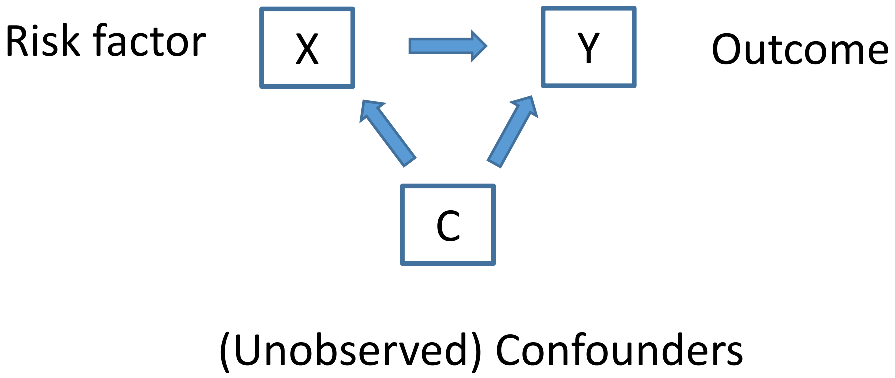
Mendelian Randomization

Drawback with observational studies



Risk factor

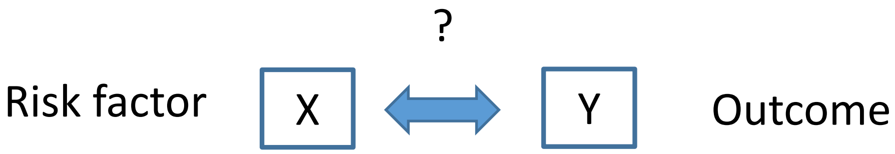
X

Y

Outcome

C

(Unobserved) Confounders



?

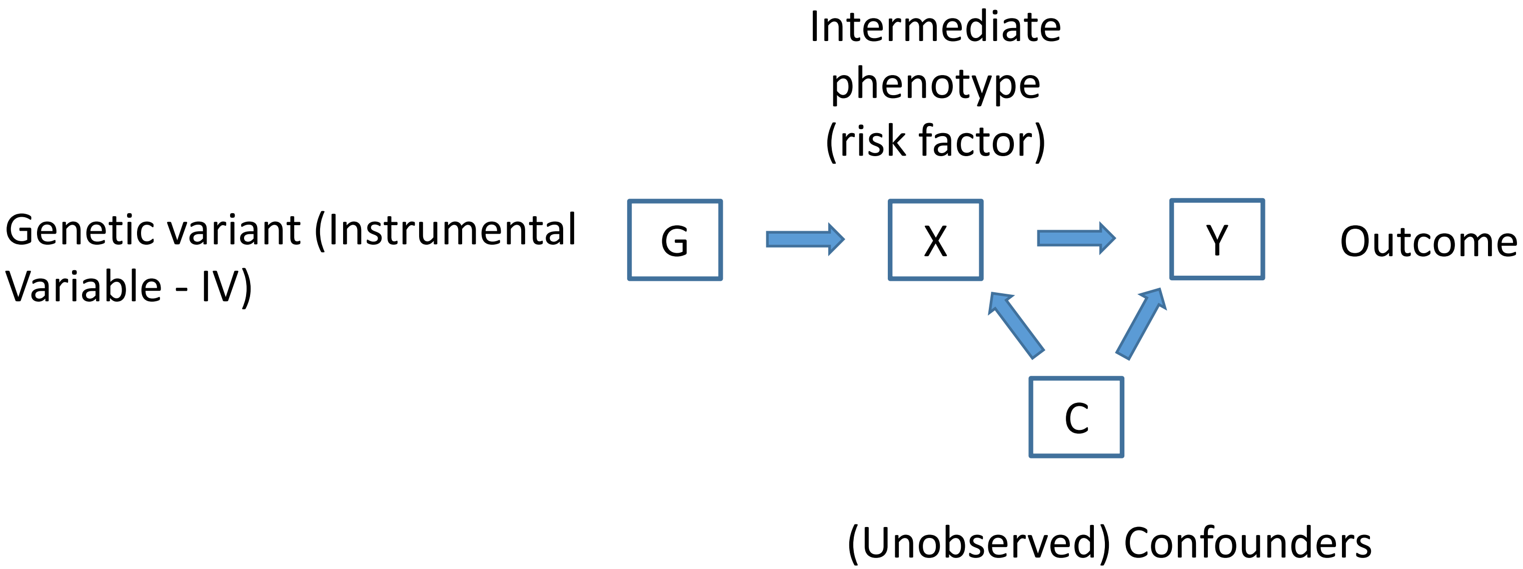
Risk factor

X

Y

Outcome

We can leverage genetic variation to (partly) overcome these issues



Intermediate phenotype (risk factor)

Genetic variant (Instrumental Variable - IV)

Outcome

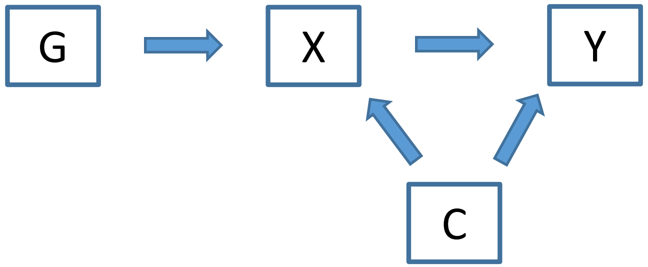
(Unobserved) Confounders

Mendelian Randomization

- *Basic principle: "genetic variants which mirror the biological effects of a modifiable environmental exposure and alters disease risk should be associated with disease risk to the extent predicted by their influence on exposure to the risk factor."*
- The random allocation of genetic variants from parents to offspring means these variants will generally be unrelated to other factors which affect the outcome.
- Furthermore, associations between the genotype and the outcome will not be affected by reverse causation because disease does not affect genotype

Ebrahim & Davey Smith, Hum Genet 2008
Davey Smith & Ebrahim, Int J Epi 2004

Three key assumptions in MR analysis



1. G (SNP or a combination of multiple SNPs) is robustly associated with X (risk factor)

2. G is unrelated to any confounders C, that can bias the relationship between G and Y (outcome). In other words, there are no common causes of G and Y (e.g. population stratification)

3. G is related to Y only through its association with X (i.e. no pleiotropy)

Assumption 1: G is robustly associated with X

- Under certain conditions, the relative bias of the instrument variable (IV) estimate is ~1/F. A “weak” IV has been defined as having F<10, where

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

R^2 is variance in X explained by the IV(s), n is sample size and k is number of IVs

- Weak IVs can lead to biased effect estimates (in the direction of the observed X-Y association) in the presence of confounding of the X–Y relationship.

Pierce, IJE 2011

Assumption 2: No confounding

- G is independent of factors (measured and unmeasured) that confound the X-Y relation

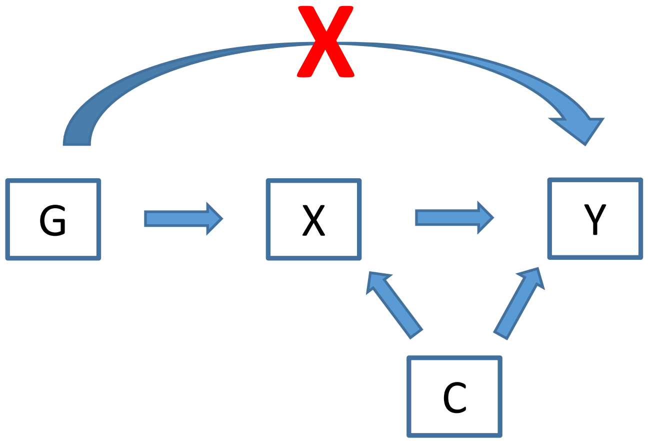
- Since G is randomized at birth and thus is independent of non-genetic confounders and is not modified by the course of disease, the one main concern here is population stratification – i.e. if ancestry is related both to G and Y.

- If you have individual-level data, you can test for this (e.g. PCA)

Assumption 3: No pleiotropy

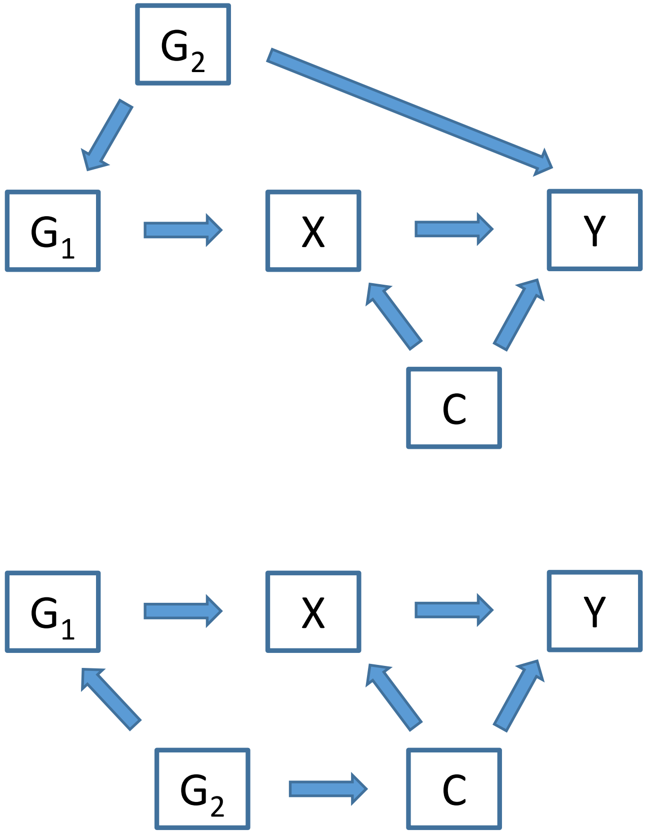
- This assumption is the trickiest

- Assumes that G is only associated with Y via X and thus the association between G and Y is fully mediated by X and not through any unmeasured factor(s). Needs to be true for SNPs in LD too

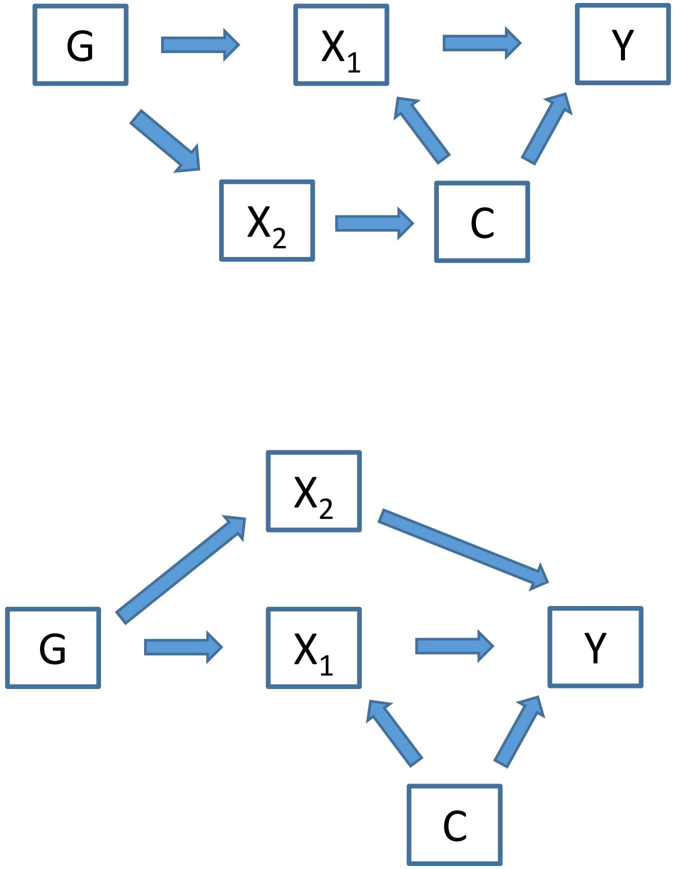


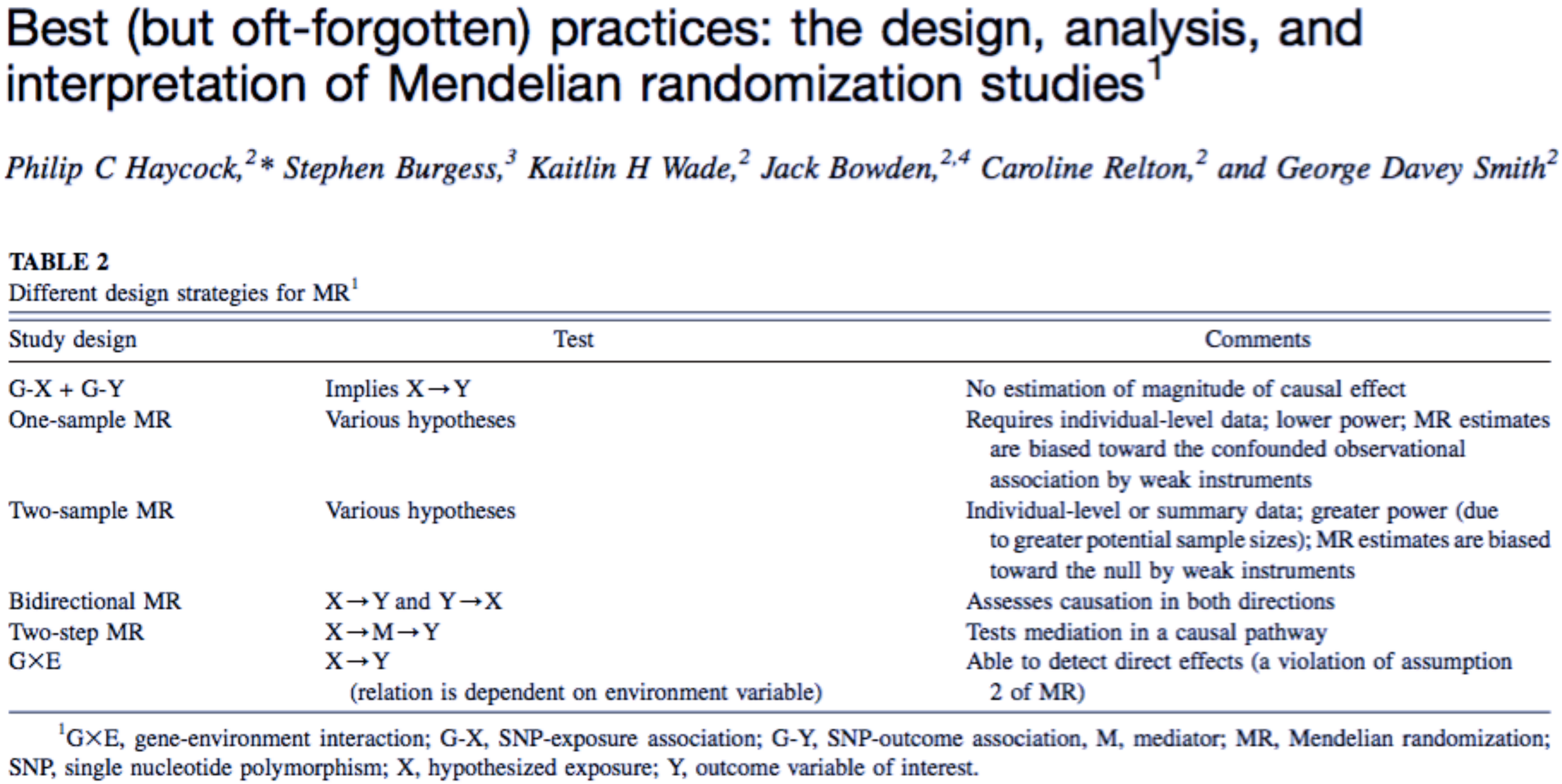
Scenarios invalidating assumption 3

LD



Pleiotropy





Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies[1]

Philip C Haycock,[2] Stephen Burgess,[3] Kaitlin H Wade,[2] Jack Bowden,[2,4] Caroline Relton,[2] and George Davey Smith[2]*

TABLE 2
Different design strategies for MR[1]

Study design	Test	Comments
G-X + G-Y	Implies X→Y	No estimation of magnitude of causal effect
One-sample MR	Various hypotheses	Requires individual-level data; lower power; MR estimates are biased toward the confounded observational association by weak instruments
Two-sample MR	Various hypotheses	Individual-level or summary data; greater power (due to greater potential sample sizes); MR estimates are biased toward the null by weak instruments
Bidirectional MR	X→Y and Y→X	Assesses causation in both directions
Two-step MR	X→M→Y	Tests mediation in a causal pathway
G×E	X→Y (relation is dependent on environment variable)	Able to detect direct effects (a violation of assumption 2 of MR)

[1] G×E, gene-environment interaction; G-X, SNP-exposure association; G-Y, SNP-outcome association, M, mediator; MR, Mendelian randomization; SNP, single nucleotide polymorphism; X, hypothesized exposure; Y, outcome variable of interest.

Haycock et al, Am J Clin Nutr 2016

Individual-level data in one sample

- Access to SNPs, risk factor, and outcome for all participants

- The causal effect of X on Y can be estimated using 2-stage least-squares (2SLS) regression:

1. X = $\mathrm{a} + \gamma G$
2. Y = $\mathrm{c} + \beta X^{*}$, where X^{*} are the genetically predicted exposure levels as measured in (1)

- The causal estimate is given by β
- Can be implemented in R using the “ivpack” package
- Weak instruments cause bias towards the observed confounded association

Summary data from two samples

- The G-X and the G-Y associations are estimated in two different samples.

- Assumes no overlap among samples and that the two populations are similar (ethnicity, age, sex, etc.)

- Here, bias due to weak IVs will be towards the null

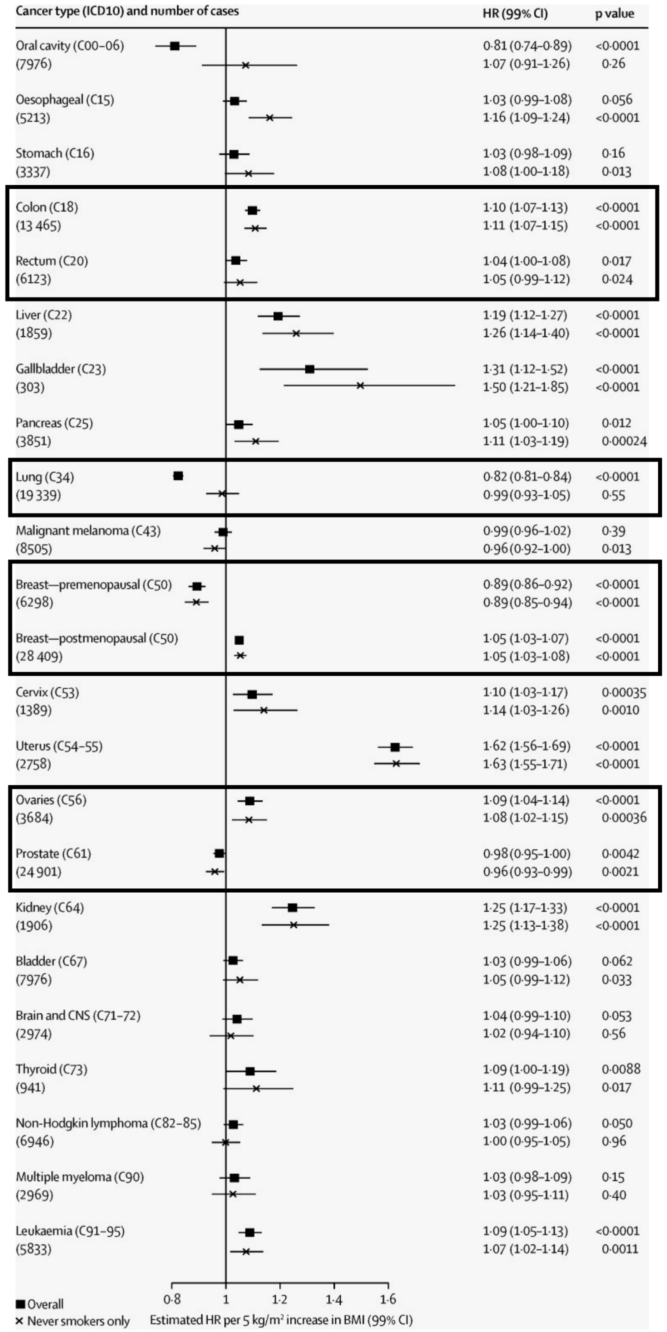
- Note: The G-X and G-Y associations need to be coded using the same effect allele

Summary data from two samples

$$\hat{\beta} = \frac{\sum_k \beta_{1k}\beta_{2k}\sigma_{\beta_{2k}}^{-2}}{\sum_k \beta_{1k}\sigma_{\beta_{2k}}^{-2}}$$

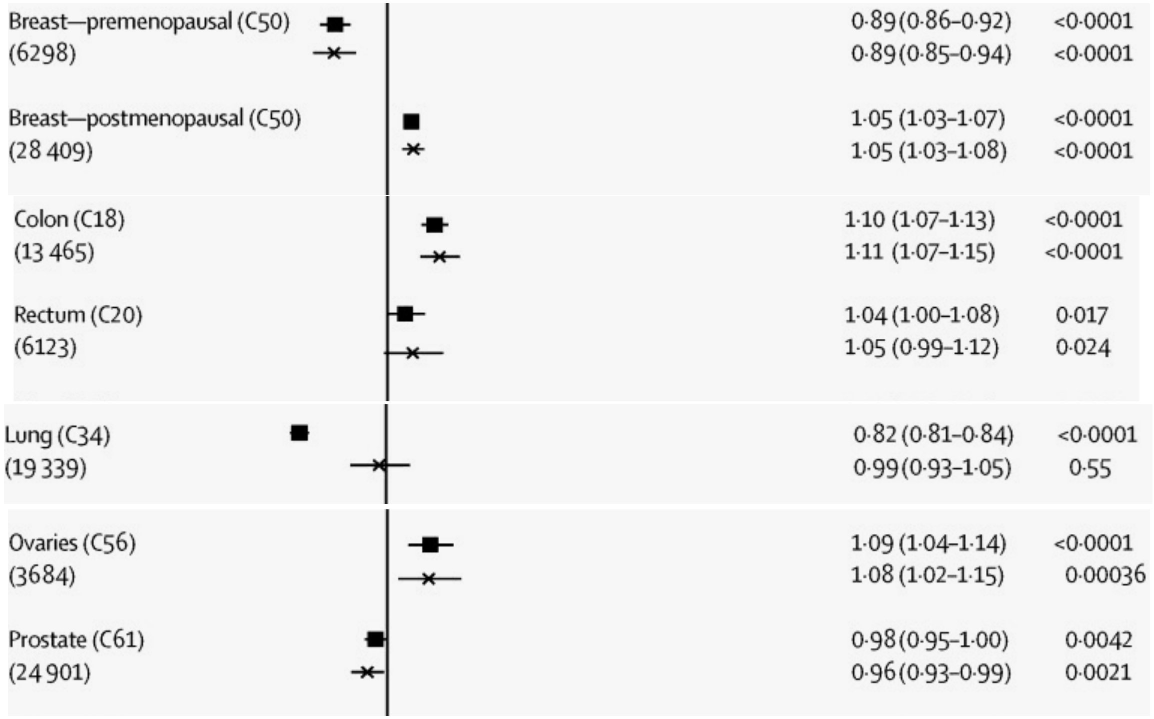
$$se(\hat{\beta}) = \sqrt{\frac{1}{\sum_k \beta_{1k}^{2}\sigma_{\beta_{2k}}^{-2}}}$$

β_{1k} is the mean change in X per allele for SNP k, β_{2k} is the mean change in Y per allele for SNP k, σ_{2k}^{-2} is the inverse variance for the G-Y association.



Cancer type (ICD10) and number of cases	HR (99% CI)	p value
Oral cavity (C00–06)	0·81 (0·74–0·89)	<0·0001
(7976)	1·07 (0·91–1·26)	0·26
Oesophageal (C15)	1·03 (0·99–1·08)	0·056
(5213)	1·16 (1·09–1·24)	<0·0001
Stomach (C16)	1·03 (0·98–1·09)	0·16
(3337)	1·08 (1·00–1·18)	0·013
Colon (C18)	1·10 (1·07–1·13)	<0·0001
(13 465)	1·11 (1·07–1·15)	<0·0001
Rectum (C20)	1·04 (1·00–1·08)	0·017
(6123)	1·05 (0·99–1·12)	0·024
Liver (C22)	1·19 (1·12–1·27)	<0·0001
(1859)	1·26 (1·14–1·40)	<0·0001
Gallbladder (C23)	1·31 (1·12–1·52)	<0·0001
(303)	1·50 (1·21–1·85)	<0·0001
Pancreas (C25)	1·05 (1·00–1·10)	0·012
(3851)	1·11 (1·03–1·19)	0·00024
Lung (C34)	0·82 (0·81–0·84)	<0·0001
(19 339)	0·99 (0·93–1·05)	0·55
Malignant melanoma (C43)	0·99 (0·96–1·02)	0·39
(8505)	0·96 (0·92–1·00)	0·013
Breast—premenopausal (C50)	0·89 (0·86–0·92)	<0·0001
(6298)	0·89 (0·85–0·94)	<0·0001
Breast—postmenopausal (C50)	1·05 (1·03–1·07)	<0·0001
(28 409)	1·05 (1·03–1·08)	<0·0001
Cervix (C53)	1·10 (1·03–1·17)	0·00035
(1389)	1·14 (1·03–1·26)	0·0010
Uterus (C54–55)	1·62 (1·56–1·69)	<0·0001
(2758)	1·63 (1·55–1·71)	<0·0001
Ovaries (C56)	1·09 (1·04–1·14)	<0·0001
(3684)	1·08 (1·02–1·15)	0·00036
Prostate (C61)	0·98 (0·95–1·00)	0·0042
(24 901)	0·96 (0·93–0·99)	0·0021
Kidney (C64)	1·25 (1·17–1·33)	<0·0001
(1906)	1·25 (1·13–1·38)	<0·0001
Bladder (C67)	1·03 (0·99–1·06)	0·062
(7976)	1·05 (0·99–1·12)	0·033
Brain and CNS (C71–72)	1·04 (0·99–1·10)	0·053
(2974)	1·02 (0·94–1·10)	0·56
Thyroid (C73)	1·09 (1·00–1·19)	0·0088
(941)	1·11 (0·99–1·25)	0·017
Non-Hodgkin lymphoma (C82–85)	1·03 (0·99–1·06)	0·050
(6946)	1·00 (0·95–1·05)	0·96
Multiple myeloma (C90)	1·03 (0·98–1·09)	0·15
(2969)	1·03 (0·95–1·11)	0·40
Leukaemia (C91–95)	1·09 (1·05–1·13)	<0·0001
(5833)	1·07 (1·02–1·14)	0·0011

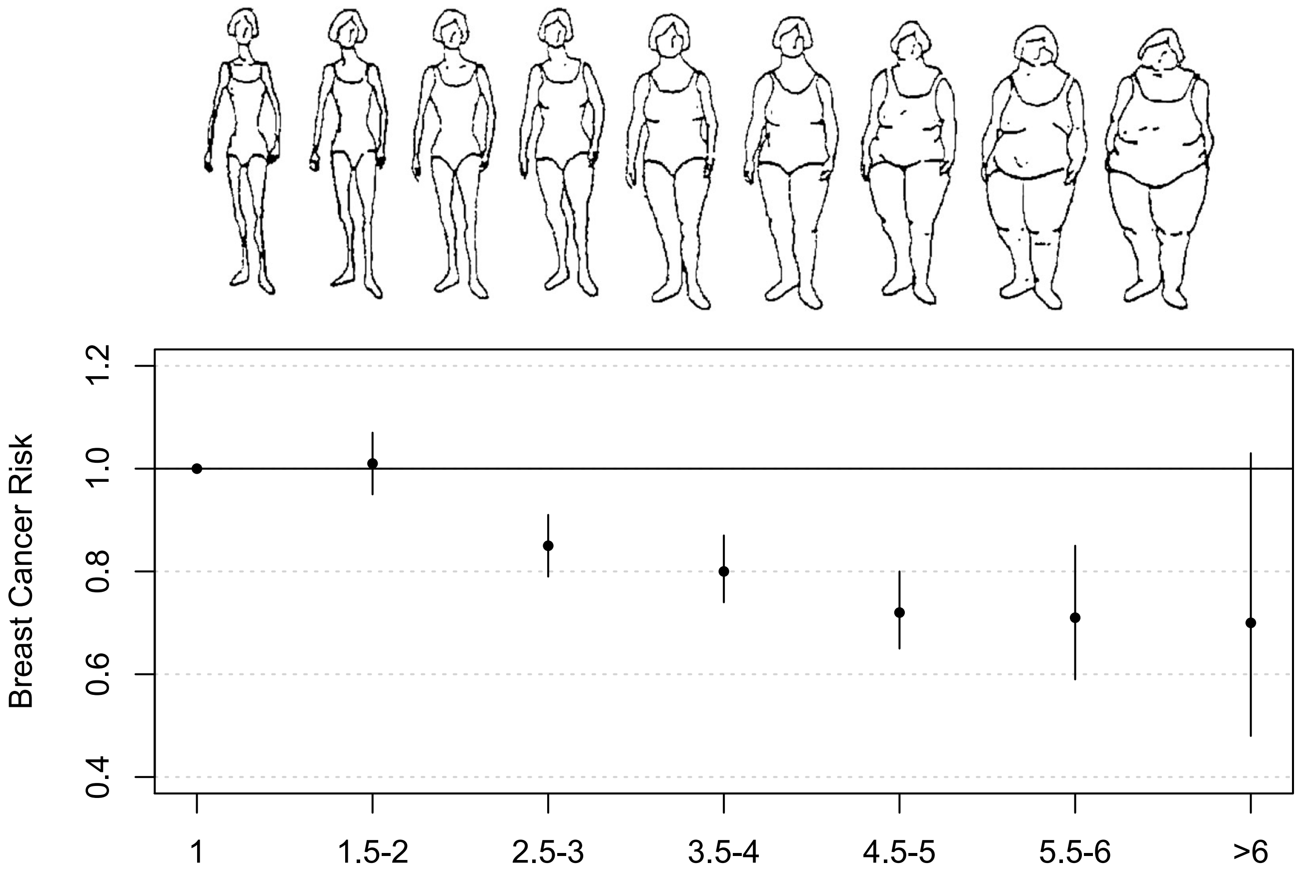
- Association between BMI and cancer risk was assessed for 22 cancers
- 5.24 million individuals (166,996 cancer cases)



Breast—premenopausal (C50)	0·89 (0·86–0·92)	<0·0001
(6298)	0·89 (0·85–0·94)	<0·0001
Breast—postmenopausal (C50)	1·05 (1·03–1·07)	<0·0001
(28 409)	1·05 (1·03–1·08)	<0·0001
Colon (C18)	1·10 (1·07–1·13)	<0·0001
(13 465)	1·11 (1·07–1·15)	<0·0001
Rectum (C20)	1·04 (1·00–1·08)	0·017
(6123)	1·05 (0·99–1·12)	0·024
Lung (C34)	0·82 (0·81–0·84)	<0·0001
(19 339)	0·99 (0·93–1·05)	0·55
Ovaries (C56)	1·09 (1·04–1·14)	<0·0001
(3684)	1·08 (1·02–1·15)	0·00036
Prostate (C61)	0·98 (0·95–1·00)	0·0042
(24 901)	0·96 (0·93–0·99)	0·0021

Bhaskaran et al, Lancet 2014

Childhood body fatness is inversely associated with breast cancer risk



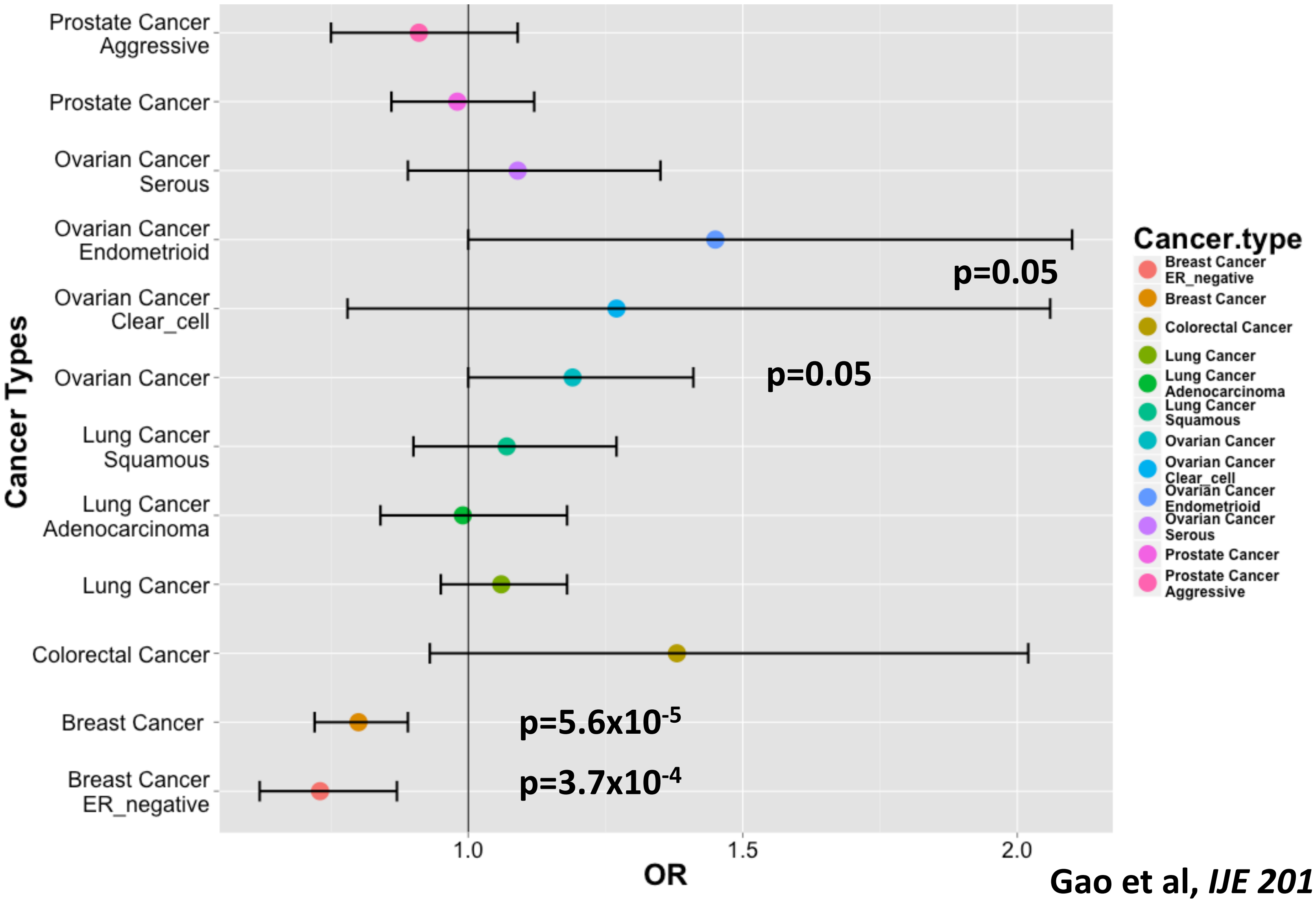
Baer et al, AJE 2010

Expansion to other cancer types within GAME-ON

Cancer Type	**Cases**	**Controls**	**GWAS studies**
Breast			
All	15,569	18,204	11
ER-negative	4,760	13,248	8
Colorectal	5,100	4,831	6
Lung[a]			
All	12,527	17,285	6
Adenocarcinoma	3,804	16,289	6
Squamous	3,546	16,434	6
Ovarian[a]			
All	4,369	9,123	3
Clear-cell	356	9,123	3
Endometrioid	715	9,123	3
Serous	2,556	9,123	3
Prostate			
All	14,160	12,712	6
Aggressive	4,446	12,724	6
Total	**51,725**	**62,155**	

Gao et al, *IJE 2016*

Childhood body fatness (9 SNPs)



p=0.05

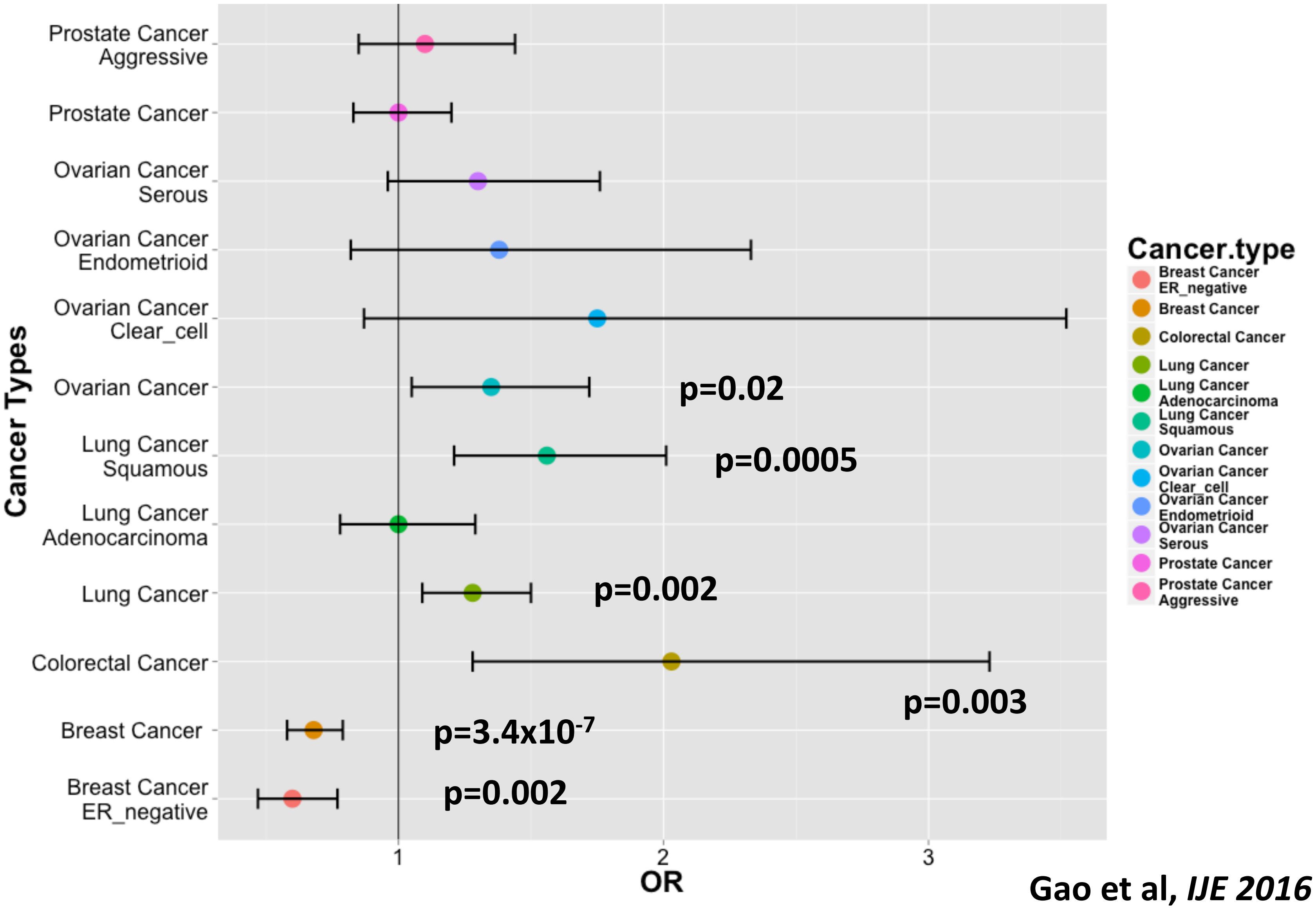
p=0.05

$p=5.6x10^{-5}$

$p=3.7x10^{-4}$

Gao et al, *IJE 2016*

Adult BMI (77 SNPs)



p=0.02

p=0.0005

p=0.002

p=0.003

p=3.4x10^{-7}

p=0.002

Gao et al, *IJE 2016*

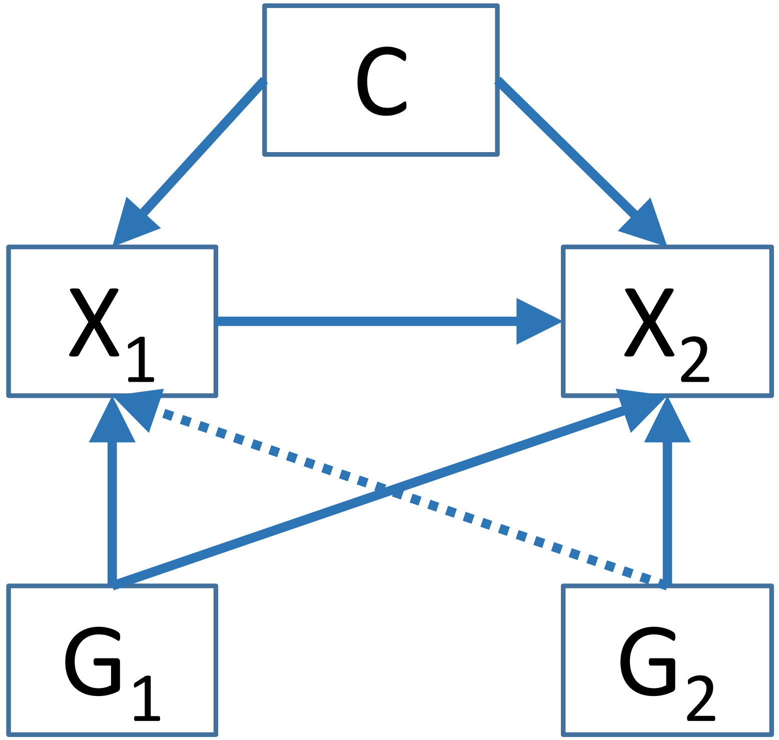
Bidirectional MR analysis

- Approach to overcome reverse causation

- IVs for both X_1 and X_2 are used to assess the causal association in both directions

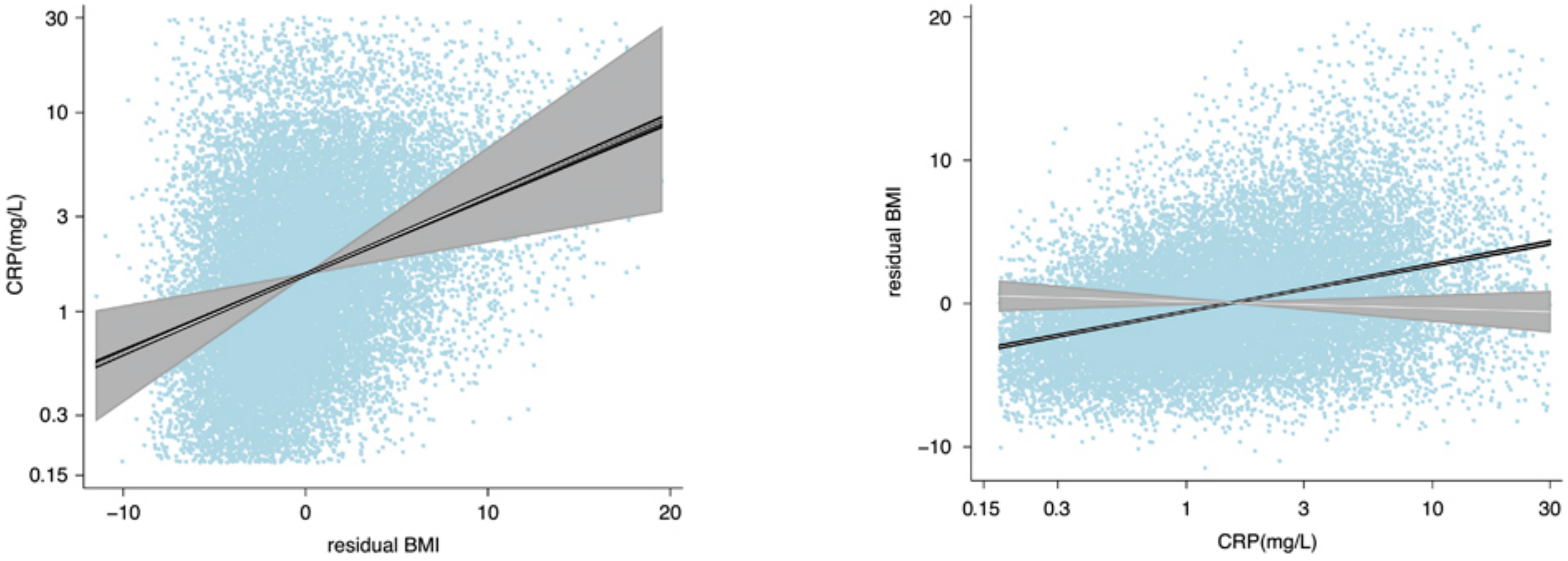
1. Is G_1 associated with X_2?
2. Is G_2 associated with X_1?

(Also confirm that G_1 is associated with X_1 and that G_2 is associated with X_2



BMI and CRP – what causes what?

- There is a consistent observed association between high BMI and high CRP levels



Light grey points represent a scatter plot of the correlation between circulating CRP and residual BMI. Gray areas represent 95% confidence regions around IV estimates. Black area represents 95% confidence regions around simple linear regression estimates.

Timpson et al, Int J Obesity 2011

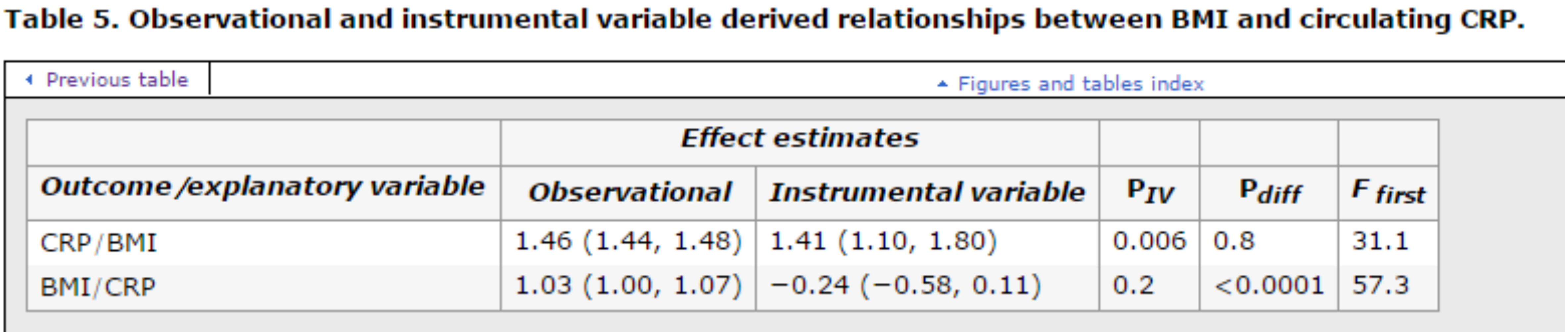


Table 5. Observational and instrumental variable derived relationships between BMI and circulating CRP.

	Effect estimates				
Outcome/explanatory variable	***Observational***	***Instrumental variable***	P_{IV}	P_{diff}	F_{first}
CRP/BMI	1.46 (1.44, 1.48)	1.41 (1.10, 1.80)	0.006	0.8	31.1
BMI/CRP	1.03 (1.00, 1.07)	−0.24 (−0.58, 0.11)	0.2	<0.0001	57.3

These data suggest that the observed association between circulating CRP and measured BMI is likely to be driven by BMI, with CRP being a marker of elevated adiposity.

Timpson et al, Int J Obesity 2011

Drawbacks with MR analysis

- Large sample sizes are needed
  - As genetic effects on risk factors are typically small, MR estimates of association have much wider confidence intervals than conventional epidemiological estimates.

- Make sure that the three key assumptions hold
  - In practice, this is very difficult, especially for the third assumption of no pleiotropy.

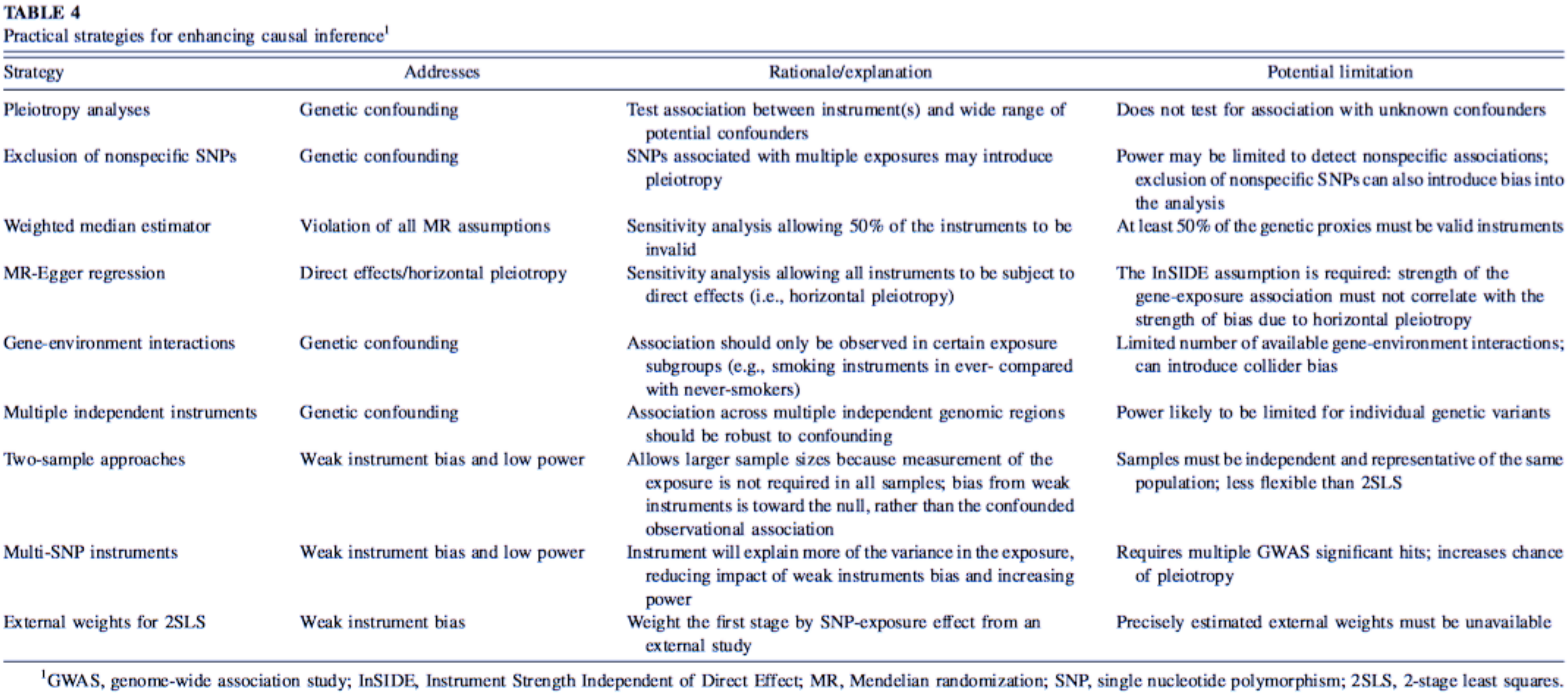


TABLE 4
Practical strategies for enhancing causal inference[1]

Strategy	Addresses	Rationale/explanation	Potential limitation
Pleiotropy analyses	Genetic confounding	Test association between instrument(s) and wide range of potential confounders	Does not test for association with unknown confounders
Exclusion of nonspecific SNPs	Genetic confounding	SNPs associated with multiple exposures may introduce pleiotropy	Power may be limited to detect nonspecific associations; exclusion of nonspecific SNPs can also introduce bias into the analysis
Weighted median estimator	Violation of all MR assumptions	Sensitivity analysis allowing 50% of the instruments to be invalid	At least 50% of the genetic proxies must be valid instruments
MR-Egger regression	Direct effects/horizontal pleiotropy	Sensitivity analysis allowing all instruments to be subject to direct effects (i.e., horizontal pleiotropy)	The InSIDE assumption is required: strength of the gene-exposure association must not correlate with the strength of bias due to horizontal pleiotropy
Gene-environment interactions	Genetic confounding	Association should only be observed in certain exposure subgroups (e.g., smoking instruments in ever- compared with never-smokers)	Limited number of available gene-environment interactions; can introduce collider bias
Multiple independent instruments	Genetic confounding	Association across multiple independent genomic regions should be robust to confounding	Power likely to be limited for individual genetic variants
Two-sample approaches	Weak instrument bias and low power	Allows larger sample sizes because measurement of the exposure is not required in all samples; bias from weak instruments is toward the null, rather than the confounded observational association	Samples must be independent and representative of the same population; less flexible than 2SLS
Multi-SNP instruments	Weak instrument bias and low power	Instrument will explain more of the variance in the exposure, reducing impact of weak instruments bias and increasing power	Requires multiple GWAS significant hits; increases chance of pleiotropy
External weights for 2SLS	Weak instrument bias	Weight the first stage by SNP-exposure effect from an external study	Precisely estimated external weights must be unavailable

[1] GWAS, genome-wide association study; InSIDE, Instrument Strength Independent of Direct Effect; MR, Mendelian randomization; SNP, single nucleotide polymorphism; 2SLS, 2-stage least squares.

Haycock et al, Am J Clin Nutr 2016