subsets of these,
  - ▶ including cohort studies, case-cohort studies, prospective and retrospective case-control studies,
  - ▶ studies with and without individual CRP measurements . . .
  - ▶ 100 000 subjects, 14 000 cases (so far!).
  - ▶ How to include all of the data?

# Existing methodology

Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression
- ▶ We calculate the ratio of these associations as our causal estimate

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression
- ▶ We calculate the ratio of these associations as our causal estimate

## Two stage least squares:

- ▶ for multiple, polychotomous SNPs in one study with continuous outcomes

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression
- ▶ We calculate the ratio of these associations as our causal estimate

## Two stage least squares:

- ▶ for multiple, polychotomous SNPs in one study with continuous outcomes
- ▶ We calculate fitted values of X in the first stage G-X regression

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression
- ▶ We calculate the ratio of these associations as our causal estimate

## Two stage least squares:

- ▶ for multiple, polychotomous SNPs in one study with continuous outcomes
- ▶ We calculate fitted values of X in the first stage G-X regression
- ▶ We use these fitted values  $\hat{X}$  in a second stage X-Y regression

# Existing methodology

## Ratio method:

- ▶ can be used for one SNP in one study with continuous or binary outcomes
- ▶ We calculate the G-X and G-Y associations by regression
- ▶ We calculate the ratio of these associations as our causal estimate

## Two stage least squares:

- ▶ for multiple, polychotomous SNPs in one study with continuous outcomes
- ▶ We calculate fitted values of X in the first stage G-X regression
- ▶ We use these fitted values  $\hat{X}$  in a second stage X-Y regression
- ▶ Standard error is calculated using sandwich variance estimators

# Simulated example

Confounded association - for individual  $i$ :

$$x_i = \alpha_1 g_i + \alpha_2 u_i + \epsilon_{xi}$$

$$y_i = \beta_1 x_i + \beta_2 u_i + \epsilon_{yi}$$

$$u_i \sim \mathcal{N}(0, 1)$$

$$\epsilon_{xi}, \epsilon_{yi} \sim \mathcal{N}(0, \sigma^2)$$

$$g_i \in \{0, 1, 2\}$$

# Simulated example

Confounded association - for individual  $i$ :

$$x_i = 0.5 g_i + 1 u_i + \epsilon_{xi}$$

$$y_i = 2 x_i - 3 u_i + \epsilon_{yi}$$

$$u_i \sim \mathcal{N}(0, 1)$$

$$\epsilon_{xi}, \epsilon_{yi} \sim \mathcal{N}(0, 0.25)$$

$$g_i \in \{0, 1, 2\}$$

# Simulated example

Confounded association - for individual  $i$ :

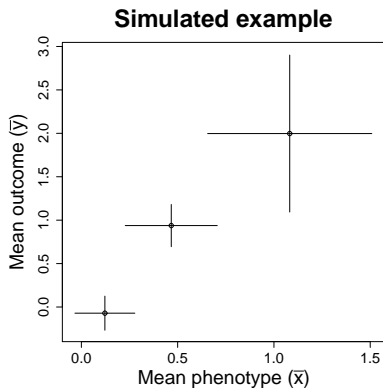
$$x_i = 0.5 g_i + 1 u_i + \epsilon_{xi}$$

$$y_i = 2 x_i - 3 u_i + \epsilon_{yi}$$

$$u_i \sim \mathcal{N}(0, 1)$$

$$\epsilon_{xi}, \epsilon_{yi} \sim \mathcal{N}(0, 0.25)$$

$$g_i \in \{0, 1, 2\}$$



## Simulated examples: Bayesian solution

Re-form the problem as  
regression with heterogeneous  
error in  $x$  - for genotypic group

$j$ :

## Simulated examples: Bayesian solution

Re-form the problem as  
regression with heterogeneous  
error in  $x$  - for genotypic group

$j$ :

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

## Simulated examples: Bayesian solution

Re-form the problem as  
regression with heterogeneous  
error in  $x$  - for genotypic group  
 $j$ :

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

- ▶ We estimate  $\sigma_{xj}^2$  and  $\sigma_{yj}^2$   
for data and set vague  
priors on all other  
parameters.

## Simulated examples: Bayesian solution

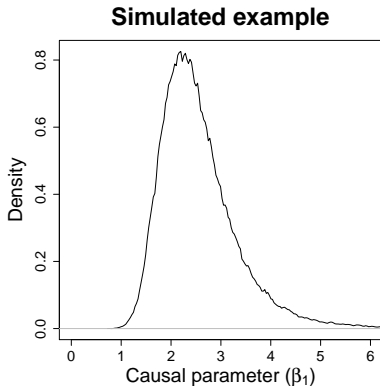
Re-form the problem as regression with heterogeneous error in  $x$  - for genotypic group  $j$ :

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

- ▶ We estimate  $\sigma_{xj}^2$  and  $\sigma_{yj}^2$  for data and set vague priors on all other parameters.
- ▶ Run in WinBUGS using MCMC sampling.



## Simulated examples: Results

(true value = 2)	Causal Estimate	95% CI
Weak (ratio)	1.637	0.563 to 6.582
Weak (Bayesian)	1.496	0.536 to 7.190
Moderate (ratio)	2.555	1.481 to 6.007
Moderate (Bayesian)	2.417	1.473 to 4.592
Strong (ratio)	2.139	1.814 to 2.554
Strong (Bayesian)	2.018	1.749 to 2.347

## Group-based method

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{x_j}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{y_j}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

## Group-based method

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

- ▶ We extend to use multiple genes, taking each genotype as a separate category in the stratification.

## Group-based method

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

- ▶ We extend to use multiple genes, taking each genotype as a separate category in the stratification.
- ▶ This gives a more detailed model of the G-X association.

## Group-based method

$$\bar{x}_j \sim \mathcal{N}(\xi_j, \sigma_{xj}^2)$$

$$\bar{y}_j \sim \mathcal{N}(\eta_j, \sigma_{yj}^2)$$

$$\eta_j = \beta_0 + \beta_1 \xi_j$$

- ▶ We extend to use multiple genes, taking each genotype as a separate category in the stratification.
  - ▶ This gives a more detailed model of the G-X association.
- ▶ If the size of groups are small, exact knowledge of  $\sigma_{xj}^2, \sigma_{yj}^2$  will not be valid.

# Individual- and additive-based methods

We can take population variances  $\sigma_x^2, \sigma_y^2$  and model X and Y on an individual level - for individual  $i$  in genotypic group  $j$ :

$$x_{ij} \sim \mathcal{N}(\xi_j, \sigma_x^2)$$

$$y_{ij} \sim \mathcal{N}(\eta_j, \sigma_y^2)$$

# Individual- and additive-based methods

We can take population variances  $\sigma_x^2, \sigma_y^2$  and model X and Y on an individual level - for individual  $i$  in genotypic group  $j$ :

$$x_{ij} \sim \mathcal{N}(\xi_j, \sigma_x^2)$$

$$y_{ij} \sim \mathcal{N}(\eta_j, \sigma_y^2)$$

If we want to introduce an additive model additive across SNPs - for individual  $i$  with  $G_{ik}$  variant alleles of SNP  $k$ :

$$\xi_i = \alpha_0 + \sum_k G_{ik} \alpha_k$$

$$x_i \sim \mathcal{N}(\xi_i, \sigma_x^2)$$

# Bayesian methodology vs Two stage least squares (2SLS)

- ▶ Both methods involve fitting a  $G-X$  regression, and then using these fitted values in a  $\hat{X}-Y$  regression

# Bayesian methodology vs Two stage least squares (2SLS)

- ▶ Both methods involve fitting a  $G-X$  regression, and then using these fitted values in a  $\hat{X}-Y$  regression
- ▶ Bayesian method fits the whole model simultaneously allowing feedback through the joint posterior

# Bayesian methodology vs Two stage least squares (2SLS)

- ▶ Both methods involve fitting a  $G-X$  regression, and then using these fitted values in a  $\hat{X}-Y$  regression
- ▶ Bayesian method fits the whole model simultaneously allowing feedback through the joint posterior
- ▶ 2SLS uses sandwich variance estimators making assumption of asymptotic normality

# Bayesian methodology vs Two stage least squares (2SLS)

- ▶ Both methods involve fitting a  $G-X$  regression, and then using these fitted values in a  $\hat{X}-Y$  regression
- ▶ Bayesian method fits the whole model simultaneously allowing feedback through the joint posterior
- ▶ 2SLS uses sandwich variance estimators making assumption of asymptotic normality
- ▶ Bayesian method uses MCMC sampling to find standard errors, confidence intervals

# Hierarchical meta-analysis

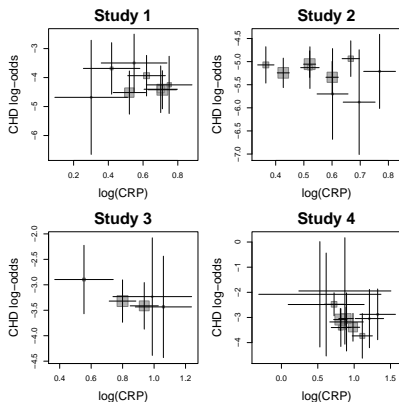
- ▶ As the causal association does not depend on the choice of genes, we impose a hierarchical model on it.

# Hierarchical meta-analysis

- ▶ As the causal association does not depend on the choice of genes, we impose a hierarchical model on it.
- ▶ This is valid when different genes or different numbers of genes are measured.

# Hierarchical meta-analysis

- ▶ As the causal association does not depend on the choice of genes, we impose a hierarchical model on it.
- ▶ This is valid when different genes or different numbers of genes are measured.



# Fixed/random-effect meta-analysis

Fixed-effect meta-analysis in  
group based method - for  
group  $j$ , study  $m$ :

# Fixed/random-effect meta-analysis

Fixed-effect meta-analysis in  
group based method - for  
group  $j$ , study  $m$ :

$$\begin{aligned}x_{jm} &\sim \mathcal{N}(\xi_{jm}, \sigma_{xjm}^2) \\y_{jm} &\sim \mathcal{N}(\eta_{jm}, \sigma_{yjm}^2) \\ \eta_{jm} &= \beta_{0m} + \beta_1 \xi_{jm} \quad (1)\end{aligned}$$

# Fixed/random-effect meta-analysis

Fixed-effect meta-analysis in  
group based method - for  
group  $j$ , study  $m$ :

$$\begin{aligned}x_{jm} &\sim \mathcal{N}(\xi_{jm}, \sigma_{x_{jm}}^2) \\y_{jm} &\sim \mathcal{N}(\eta_{jm}, \sigma_{y_{jm}}^2) \\ \eta_{jm} &= \beta_{0m} + \beta_1 \xi_{jm} \quad (1)\end{aligned}$$

Or for random-effect  
meta-analysis, line (1) is  
replaced by:

$$\begin{aligned}\eta_{jm} &= \beta_{0m} + \beta_{1m} \xi_{jm} \\ \beta_{1m} &\sim \mathcal{N}(\mu_\beta, \psi^2)\end{aligned}$$

# Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.

# Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group

# Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group
- ▶ ... allowing for an additive structure between these values if appropriate.

## Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group
- ▶ ... allowing for an additive structure between these values if appropriate.
- ▶ We simultaneously estimate the mean value of outcome ( $\eta_j$ ) in the group

## Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group
- ▶ ... allowing for an additive structure between these values if appropriate.
- ▶ We simultaneously estimate the mean value of outcome ( $\eta_j$ ) in the group
- ▶ ... under the constraint of a linear relationship between mean phenotype and mean outcome level ( $\eta_j = \beta_0 + \beta_1 \xi_j$ ).

# Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group
- ▶ ... allowing for an additive structure between these values if appropriate.
- ▶ We simultaneously estimate the mean value of outcome ( $\eta_j$ ) in the group
- ▶ ... under the constraint of a linear relationship between mean phenotype and mean outcome level ( $\eta_j = \beta_0 + \beta_1 \xi_j$ ).
- ▶ We set a hierarchical model on our causal parameter between studies.

# Summary of Bayesian methodology

- ▶ We stratify the population into genotypic groups.
- ▶ For each genotypic group, we estimate the mean value of phenotype ( $\xi_j$ ) in that group
- ▶ ... allowing for an additive structure between these values if appropriate.
- ▶ We simultaneously estimate the mean value of outcome ( $\eta_j$ ) in the group
- ▶ ... under the constraint of a linear relationship between mean phenotype and mean outcome level ( $\eta_j = \beta_0 + \beta_1 \xi_j$ ).
- ▶ We set a hierarchical model on our causal parameter between studies.
- ▶ We draw samples from the posterior distribution using WinBUGS.

## CCGC study

- ▶ We use three SNPs measured in the majority of studies.

# CCGC study

- ▶ We use three SNPs measured in the majority of studies.
- ▶ For participant  $i$  in genotypic group  $j$  with  $N_j$  participants,  $n_j$  cases, with  $G_{kjm}$  variant alleles of SNP  $k$  from study  $m$ :

$$\xi_{jm} = \alpha_{0m} + \alpha_1 G_{1jm} + \alpha_2 G_{2jm} + \alpha_3 G_{3jm}$$

$$x_{ijm} \sim \mathcal{N}(\xi_{jm}, \sigma_{xm}^2)$$

$$n_j \sim \mathcal{B}(N_j, \pi_j)$$

$$\eta_j = \text{logit}(\pi_j) = \beta_0 + \beta_1 \xi_j$$

# Conclusion

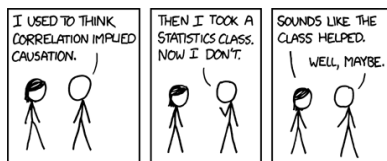
- ▶ The Bayesian method gives similar results to other established methods.

# Conclusion

- ▶ The Bayesian method gives similar results to other established methods.
- ▶ The Bayesian method is flexible to deal with situations existing methods cannot deal with:
- ▶ . . . meta-analysis, missing data, binary outcomes, uncertainty in haplotype assignment.

# References and Acknowledgements

1. Greenland S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* 2000;29:722-729.
2. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* 2007;16:309-330.
3. McKeigue P et al. Bayesian methods for instrumental variable analysis with genetic instruments ("Mendelian randomization"): example with urate transporter SLC2A9 as instrumental variable for effect of urate levels on metabolic syndrome. Accessed at <http://homepages.ed.ac.uk/pmckeigu/-mendelrand/instrumuric.pdf> on 04/03/09.
4. Baum CF, Schaffer ME, Stillman S. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 2003;3:1-31.
5. Munroe, R. "Correlation" xkcd - A Webcomic, 2009:552. Accessed at <http://xkcd.com/552/> on 06/04/09.



## Thanks to:

- ▶ The Medical Research Council
- ▶ The CRP CHD Genetics Collaboration and especially Frances Wensley for data collation.