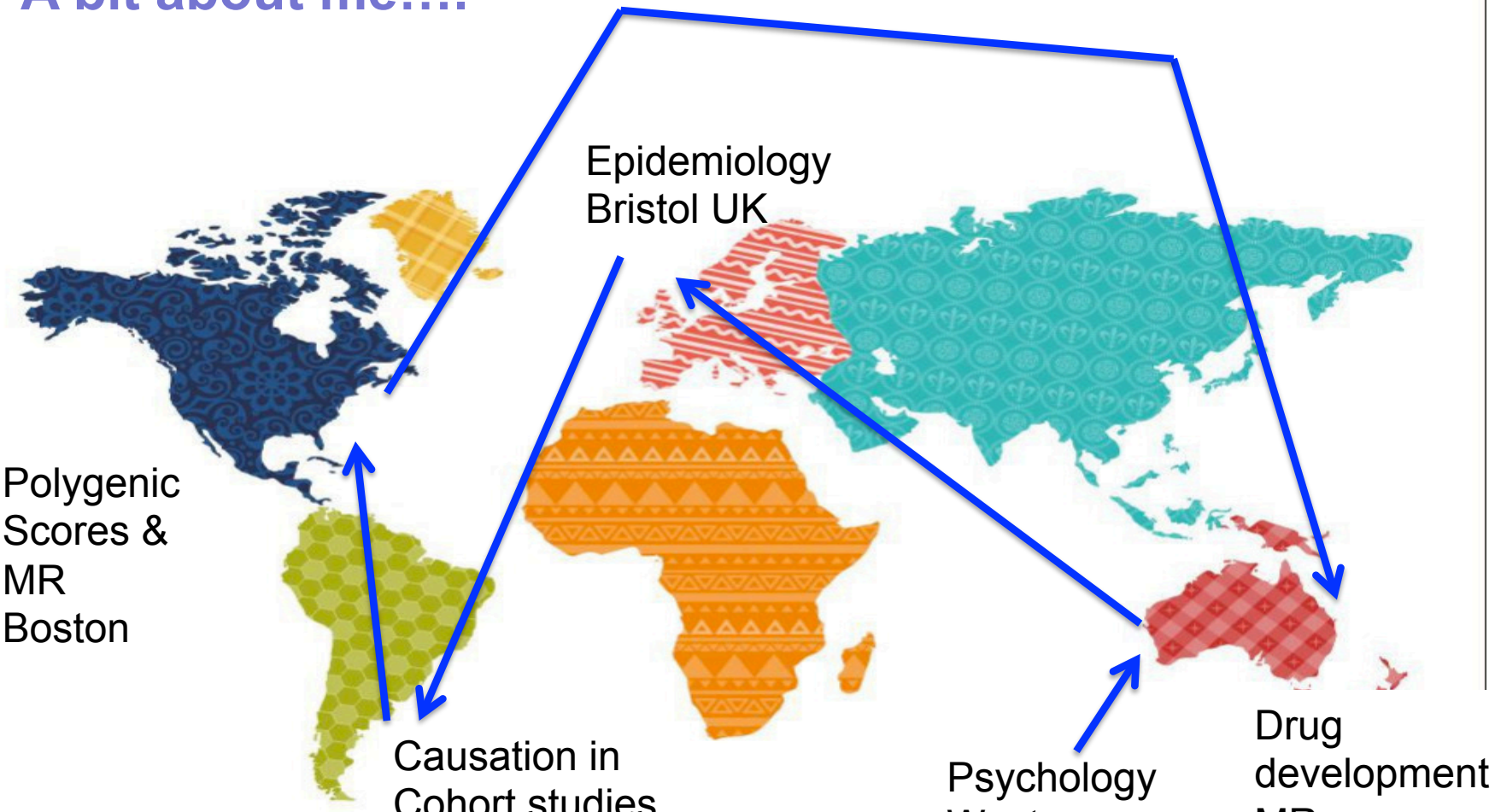
Mendelian Randomization: Using genes to test for causal traits

CNSG tutorial
June 2016

Marie-Jo Brion
Postdoctoral Research Fellow
Evans lab, UQ Diamantina Institute Genomic Medicine
CNSG, Queensland Brain Institute UQ

A bit about me….



MENDELIAN RANDOMIZATION

- What's all the fuss about Mendelian Randomization
- What is Mendelian Randomization (MR)
- Standard MR methods
- Recent Extensions to address key limitations
- Additional useful concepts to understand in MR (if there's time!)

WHATS ALL THE FUSS ABOUT MR?

A Mendelian Randomization Study of Circulating Uric Acid and Type 2 Diabetes

Ivonne Sluijs[1], Michael V. Holmes[2,3], Yvonne T. van der Schouw[1], Joline W.J. Beulens[1], Folkert W. Asselbergs[1,4,5], José María Huerta[6,7], Tom M. Palmer[8], Larraitz Arriola[7,9,10], Beverley Balkau[11,12],

Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study

Lancet 2012; 380: 572–80

Benjamin F Voight*, Gina M Peloso*, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Bar … ma Hólm, Eric L Ding, Toby Johnson, Heribert Schunkert, Nilesh J Samani, Robert Clarke, Jemma C Hopewell, John F Thompson, Mingyao Li,

PLOS | MEDICINE

Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data

OPEN ACCESS BMJ 2014;349:g4164 doi: 10.1136/bmj.g4164

Michael V Holmes *assistant professor (joint first author)*[1 2 3], Caroline E Dale *research fellow (joint first author)*[4], Luisa Zuccolo *population health scientist fellow*[5], Richard J Silverwood *lecturer in*

RESEARCH ARTICLE

Obesity and Multiple Sclerosis: A Mendelian Randomization Study

Lauren E. Mokry[1,2☯], Stephanie Ross[2☯], Nicholas J. Timpson[3], Stephen Sawcer[4], George Davey Smith[3], J. Brent Richards[1,2,5,6,7*]

PLOS | MEDICINE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study

Lauren E. Mokry, Stephanie Ross, Omar S. Ahmad, Vincenzo Forgetta, George Davey Smith, Aaron Leong, Celia M. T. Greenwood, George Thanassoulis, J. Brent Richards

OPEN ACCESS Freely available online

PLOS | MEDICINE

Serum Iron Levels and the Risk of Parkinson Disease: A Mendelian Randomization Study

Irene Pichler[1☯]*, Fabiola Del Greco M.[1☯], Martin Gögele[1], Christina M. Lill[2,3], Lars Bertram[2], Chuong B. Do[4], Nicholas Eriksson[4], Tatiana Foroud[5], Richard H. Myers[6], PD GWAS Consortium[¶],

Association of plasma uric acid with ischaemic heart disease and blood pressure: mendelian randomisation analysis of two large cohorts

OPEN ACCESS

BMJ 2013;347:f4262 doi: 10.1136/bmj.f4262

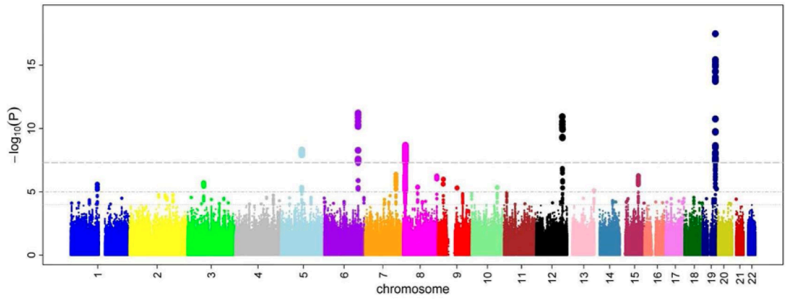
Tom M Palmer *assistant professor*[1]. Børge G Nordestgaard …

C-reactive protein and its role in metabolic syndrome: mendelian randomisation study

Nicholas J Timpson, Debbie A Lawlor, Roger M Harbord, Tom R Gaunt, Ian N M Day, Lyle J Palmer, Andrew T Hattersley, Shah Ebrahim, Gordon D O Lowe, Ann Rumley, George Davey Smith

ANALOGY: GENETIC STUDIES VS EPIDEMIOLOGY

- GWAS:
  - 500,000 SNP-trait associations
  - Small SNP effects, independent outside LD blocks
  - Identify only small numbers



- Epidemiology hypothetical “T-WAS”
  - 500,000 trait-trait associations
  - **A huge number** will come up as associated
    - human traits of health and disease are extremely highly intercorrelated

  - Big problem for epidemiological association is (not discovery of new hits)
    - How to distinguish which of the thousands are causal relationships we can intervene on and which are non-causal correlations

THE PROBLEM WITH EPIDEMIOLOGICAL ASSOCIATIONS

RESEARCH ARTICLE

Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology

George Davey Smith, Debbie A Lawlor, Roger Harbord, Nic Timpson, Ian Day, Shah Ebrahim

PLOS | MEDICINE

December 11, 2007 • http://dx.doi.org/10.1371/journal.pmed.0040352

We demonstrate that behavioural, socioeconomic, and physiological factors are strongly interrelated, with 45% of all possible pairwise associations between 96 nongenetic characteristics (n = 4,560 correlations) being significant at the $p < 0.01$ level

THE PROBLEM WITH EPIDEMIOLOGICAL ASSOCIATIONS

No reliable methods for fully controlling for confounding in standard observational studies

- Statistical covariate adjustment shown to be **completely inadequate**
- Action frequently taken in public health based on extremely poor evidence

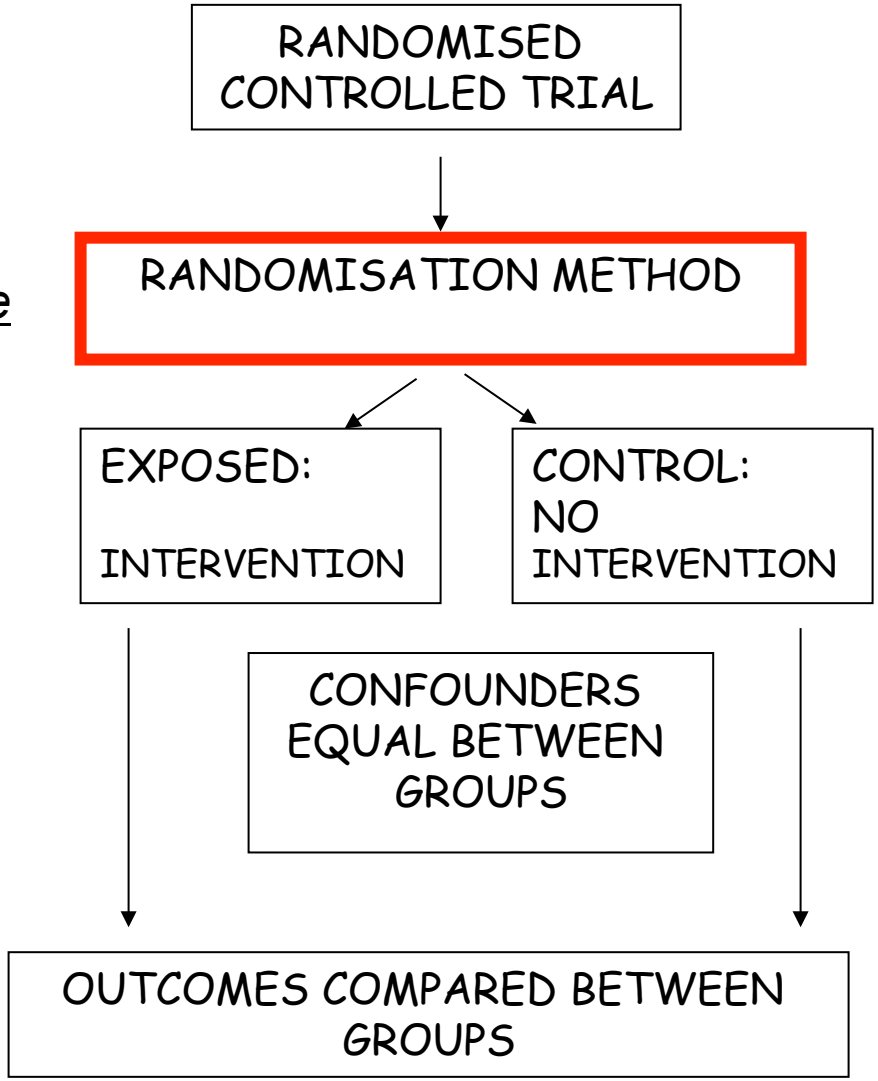
Serious & widespread effects

- Ineffective (harmful) medical and health interventions & policies
- Misleading public health information & advice
- Failed drug development research (95% failure rate)

HOW CAN WE DO A BETTER JOB AT IDENTIFYING CAUSAL EFFECTS?

RCTS: THE 'GOLD STANDARD' FOR CAUSALITY

Randomisation makes causal inference possible



WHY NOT JUST RELY ON RANDOMISED CLINICAL TRIALS?

Ethically:

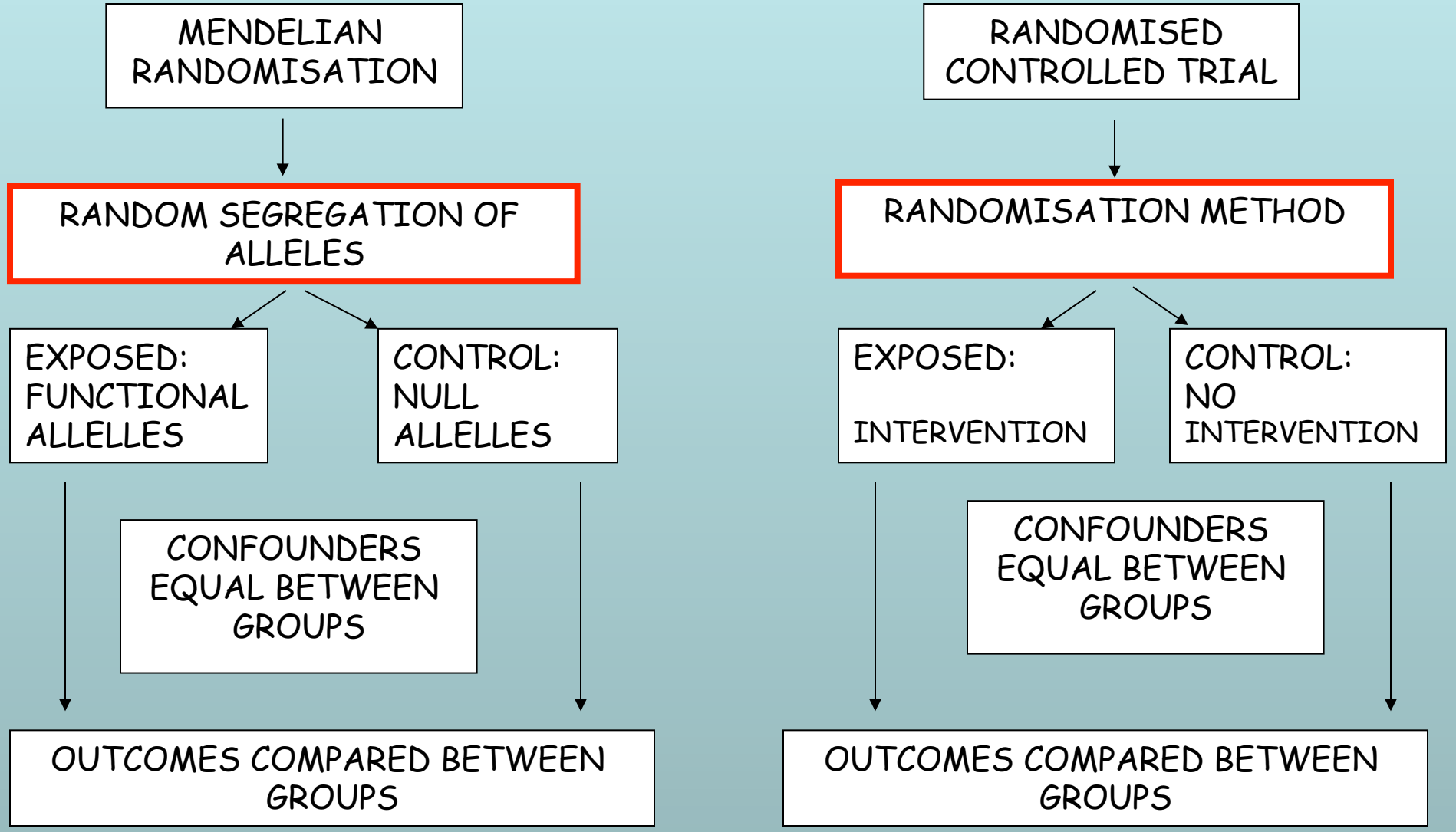
1. RCTs cannot be undertaken for many traits of interest

(anything adverse) Most human studies need to be observational

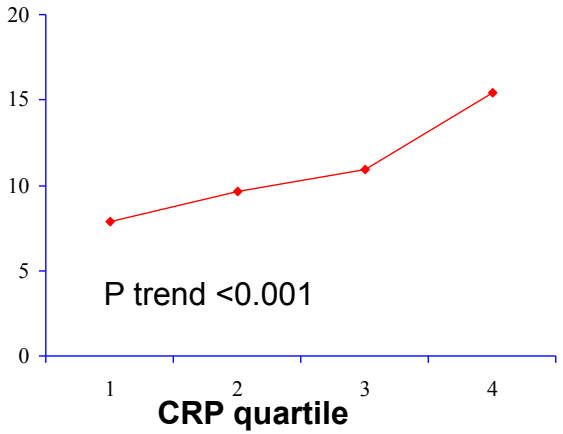
2. RCTs need to be undertaken AFTER there is already good evidence for causality in humans

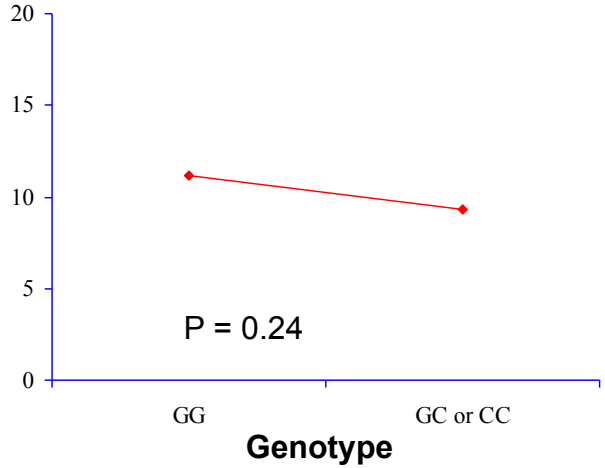
(before subjecting them to experiments & investing millions of dollars)

MENDELIAN RANDOMISATION AND RCTS

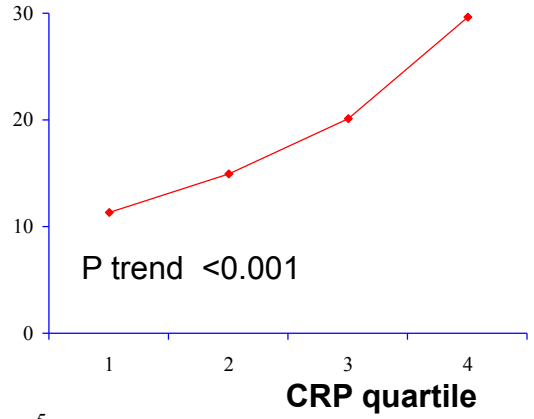


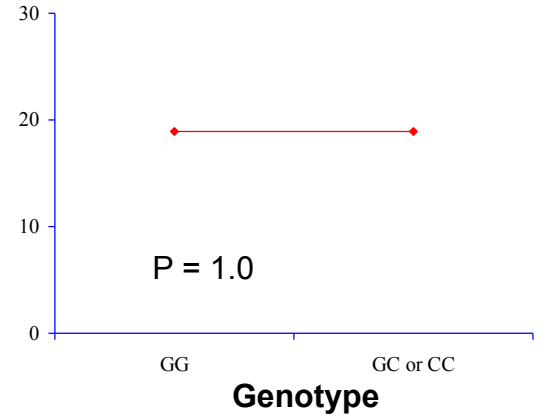
CRP & Smoking



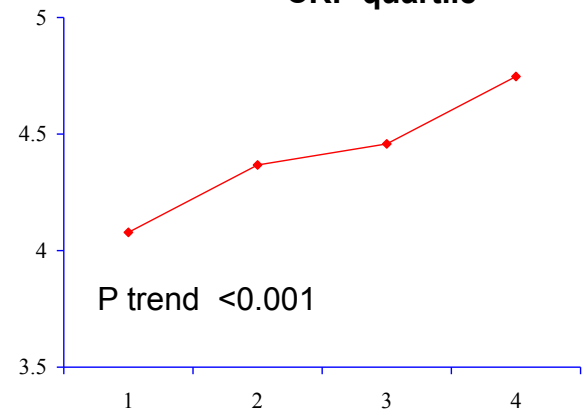


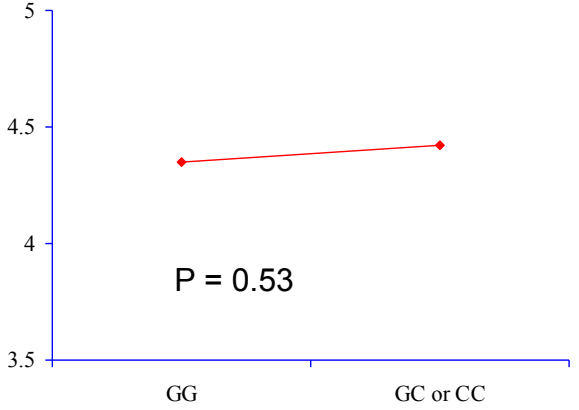
CRP & Physical inactivity





CRP & Socio-economic position





Davey Smith et al, ATVB 2005

WHAT DOES MENDELIAN RANDOMIZATION ACTUALLY DO?

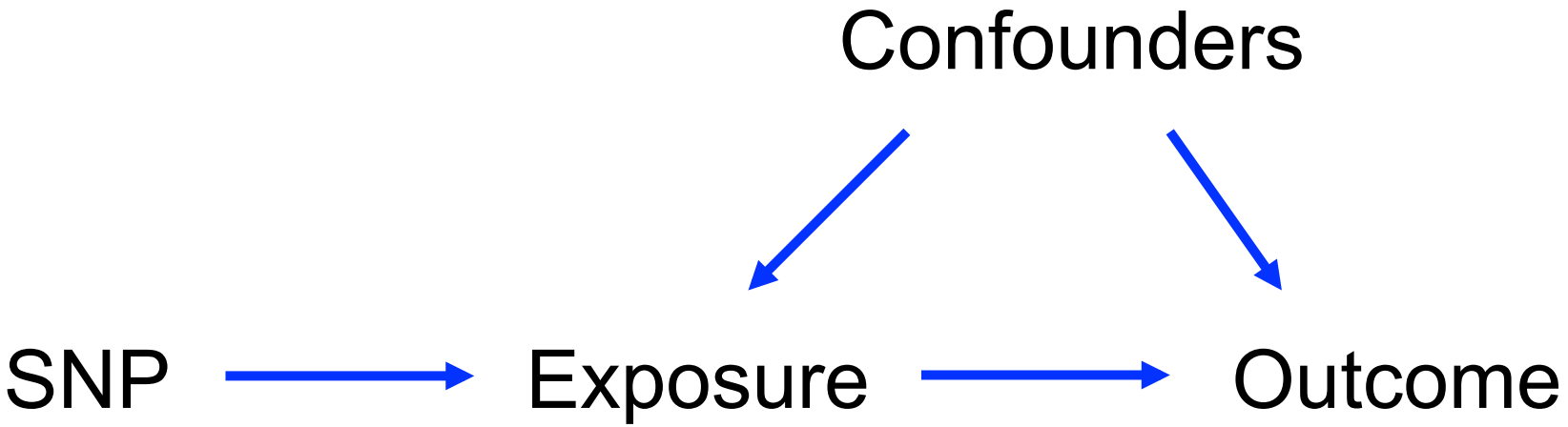
Based on concept that alleles segregate randomly with respect to environmental factors and genetic variants for different traits assort independently:

1. Tests for the presence of a causal relationship between two variables

2. Estimates magnitude of a causal effect

Provided 3 core assumptions are met……

3 CORE REQUIREMENTS FOR MENDELIAN RANDOMIZATION TO BE VALID



(1) SNP is reliably associated with the exposure

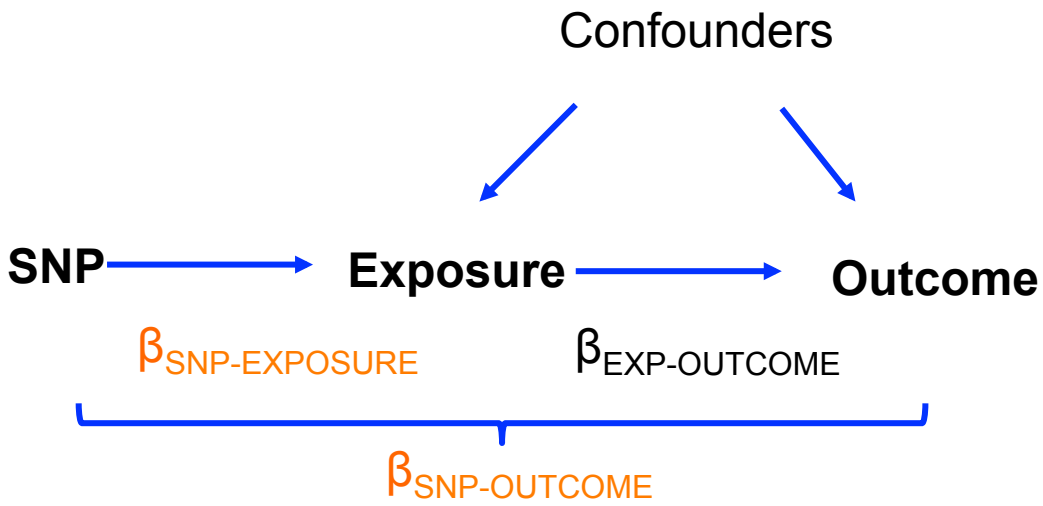
(2) SNP is not associated with confounding variables

(3) SNP only associated with outcome through the exposure *

MENDELIAN RANDOMIZATION

- What's all the fuss about Mendelian Randomization
- What is Mendelian Randomization (MR)
- Standard MR methods
- Recent Extensions to address key limitations
- Additional useful concepts to understand in MR (if there's time!)

STANDARD MR – USING INDIVIDUAL LEVEL DATA



TSLS:

1) Regress exposure on SNP & obtain predicted values
2) Regress outcome on **predicted** exposure (from 1st stage regression)

Wald Test* : $\dfrac{\beta_{SNP\text{-}OUTCOME}}{\beta_{SNP\text{-}EXPOSURE}}$

$\dfrac{\beta_{SNP\text{-}EXPOSURE} \times \beta_{EXP\text{-}OUTCOME}}{\beta_{SNP\text{-}EXPOSURE}}$

* Can also use summary data

EXAMPLE OF TSLS IN R

```
#R package needed for two stage least squares analysis
    library(AER)

#Ordinary least squares regression (contains CONFOUNDING)
    summary(lm(Y~X))

#Mendelian randomization analysis
    summary(ivreg(Y ~ X |  Z))

  #Single-SNP TSLS MR
    summary( ivreg(bmi ~ crp | rs12037, data=mrtest)

  #Multi-SNP TSLS MR
    summary( ivreg(bmi ~ hscrp | rs12037 + rs4206 + rs4129 + rs2794, data=mrtest)

  #Allelic-score TSLS MR
    # First generate (weighted or unweighted) allele scores in PLINK/R
    summary( ivreg(bmi ~ crp | CRPscore, data=mrtest)
```

TSLS IN R: EXAMPLE OUTPUT

Assessing the causal effect of CRP on BMI, using CRP allele score

ORDINARY LEAST SQUARES phenotypic association

```
Call:
lm(formula = mr$bmi ~ mr$crp)
Coefficients:
             Estimate      SE        Pr(>|t|)
crp           0.348        0.0137    <2e-16 ***
```

TSLS Mendelian randomization

```
Call:
lm(formula = mr$bmi ~ mr$crp | mr$allelescore)
Coefficients:
             Estimate      SE        Pr(>|t|)
crp          0.0512        0.0941    0.833
```

BOTH RETURN CHANGE IN : BMI (OUTCOME)

PREDICTED BY : UNIT CHANGE IN CRP (EXPOSURE)

BUT TSLS = **CAUSAL**

MENDELIAN RANDOMIZATION METHODS

- **Standard MR methods :**
    - Two-stage least squares (TSLS) on individual level data
        - Single SNP MR
        - Multi-SNP MR
        - Allelic score MR
- **Recent Extensions:**
    - Summary statistic & two sample MR
        - Inverse-variance weighted (IVW) MR – maximise power
        - Egger MR – address pleiotropy

MR FOR SUMMARY STATISTIC & TWO-SAMPLE DATA

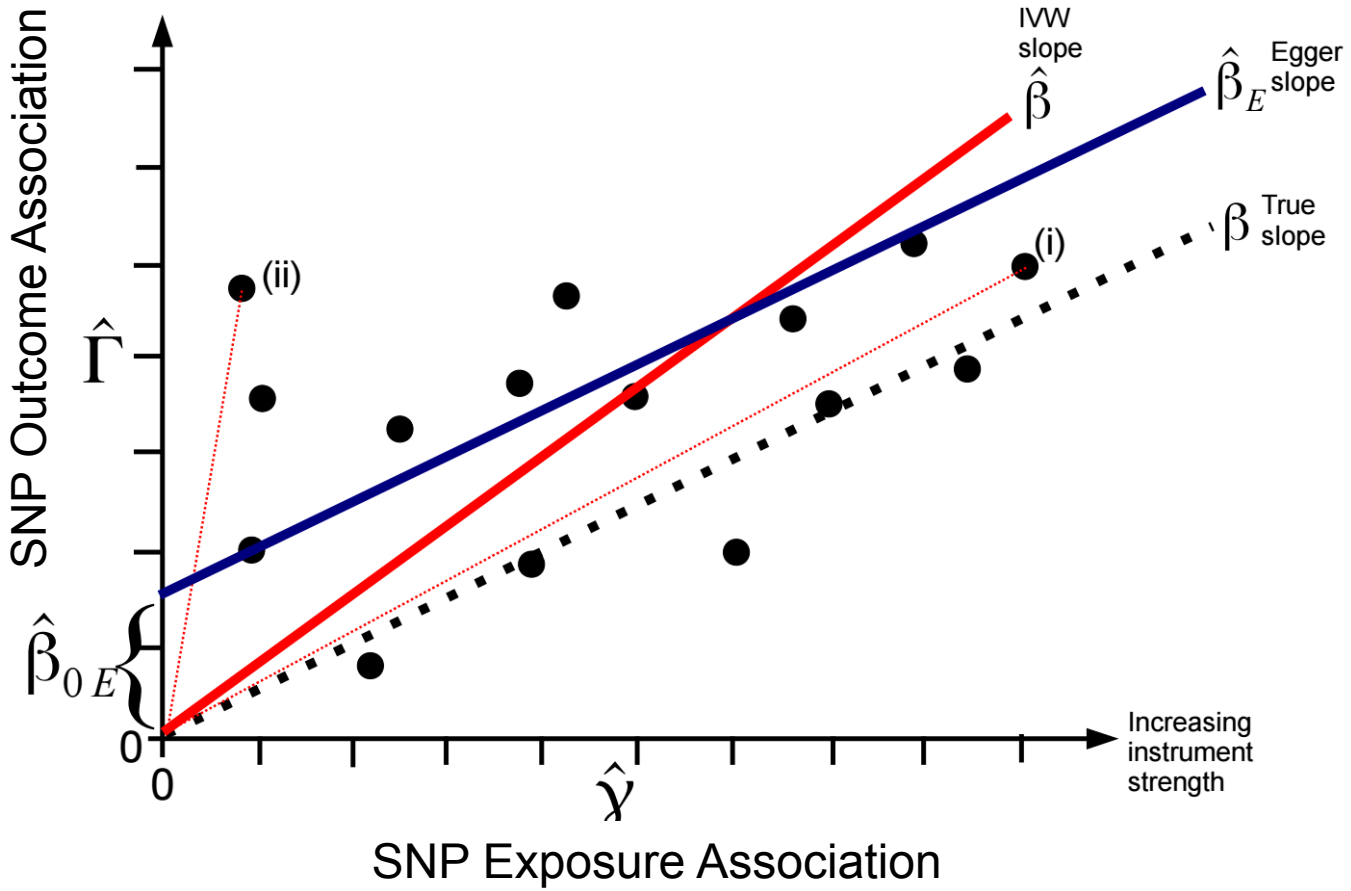
1. Inverse-variance weighted (IVW) MR

- Summary-level SNP estimates from multiple genetic variants
  - Can be from two different GWAS MAs
    (one for exposure one for outcome
- Fixed effects IVW meta-analysis across different SNPs
  - For their the causal IV estimate
  (ratio of SNP effect on outcome divided by SNP effect on exposure)
- Equivalent to doing an IVW regression analysis of SNP outcome on SNP exposure

2. MR Egger

- Similar to IVW but in the regression allows intercept to vary from zero

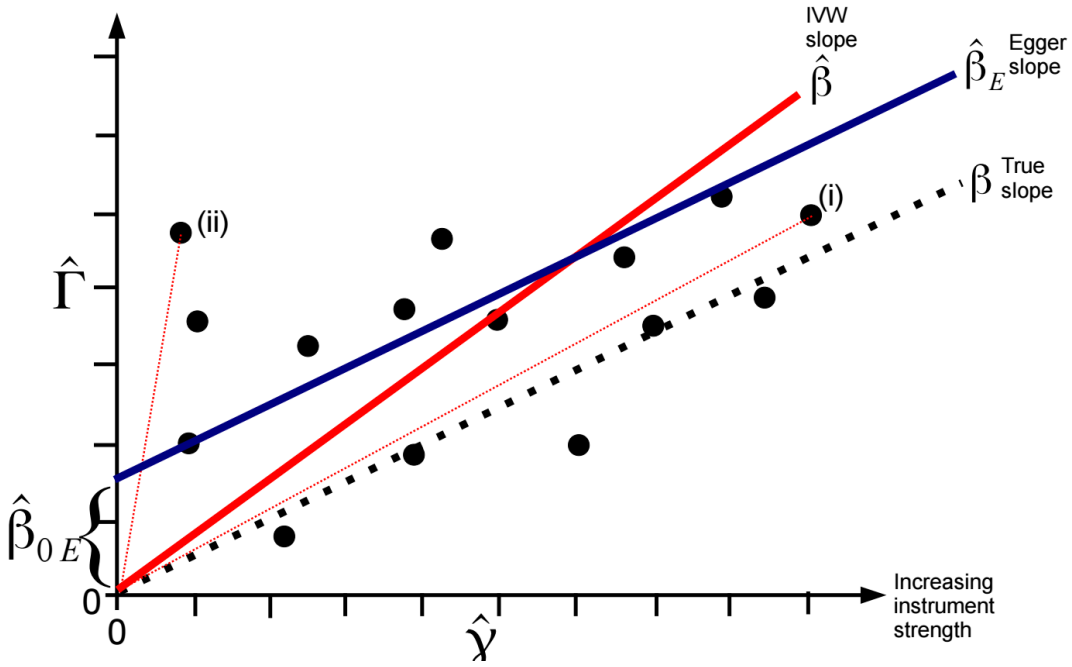
IVW MR AND EGGER



Regression beta = weighted average of SNP_outcome/SNP_exposure)

Causal estimate of change in outcome per unit change in exposure

IVW AND EGGER MR IN R



 # IVW MR

```
ivw.r  <- lm(b_out ~ - 1 + b_exp, weights = (1 / (se_out)^2)
```

 # MR Egger

```
egg.r  <- lm(b_out ~ b_exp, weights = (1 / (se_out)^2))
```

IVW AND EGGER R OUTPUT

 # IVW

```
lm(mr$b_schz ~ -1 + mr$b_crp, weights = 1 / (mr$se_schz)^2)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
b_crp  -0.1388     0.0438  -3.168  0.00562 **
```

 # EGGER

```
lm(mr$b_schz ~ mr$b_crp, weights = 1 / (mr$se_schz)^2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002090   0.004326   0.483   0.6355
b_crp       -0.131447   0.047305  -2.779   0.0134 *
```

MR FOR SUMMARY STATISTIC & TWO-SAMPLE DATA

2. MR Egger

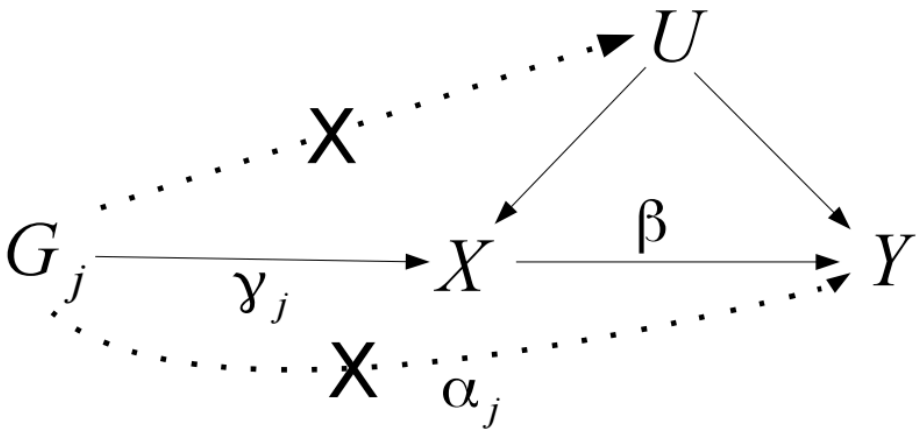
Advantages:

1. Two key elements to Egger:
   - Provides causal effect estimate that is less biased in the presence of pleiotropy
   - Tests statistically for the presence of pleiotropy

2. Egger enables an MR assumption to be relaxed

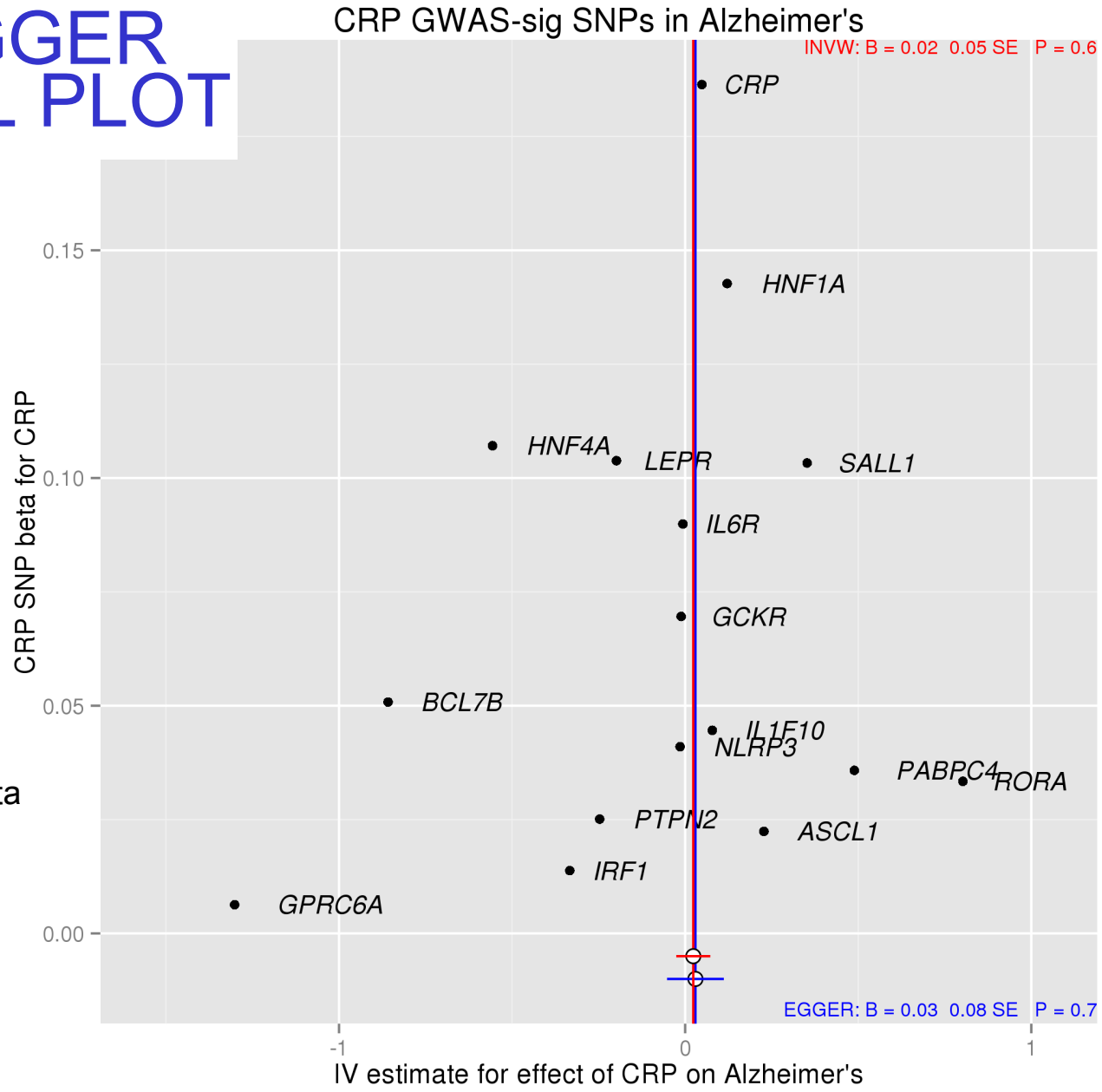
EXCLUSION RESTRICTION VS INSIDE ASSUMPTION

Standard MR assumption
'Exclusion Restriction'
(i.e. NO directional pleiotropy
No α_j)



Egger MR assumption
'INSIDE assumption'
(i.e. No correlation between α_j and γ_j
across instruments)

MR EGGER FUNNEL PLOT

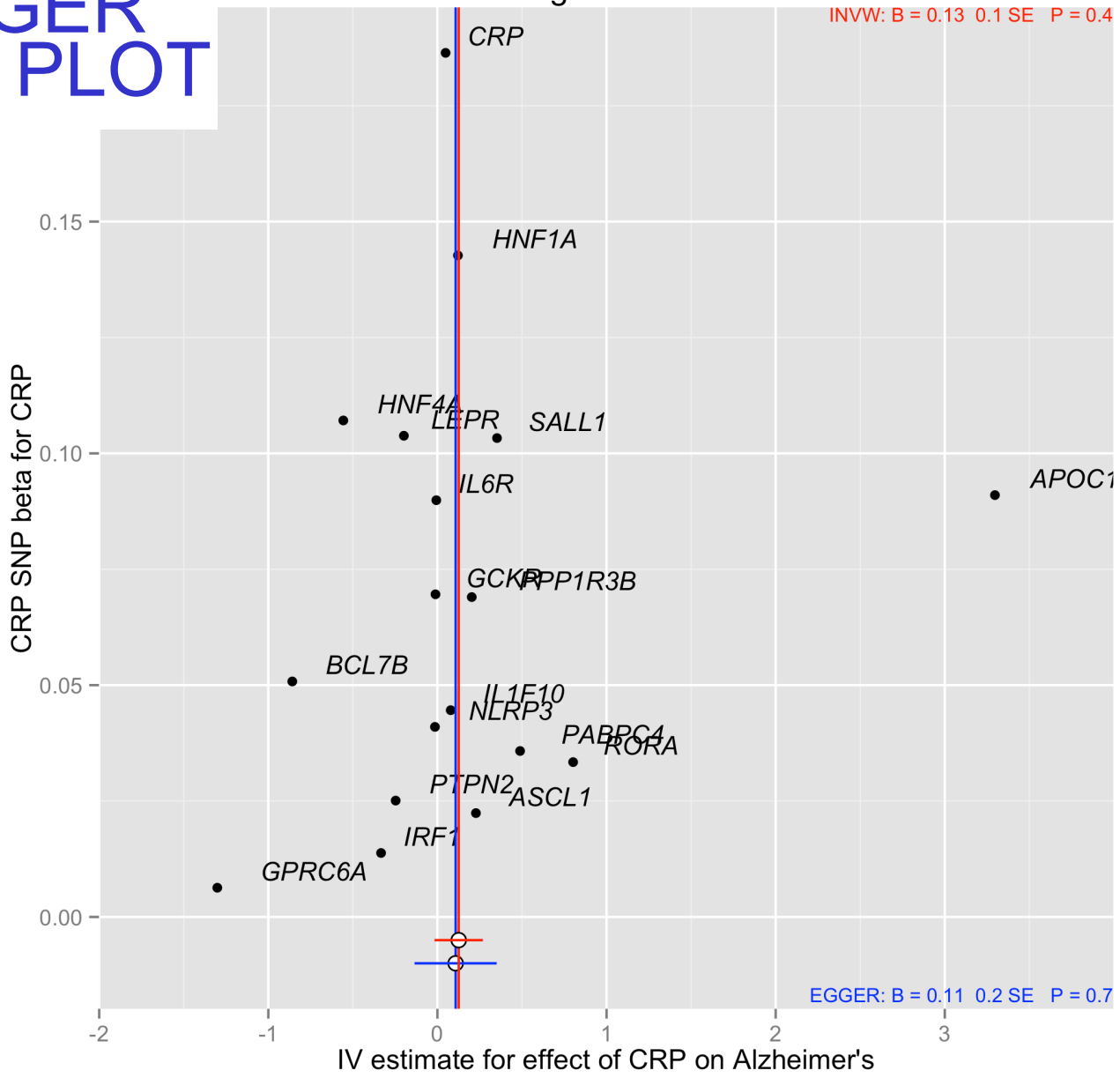


"MAF-corrected" exposure SNP beta

(1/se(Beta_IV)

ggplot(data, aes(y = b_exp_maf, x = b_iv))

MR EGGER FUNNEL PLOT



"MAF-corrected" exposure SNP beta

(1/se(Beta_IV)

```
ggplot(data, aes(y = b_exp_maf, x = b_iv))
```

SUMMARY STATISTIC IVW AND MR EGGER

Overall aims to maximise statistical power for MR by using summary-level SNP effects from very large GWAS studies

IVW MR - better statistical power
- more biased in the presence of pleiotropy
- equivalent results to individual-level multi-SNP TSLS MR

Egger MR - lower statistical power
- less biased in the presence of pleiotropy

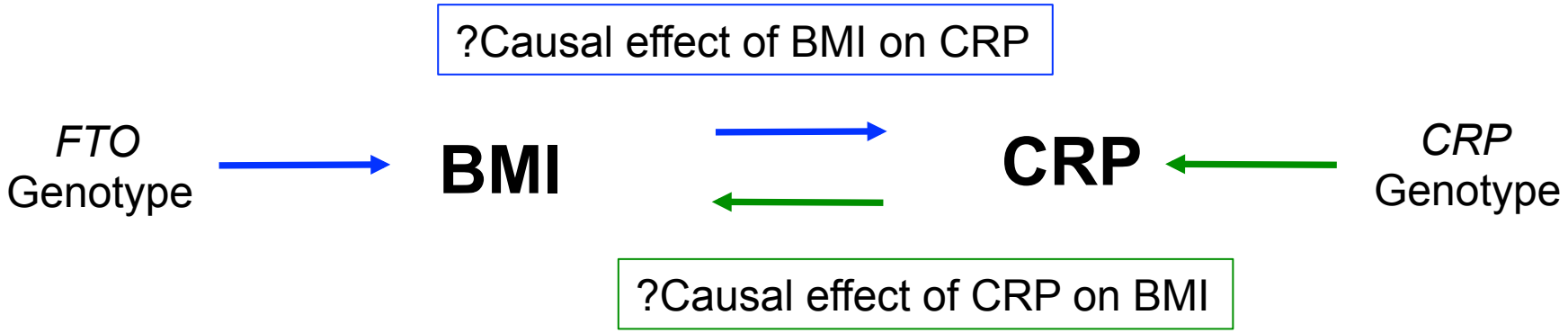
Best to implement BOTH IVW and Egger interpret the estimates together

MENDELIAN RANDOMIZATION

- What's all the fuss about MR
- What is Mendelian Randomization (MR)
- Standard MR methods
- Recent Extensions to address power and pleiotropy
- Additional useful concepts to understand in MR (if there's time!)

ADDITIONAL USEFUL CONCEPTS TO UNDERSTAND IN MR

"BI-DIRECTIONAL MENDELIAN RANDOMIZATION"



INSTRUMENT STRENGTH

- Weak genetic instruments biases causal estimates
    - Single sample MR: towards confounded observational estimate
    - Two-sample MR: towards the null

- Check by looking at F-statistic from the first stage regression in TSLS

$$\frac{R^2 / k}{(1-R^2)/ (n-k-1)}$$

    - F-stat >10
        - Bias <10%

    - Provided by ‘diagnostics’ in AER

Calculating Statistical Power for MR

Why is it important?

- *Very* large sample sizes are usually required to ensure adequate statistical power for MR studies
- Inadequately powered MR studies can lead to false negatives and incorrectly concluding a non-causal effect

What determines statistical power for MR?

Three main parameters:

i) amount of variance in the exposure trait explained by the genetic instrument

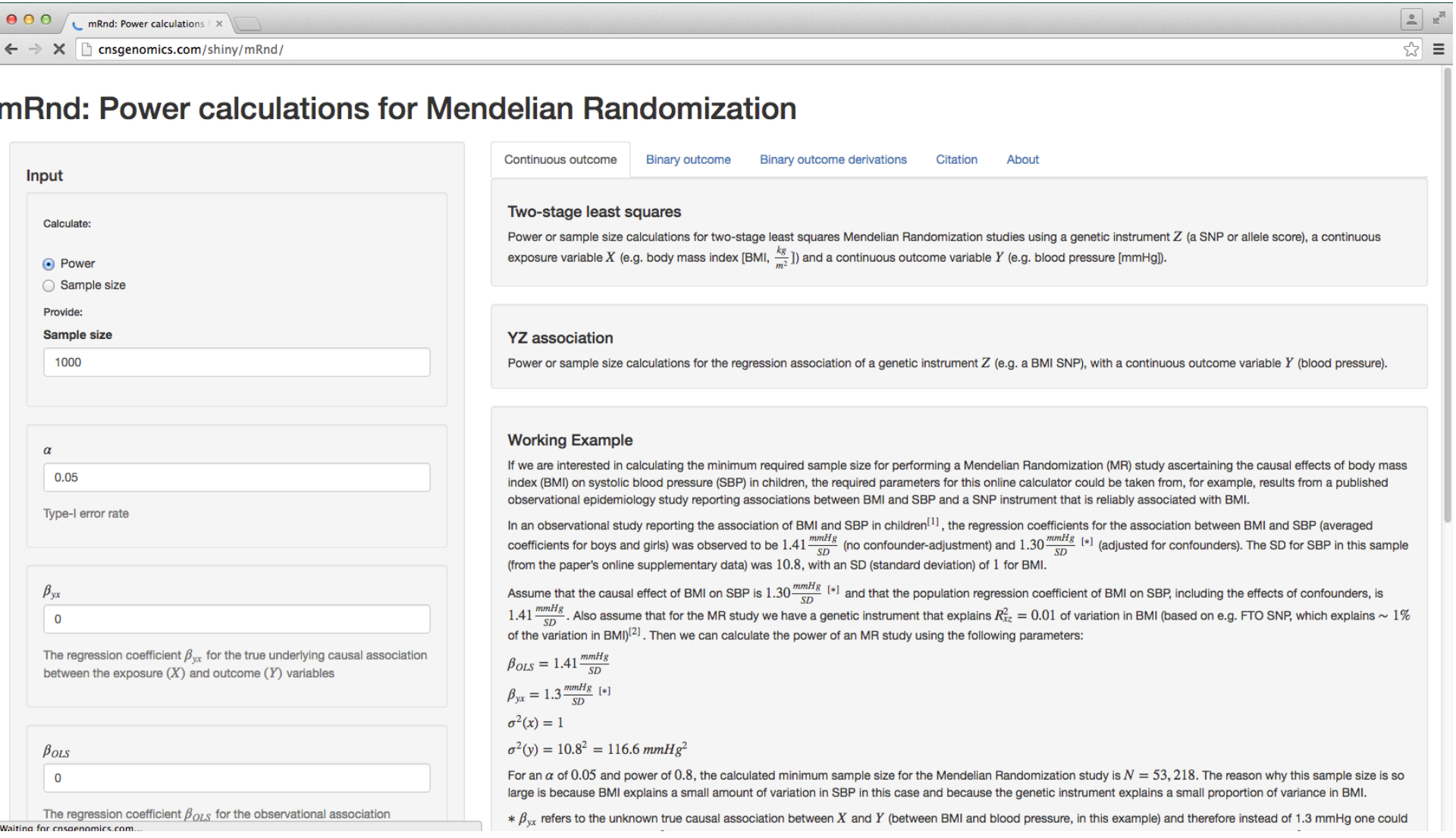
ii) study sample size,

iii) magnitude of the causal effect of the exposure on the outcome

Online Power Calculator for MR

Webpage: cnsgenomics.com/shiny/mRnd/

For details see: *Brion MJ, Shakbahzov K & Visscher P. Int J Epid (2013)*



Parameters Required to Perform Calculation

1 - **Desired level of power** *(eg 80%)* **OR available sample size** *(N)*

2 – **Alpha level** *eg 0.05*

3 – **Magnitude of causal XY association**

ie a hypothetical value estimated from literature

4 – **Magnitude of observational XY association**

ie from literature, implicitly contains confounding

5 – **Variance of X** *ie from the reported observational association*

6 – **Variance of Y** *ie from the reported observational association*

Sample Size Requirements for MR:

"Real World" Example of BMI and BP in children using *FTO*

Working Example

If we are interested in calculating the minimum required sample size for performing a Mendelian Randomization (MR) study ascertaining the causal effects of body mass index (BMI) on systolic blood pressure (SBP) in children, the required parameters for this online calculator could be taken from, for example, results from a published observational epidemiology study reporting associations between BMI and SBP and a SNP instrument that is reliably associated with BMI.

In an observational study reporting the association of BMI and SBP in children[1], the regression coefficients for the association between BMI and SBP (averaged coefficients for boys and girls) was observed to be $\mathbf{1.41}\frac{mmHg}{SD}$ (no confounder-adjustment) and $\mathbf{1.30}\frac{mmHg}{SD}$ [*] (adjusted for confounders). The SD for SBP in this sample (from the paper's online supplementary data) was $\mathbf{10.8}$, with an SD (standard deviation) of $\mathbf{1}$ for BMI.

Assume that the causal effect of BMI on SBP is $\mathbf{1.30}\frac{mmHg}{SD}$ [*] and that the population regression coefficient of BMI on SBP, including the effects of confounders, is $\mathbf{1.41}\frac{mmHg}{SD}$. Also assume that for the MR study we have a genetic instrument that explains $R^2_{xz} = \mathbf{0.01}$ of variation in BMI (based on e.g. FTO SNP, which explains $\sim \mathbf{1\%}$ of the variation in BMI)[2]. Then we can calculate the power of an MR study using the following parameters:

$\beta_{OLS} = \mathbf{1.41}\frac{mmHg}{SD}$

$\beta_{yx} = \mathbf{1.3}\frac{mmHg}{SD}$ [*]

$\sigma^2(x) = \mathbf{1}$

$\sigma^2(y) = \mathbf{10.8}^2 = \mathbf{116.6}\ mmHg^2$

For an α of $\mathbf{0.05}$ and power of $\mathbf{0.8}$, the calculated minimum sample size for the Mendelian Randomization study is $N = \mathbf{53,218}$. The reason why this sample size is so large is because BMI explains a small amount of variation in SBP in this case and because the genetic instrument explains a small proportion of variance in BMI.

Required sample size =53,218

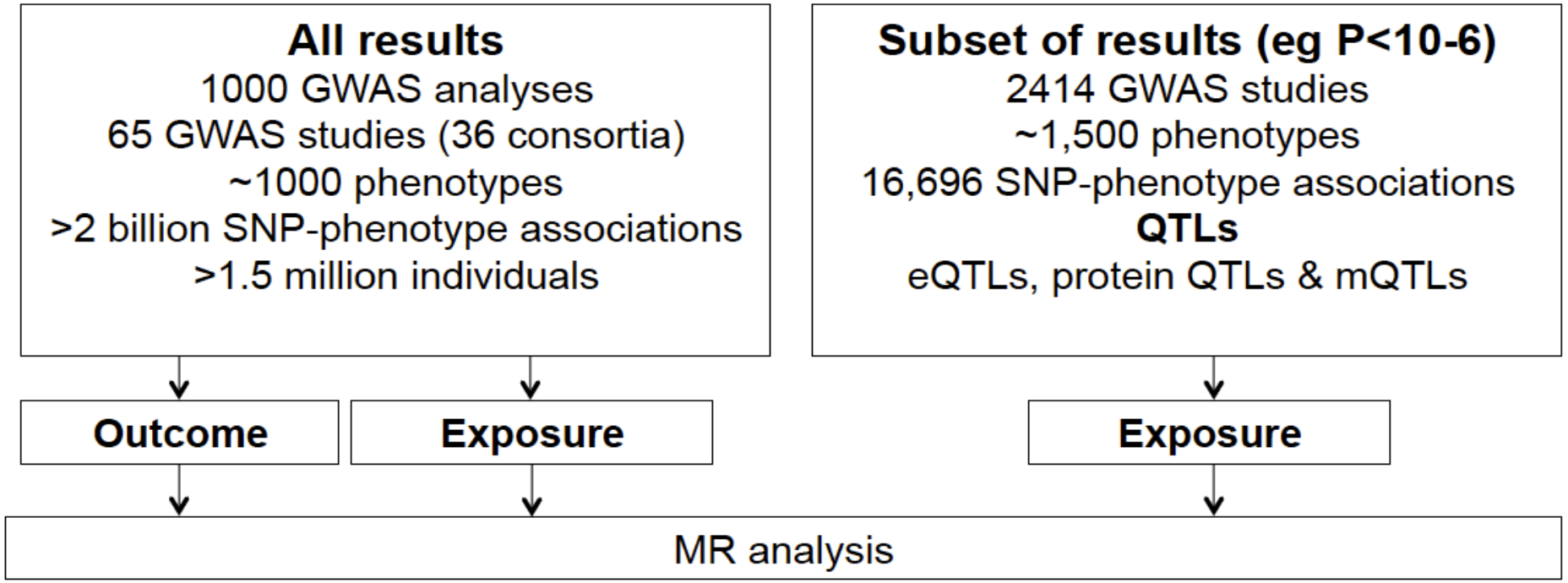
* β_{yx} refers to the unknown true causal association between X and Y (between BMI and blood pressure, in this example) and therefore instead of 1.3 mmHg one could potentially use any value of β_{yx} deemed plausible or, for example, inspect the power/sample size calculations for a range of hypothetical values of β_{yx}.

1. Lawlor DA, Benfield L, Logue J et al. Association between general and central adiposity in childhood, and change in these, with cardiovascular risk factors in adolescence: prospective cohort study. BMJ 2010; 341: c6224.

2. Frayling TM, Timpson NJ, Weedon MN et al. A Common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 2007; 316(5826): 889-894.

MRBASE

A platform for Mendelian randomisation using summary data from genome-wide association studies



University of BRISTOL

www.mrbase.org/alpha

MRC Integrative Epidemiology Unit

References

- **Davey-Smith & Ebrahim (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? IJE, 32, 1-22.**
- Palmer et al (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors *Stat Methods Med Res* 21(3): 223-242
- Bowden et al (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*, 44, 512-25.
- Davey-Smith & Hemani (2014). Mendelian randomization: genetic studies for causal inference in epidemiological studies. *Hum Mol Genet*, 23(1), R89-98.
- Evans & Davey-Smith (2015). Mendelian randomization: New applications in the coming age of hypothesis free causality. Annu Rev Genomics Hum Genet, 16, 327-50.
- Brion et al (2013). Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol*, 42(5), 1497-501.
- Lawlor et al (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology *Stat Meth* 27(8): 1133-63
- Didelez et al (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Meth Res* 16(4): 309-30