Meta-analysis of individual participant data from observational studies

Simon Thompson

University of Cambridge, UK

4. Causal estimates

Freiburg, March 2013

Observational and causal associations of risk factors with disease

Associations in observational studies affected by:

- confounding
- measurement error
- reverse causation

Still useful for risk prediction

Causal relationships can be estimated using:

- randomized trials
- instrumental variables (e.g. genetic variants)

Essential for identifying treatment targets

Causal effect is change in the outcome given a change in the risk factor

C-reactive protein (CRP) and CHD

CRP is an acute-phase protein, a marker of inflammation, strongly associated with CHD in observational prospective epidemiological studies

IPD meta-analysis based on 54 prospective studies; 10,000 CHD events

Adjustments (usual level of confounders)	Hazard ratio per 1 SD increase in usual log CRP (95% CI)
Age, sex	1.68 (1.59 to 1.78)
+ SBP, smoking, diabetes, BMI, log TG, chol, HDL-C, alcohol	1.37 (1.27 to 1.48)
+ fibrinogen	1.23 (1.07 to 1.42)

ERFC, Lancet 2010

Is CRP causally related to CHD?

Genetic variants as instrumental variables = Mendelian Randomization (MR)

Genetic variants often have only small effects on a risk factor / phenotype

Precision of individual MR studies is low

Typically require meta-analysis of MR studies (especially to ‘show’ a null effect)

CRP CHD Genetics Collaboration (CCGC)

CCGC collated individual participant data (IPD):

43 studies (cross-sectional, case-control, prospective)
160,000 participants of European descent
36,000 CHD events (MI, CHD death)

Four pre-specified genetic variants (SNPs)*
Additional SNPs in some studies*
[* on the CRP-regulatory gene on chromosome 1]

Blood CRP concentrations in most studies

Aim: To estimate the causal effect of CRP on CHD as precisely as possible

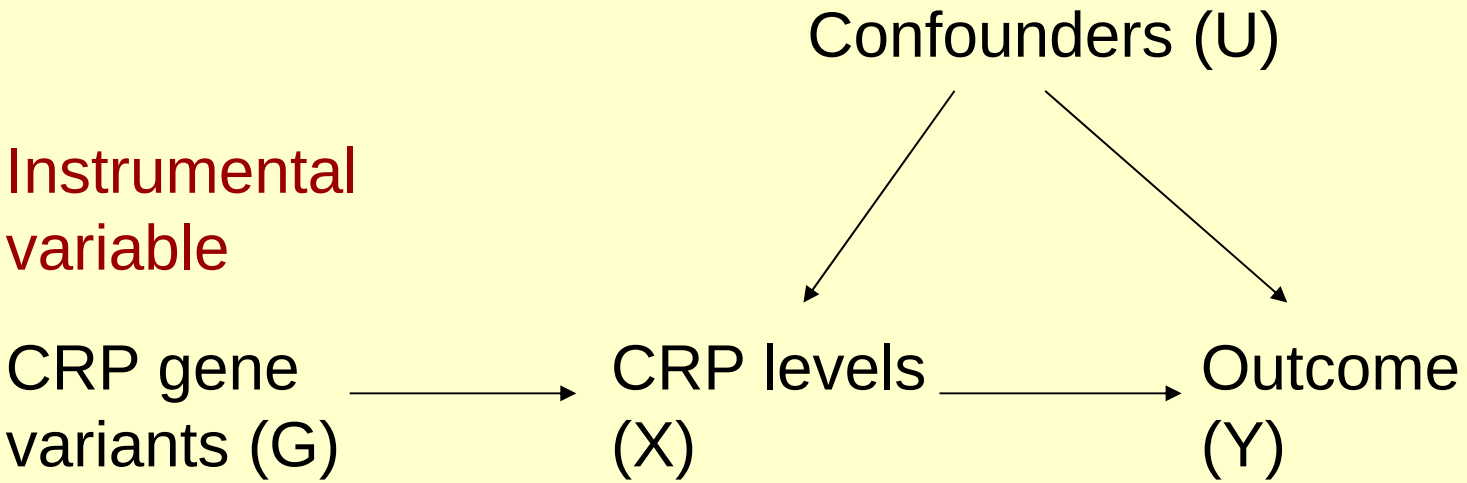
Outline of talk

1. One genetic variant in one study
2. Multiple genetic variants in one study
3. Multiple genetic variants in multiple studies

Issues:

4. Different study designs
5. Weak instrument bias
6. Lack of CRP measurements in some studies

Diagram of causal effects



Three crucial assumptions:

G affects X

G is not related to U

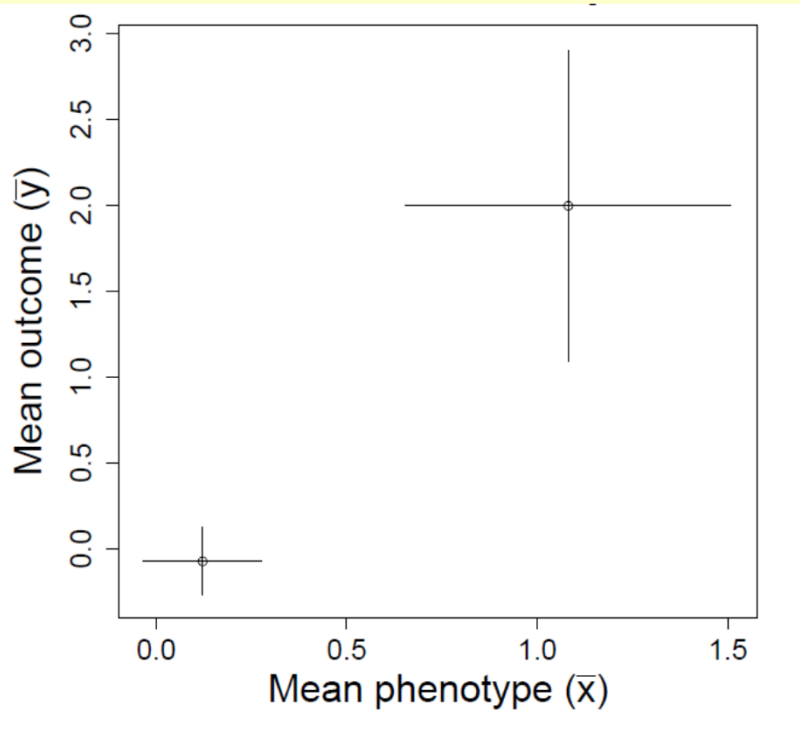
Y is conditionally independent of G given X and U

Conventional instrumental variable analysis (i)

2 genetic subgroups

Mean (95% CI) outcome and phenotype by genetic subgroup

Mean outcome = log odds of CHD



Ratio of coefficients method:

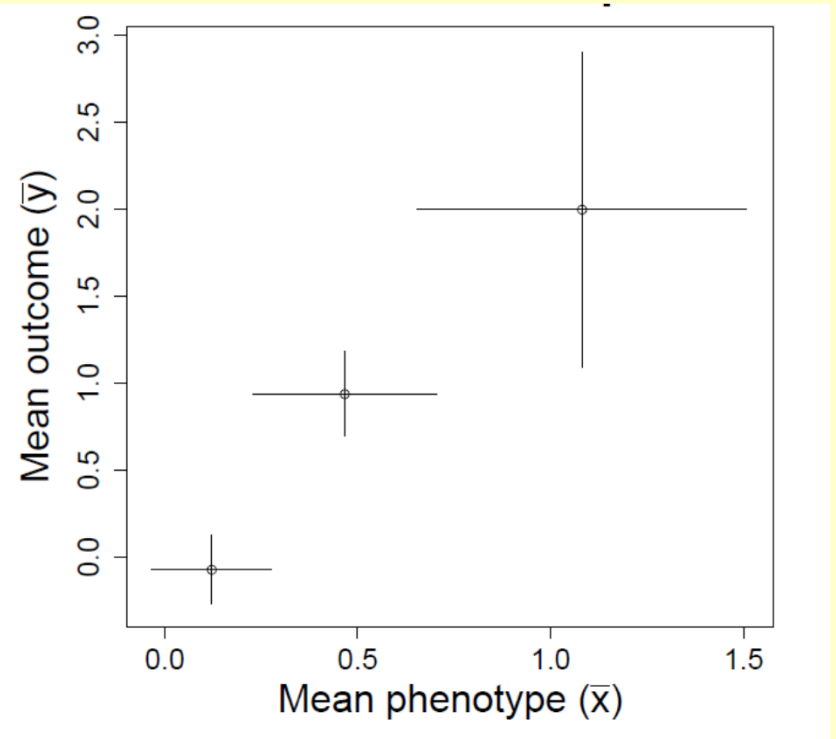
$$\text{causal effect} = \frac{\Delta \text{ log odds of CHD}}{\Delta \text{ mean phenotype}}$$

Conventional instrumental variable analysis (ii)

3 genetic subgroups

Mean (95% CI) outcome and phenotype by genetic subgroup (G=0,1,2)

Mean outcome = log odds of CHD



Two-stage method:

(i) Regress X on G = 0,1,2, giving X-pred
(ii) Regress Y on X-pred to estimate causal effect

A modelling approach

Prospective study of new incident CHD with fixed follow-up

Individual i in genetic subgroup j has phenotype x_{ij}
n_j events amongst N_j participants in subgroup j

Model at group level:

$$x_{ij} \sim N(\xi_j, \sigma^2)$$
$$n_j \sim Bin(N_j, \pi_j)$$
$$logit(\pi_j) = \beta_0 + \beta_1 \xi_j$$

β_1 is the causal effect estimate (increase in log odds of event per unit increase in phenotype)

Two-stage approach or one-stage approach

Multiple genetic markers in one study

Individual i has:

outcome $y_i = 0/1$

phenotype x_i

genetic variants $g_{ik}=0,1,2$ for $k=1\ldots K$ SNPs

Additive linear (per allele) model at individual level:

$$x_i \sim N(\xi_i, \sigma^2)$$

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

$$y_i \sim Bin(1, \pi_i)$$

$$logit(\pi_i) = \beta_0 + \beta_1 \xi_i$$

β_1 is the causal effect estimate

Multiple genetic markers in multiple studies

Individual i in study m has:

outcome y_{im} = 0/1, phenotype x_{im}

genetic variants g_{ikm}=0,1,2 for k=1…K_m

Additive linear (per allele) model at individual level:

$$x_{im} \sim N(\xi_{im}, \sigma_m^2)$$

$$\xi_{im} = \alpha_{0m} + \sum_k \alpha_{km} g_{ikm}$$

$$y_{im} \sim Bin(1, \pi_{im})$$

$$logit(\pi_{im}) = \beta_{0m} + \beta_{1m} \xi_{im}$$

$\beta_{1m} = \beta_1$ fixed-effect meta-analysis

$\beta_{1m} \sim N(\beta_1, \tau^2)$ random-effects meta-analysis

Bayesian implementation

Use grouped data as much as possible (computational efficiency)

Vague priors:

- Wide normal $N(0,100^2)$ on regression parameters
- Wide uniform U[0,20] on standard deviations

MCMC using WinBUGS

Propagates uncertainty from stage 1 to stage 2

Allows feedback from stage 2 to stage 1

'Estimate' = mean of posterior distribution

'SE' = SD of posterior distribution

'95% CI' = 2.5^{th} to 97.5^{th} percentile of posterior distribution

Different study designs

Cross-sectional prevalence study:

Use phenotype data in non-cases only

Retrospective and nested case-control studies:

Use phenotype data in controls only

Matched case-control studies:

Ignore matching

Check with sensitivity analysis

Prospective studies:

Ignore variable follow-up

Check with sensitivity analyses

Estimate of causal effect as population (marginal) log odds ratio in all studies

Can include individuals with missing phenotype data

A prospective study with both prevalent and incident cases

In genetic subgroup j :

N_{1j} individuals of whom n_{1j} are prevalent cases

N_{2j} ($=N_{1j}–n_{1j}$) non-prevalent individuals,

of whom n_{2j} have incident events

Model at group level:

$x_{ij} \sim N(\xi_j, \sigma^2)$ for $i=1…N_{2j}$ non-prevalent subjects

$n_{1j} \sim Bin(N_{1j}, \pi_{1j})$

$n_{2j} \sim Bin(N_{2j}, \pi_{2j})$

$logit(\pi_{1j}) = \beta_{01}+\beta_1\xi_j$

$logit(\pi_{2j}) = \beta_{02}+\beta_1\xi_j$

Estimate a single causal log odds ratio β_1

Weak instrument bias

Weak instruments

- explain little variation in phenotype
- small studies (finite sample bias)

F-statistic for regression of phenotype on genetic instrument(s) is a measure of instrument strength

Weak instruments give causal estimates biased in the direction of the observational association

Expected F-statistic >10 generally limits the bias in the causal estimate to less that 1/10 = 10% of the bias in the observational association

Addressing weak instrument bias in meta-analysis

Combine estimates of genetic effects on phenotype across studies which assessed the same SNPs:

$x_{im} \sim N(\xi_{im}, \sigma_m^2)$ in study m

$\xi_{im} = \alpha_{0m} + \sum_k \alpha_{km} g_{ikm}$

$\alpha_{km} = \alpha_k$ fixed-effect

$\alpha_{km} \sim N(\alpha_k, \tau_k^2)$ random-effects

CCGC (4 pre-specified SNPs):

g_1, g_2, g_3, g_4	20 studies
g_1, g_2, g_4	12 studies
g_1, g_2, g_3	5 studies
g_2	5 studies
other	1 study

Studies without phenotype data

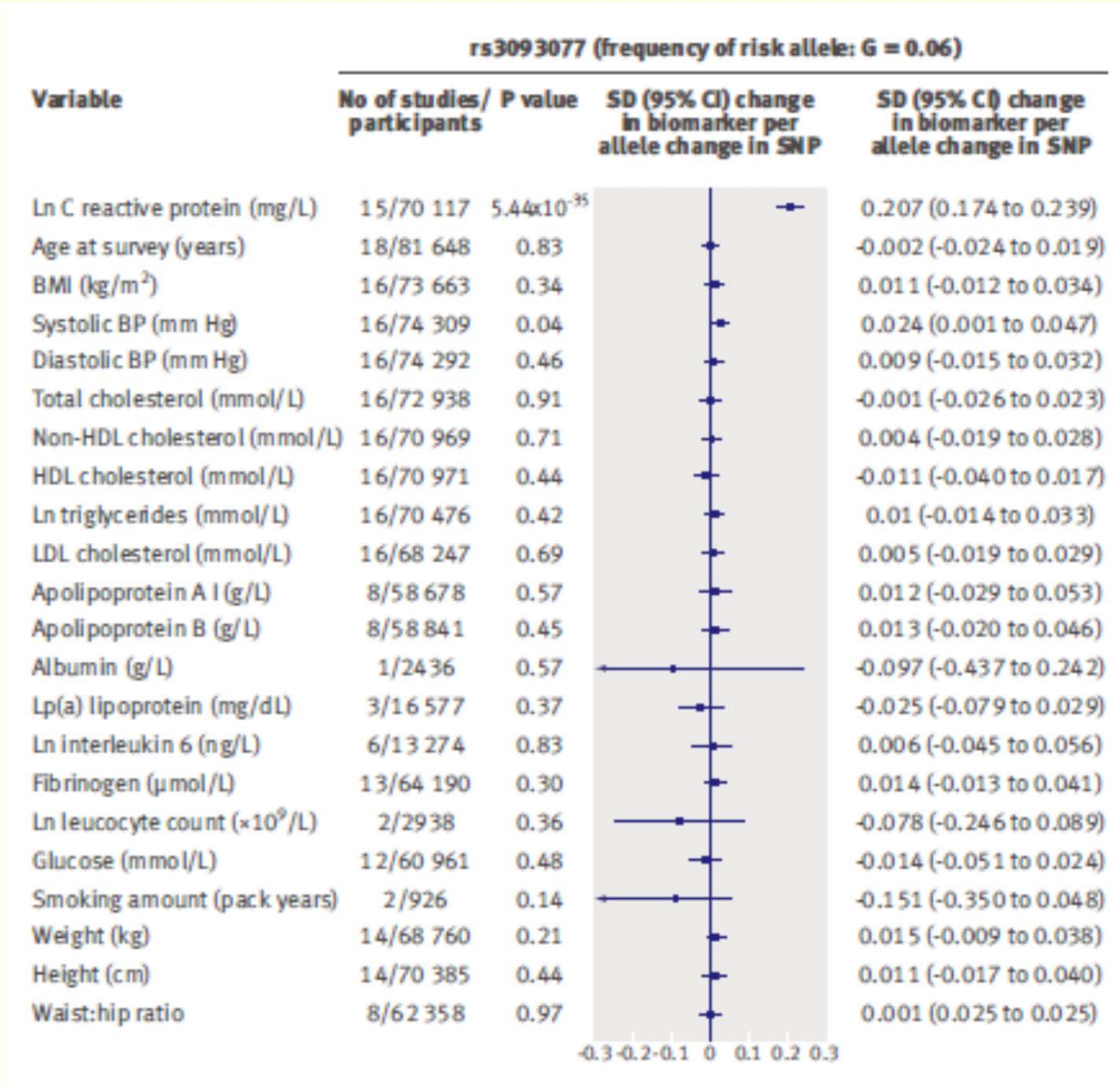
Study has genetic variants in common with other studies

Use the random-effects distributions for the genetic association parameters $\alpha_1 \ldots \alpha_K$ as a predictive distribution (implicit prior) for the unknown parameters

Requires an assumption of exchangeability: explicitly Bayesian

10 out of 43 studies had no CRP data

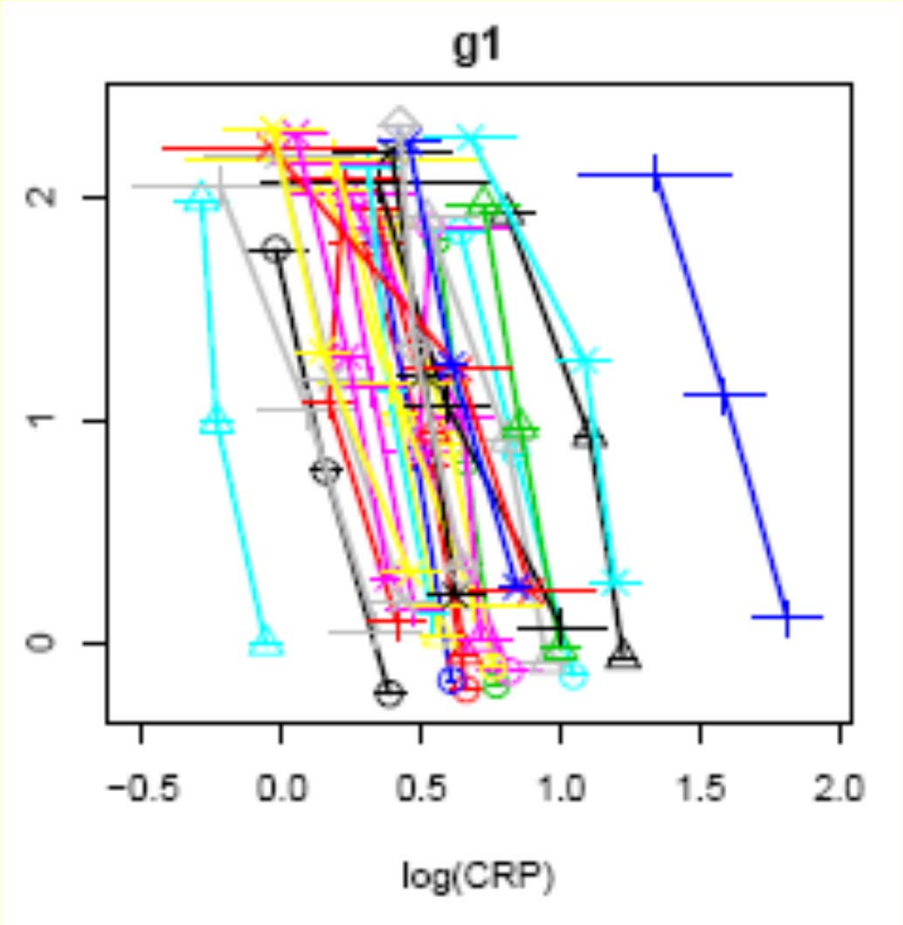
Are genetic variants instrumental variables?



rs3093077 (frequency of risk allele: G = 0.06)

Variable	No of studies/ participants	P value	SD (95% CI) change in biomarker per allele change in SNP
Ln C reactive protein (mg/L)	15/70 117	5.44×10^{-35}	0.207 (0.174 to 0.239)
Age at survey (years)	18/81 648	0.83	-0.002 (-0.024 to 0.019)
BMI (kg/m^2)	16/73 663	0.34	0.011 (-0.012 to 0.034)
Systolic BP (mm Hg)	16/74 309	0.04	0.024 (0.001 to 0.047)
Diastolic BP (mm Hg)	16/74 292	0.46	0.009 (-0.015 to 0.032)
Total cholesterol (mmol/L)	16/72 938	0.91	-0.001 (-0.026 to 0.023)
Non-HDL cholesterol (mmol/L)	16/70 969	0.71	0.004 (-0.019 to 0.028)
HDL cholesterol (mmol/L)	16/70 971	0.44	-0.011 (-0.040 to 0.017)
Ln triglycerides (mmol/L)	16/70 476	0.42	0.01 (-0.014 to 0.033)
LDL cholesterol (mmol/L)	16/68 247	0.69	0.005 (-0.019 to 0.029)
Apolipoprotein A I (g/L)	8/58 678	0.57	0.012 (-0.029 to 0.053)
Apolipoprotein B (g/L)	8/58 841	0.45	0.013 (-0.020 to 0.046)
Albumin (g/L)	1/2436	0.57	-0.097 (-0.437 to 0.242)
Lp(a) lipoprotein (mg/dL)	3/16 577	0.37	-0.025 (-0.079 to 0.029)
Ln interleukin 6 (ng/L)	6/13 274	0.83	0.006 (-0.045 to 0.056)
Fibrinogen (µmol/L)	13/64 190	0.30	0.014 (-0.013 to 0.041)
Ln leucocyte count ($\times 10^9$/L)	2/2938	0.36	-0.078 (-0.246 to 0.089)
Glucose (mmol/L)	12/60 961	0.48	-0.014 (-0.051 to 0.024)
Smoking amount (pack years)	2/926	0.14	-0.151 (-0.350 to 0.048)
Weight (kg)	14/68 760	0.21	0.015 (-0.009 to 0.038)
Height (cm)	14/70 385	0.44	0.011 (-0.017 to 0.040)
Waist:hip ratio	8/62 358	0.97	0.001 (0.025 to 0.025)

Is per-allele analysis reasonable?



Mean level of log CRP (95%CI) in non-diseased individuals in each study by number of variant alleles in g_1

Principal results of CCGC

Causal estimate = log odds ratio of CHD per unit increase in log CRP

	Studies / Cases	Causal est. (95%CI)	Heterog.
Two-stage classical analysis			
Logistic regression	33 / 24135	0.024 (-0.092 to 0.140)	I^2=13%
One-stage Bayesian analysis			
Pooled SNPs (all studies)	43 / 36463	-0.013 (-0.115 to 0.094)	τ=0.106

Interpretation of principal result

Estimate of β_1	-0.013
95% CI	(-0.115 to 0.094)
Estimate of τ	0.106

Overall OR per unit increase in log CRP:

0.99 (95%CI 0.89 to 1.10)

Overall OR per doubling in CRP:

0.99 (95%CI 0.92 to 1.07)

Predictive distribution for true OR in new study per doubling of CRP:

0.99 (95% range 0.84 to 1.16)

Conclusions

Provides a flexible framework for meta-analysis of MR studies:

- Multiple, different SNPs in each study
- Studies with prevalent and incident cases
- Studies without phenotype data
- Heterogeneity between studies
- Minimising weak instrument bias

References

Meta-analysis

Burgess S, Thompson SG. *Stat Med* 2010; 29: 1298-1311.

Weak instrument bias

Burgess S, Thompson SG. *Stat Med* 2011; 30: 1312-1323.

Burgess S, Thompson SG. *Int J Epid*, 2011; 40: 755-764.

CCGC methods

Burgess S, Thompson SG. *Stat Meth Med Res*, June 2012, e-pub.