Molecular & Genetic Epidemiology

Kostas Tsilidis, PhD

University of Ioannina & Imperial College London

ktsilidis@gmail.com

Studies in Genetic Epidemiology

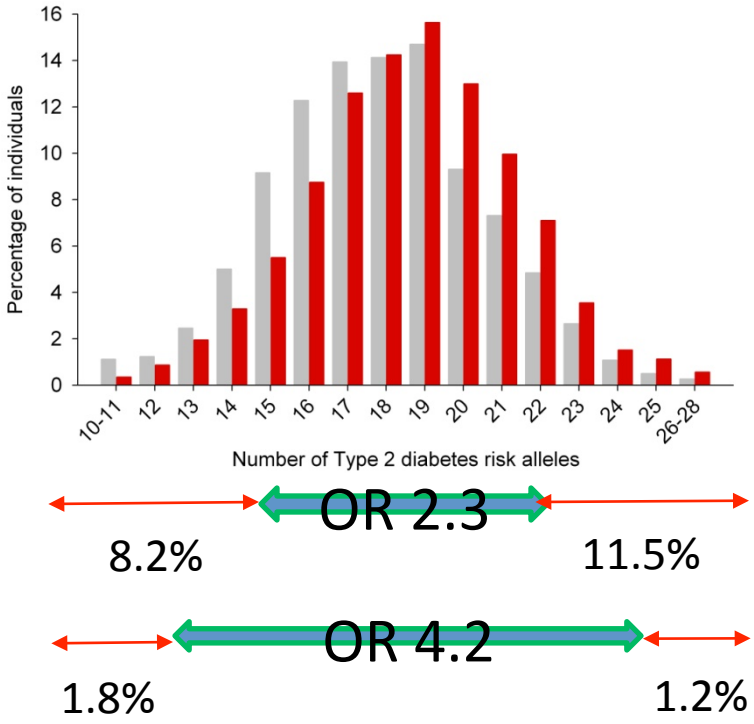
- Linkage analysis using families takes unbiased look at whole genome, but is underpowered for the size of genetic effects we expect to see for many complex genetic traits.

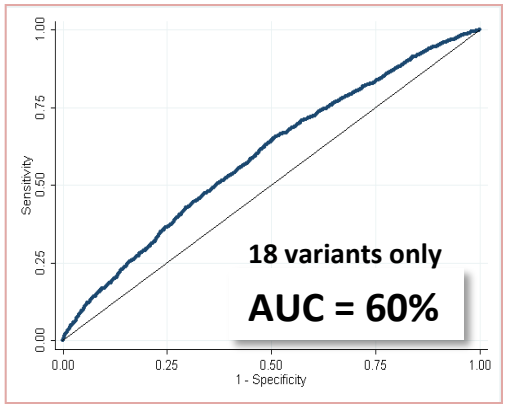
- Candidate-gene association studies have greater power to identify smaller genetic effects, but rely on *a priori* knowledge about disease etiology.

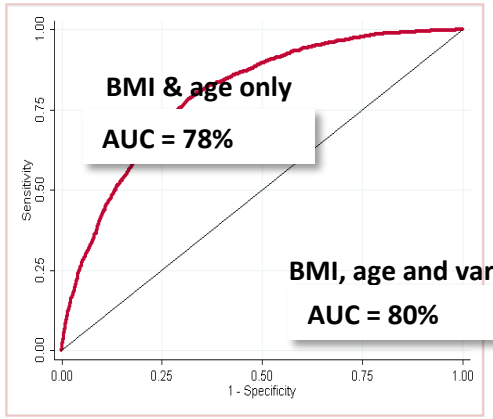
- Genome-wide association studies combine the genomic coverage of linkage analysis with the power of association studies to have much better chance of finding complex trait susceptibility variants.

    - Other advantages: agnostic search, large sample sizes, improved quality of genotyping, rigorous p-value thresholds, replication

Prediction not (yet) possible







Even with 40 genetic variants prediction is poor

Individual effects are modest

Only ~5-10% of genetic predisposition found

Weedon et al, PLOS, 2007
Lango et al, Diabetes 2008

Missing Heritability?

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Reasons for missing heritability

- "Common disease, common variant" is incorrect – study rarer variants
- Calculation of heritability effects is wrong?
- Not enough common variants of small effect detected
- Structural or other genomic variants more important
- Difficult to analyse gene-gene/gene-environment interactions and in general high-dimensional and systems biology data (i.e., combination of genomic, transcriptomic, proteomic, metabolomic data)

Ways forward…

- Further genetic discovery (denser genotyping)
- Better characterization of validated genes
- Use genes for causal inference (Mendelian randomization)
- Whole genome sequencing
- Systems biology approaches
- Development of clinically useful risk prediction models
- Other translation

Mendelian Randomization

Learning Objectives

- Be able to explain the concept of Mendelian randomization and discuss its potential value, whilst recognising its assumptions and limitations.

Outline

- Mendelian Randomization
  - Conceptual Overview
  - Assumptions
  - Effect estimation
  - Examples
  - Limitations and Current Advances

Definition of Mendelian randomization (MR)

- *"The use of genetic data on human participants in an observational setting to evaluate the potential causal nature of a modifiable risk factor"*
  - Recent examples of causal effects
    - Blood pressure, obesity, LDL-C, IL-6 and CVD
  - Recent examples of lack of causal effects
    - CRP, HDL-C and CVD

Burgess S, et al. J Compar Effect Res 2013;2:209-212

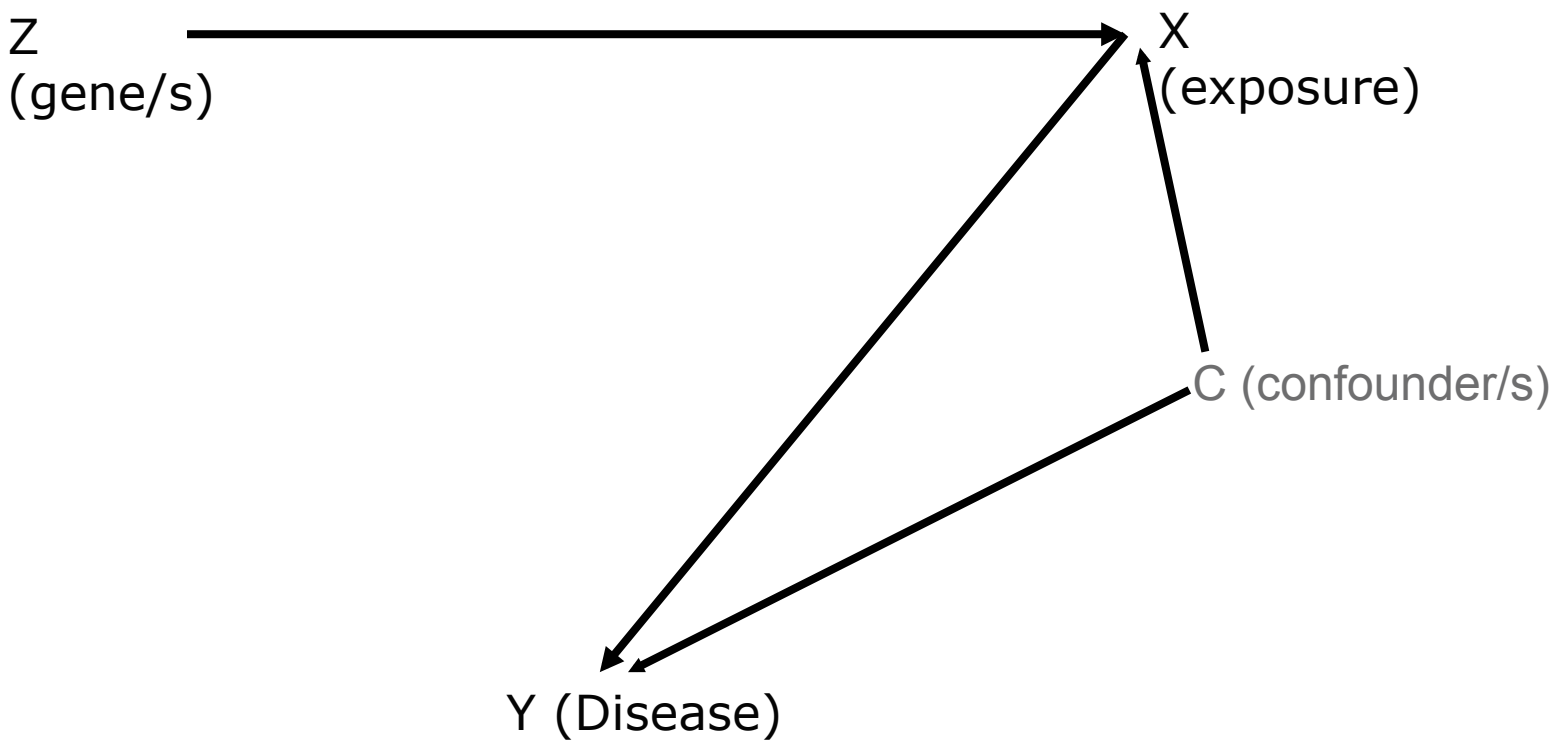
Attributes of Mendelian randomization

- Certain genetic polymorphisms produce phenotypes which mimic (reflect, serve as proxies for) the effect of environmental exposures

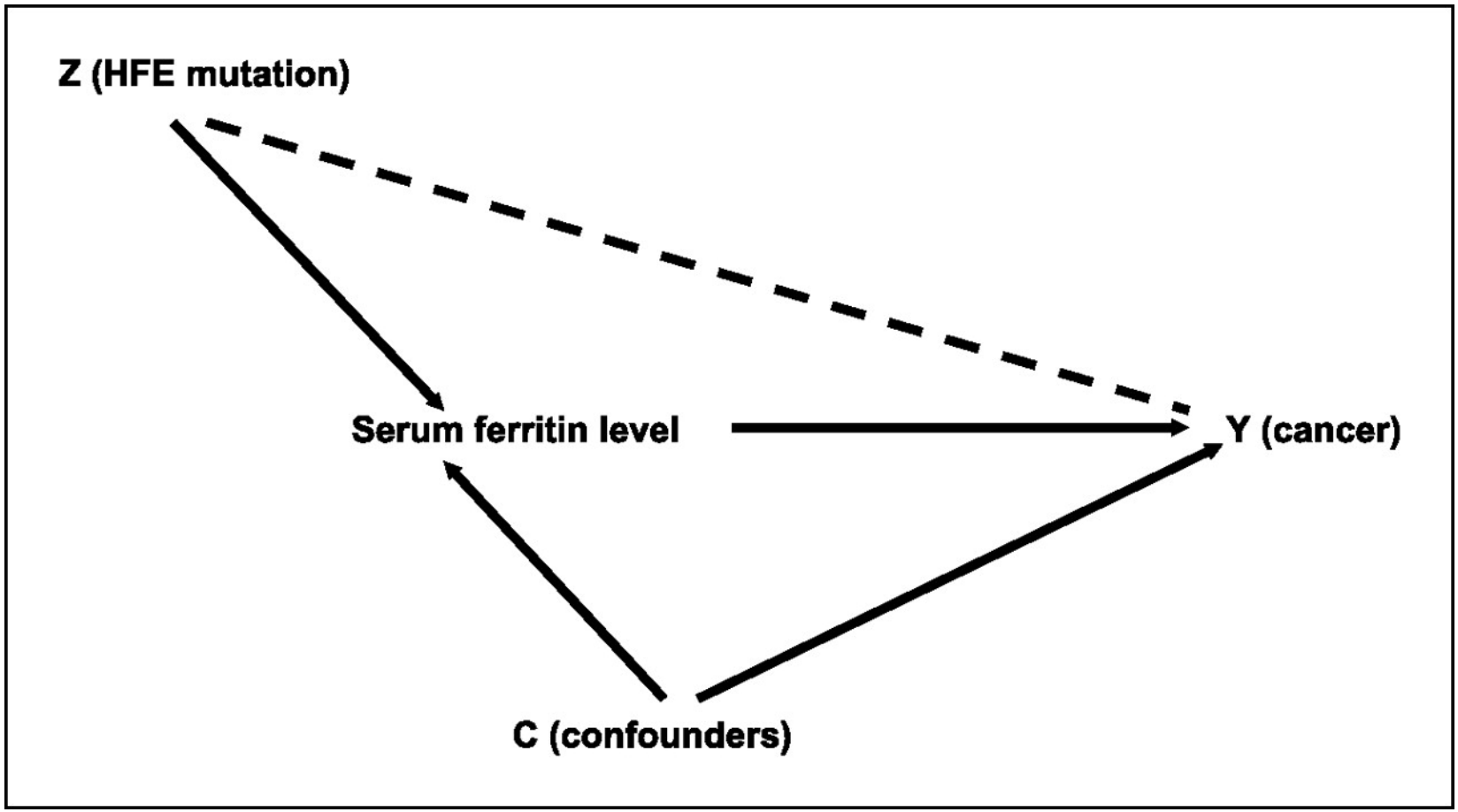
- Allelic variants mimicking environmental exposures (genetic instruments, instrumental variables (IV))
  - *IL6* gene for serum IL-6
  - Vitamin D metabolizing genes for serum 25(OH)D
  - *ALDH2* gene for alcohol intake
  - Lactase persistence gene for dairy product intake
  - *HFE* mutations for high serum iron

- Because of random assortment of alleles, MR reduces bias due to confounding
- MR also largely avoids exposure measurement error and reverse causation bias

Causal diagram in Mendelian randomization



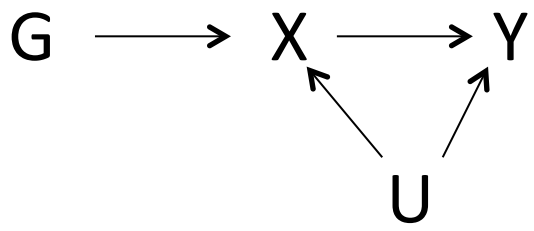
A directed acyclic graph depicting how the *HFE* gene can be used as a proxy (instrumental variable) for serum ferritin in a Mendelian randomization study of cancer.

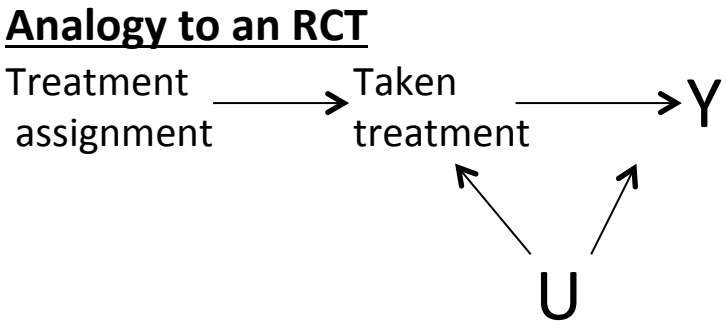


Schatzkin A et al. Cancer Prev Res 2009;2:104-113

Assumptions of MR

1. The genetic marker is associated with the exposure.
2. The genetic marker is independent of the outcome given the exposure and all confounders (exclusion restriction).
3. The genetic marker is independent of factors that confound the exposure-outcome association.





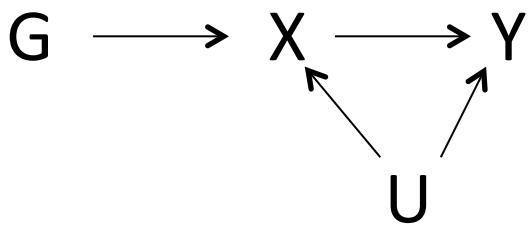
VanderWeele TJ, et al. Epidemiology 2014;25:427-435

Evaluating the MR assumptions

1. The genetic marker is associated with the exposure.
   - It can be easily evaluated in a dataset

2. The genetic marker is independent of the outcome given the exposure and all confounders.
   - Adjust for X in the G-Y association, but beware of collider bias (need to adjust also for U)

3. The genetic marker is independent of factors that confound the exposure-outcome association.
   - Test whether G is associated with measured U factors

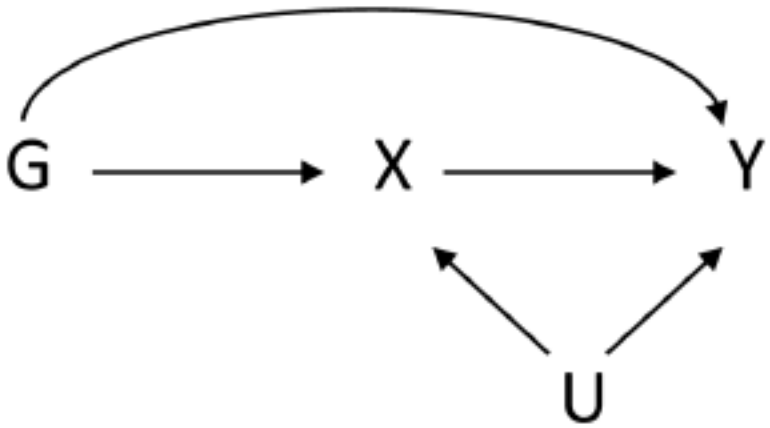


Potential violations of the exclusion restriction assumption

1. Pleiotropy (horizontal)
2. G*E interaction
3. G*G interaction
4. Linkage disequilibrium
5. Population stratification
6. Other reasons

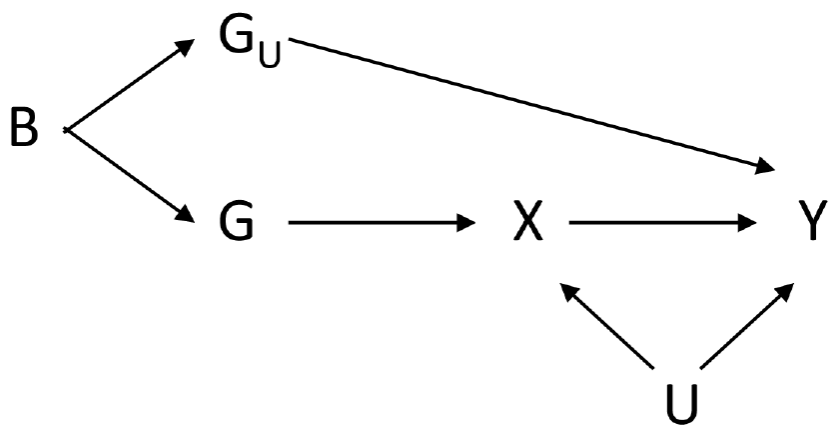
Potential violations of the exclusion restriction assumption

1. Pleiotropy (horizontal)
2. G*E interaction
3. G*G interaction
4. Linkage disequilibrium
5. Population stratification
6. Other reasons



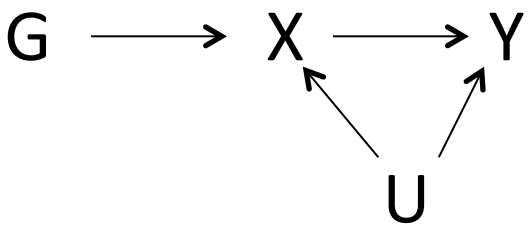
Potential violations of the exclusion restriction assumption

1. Pleiotropy (horizontal)
2. G*E interaction
3. G*G interaction
4. Linkage disequilibrium
5. Population stratification
6. Other reasons



Weak instrument bias

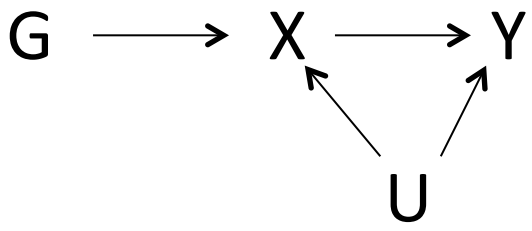
- The IV strength is measured by the F statistic from the X on G regression.
- If F<10, then weak instrument.
- Seek for parsimonious models.



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

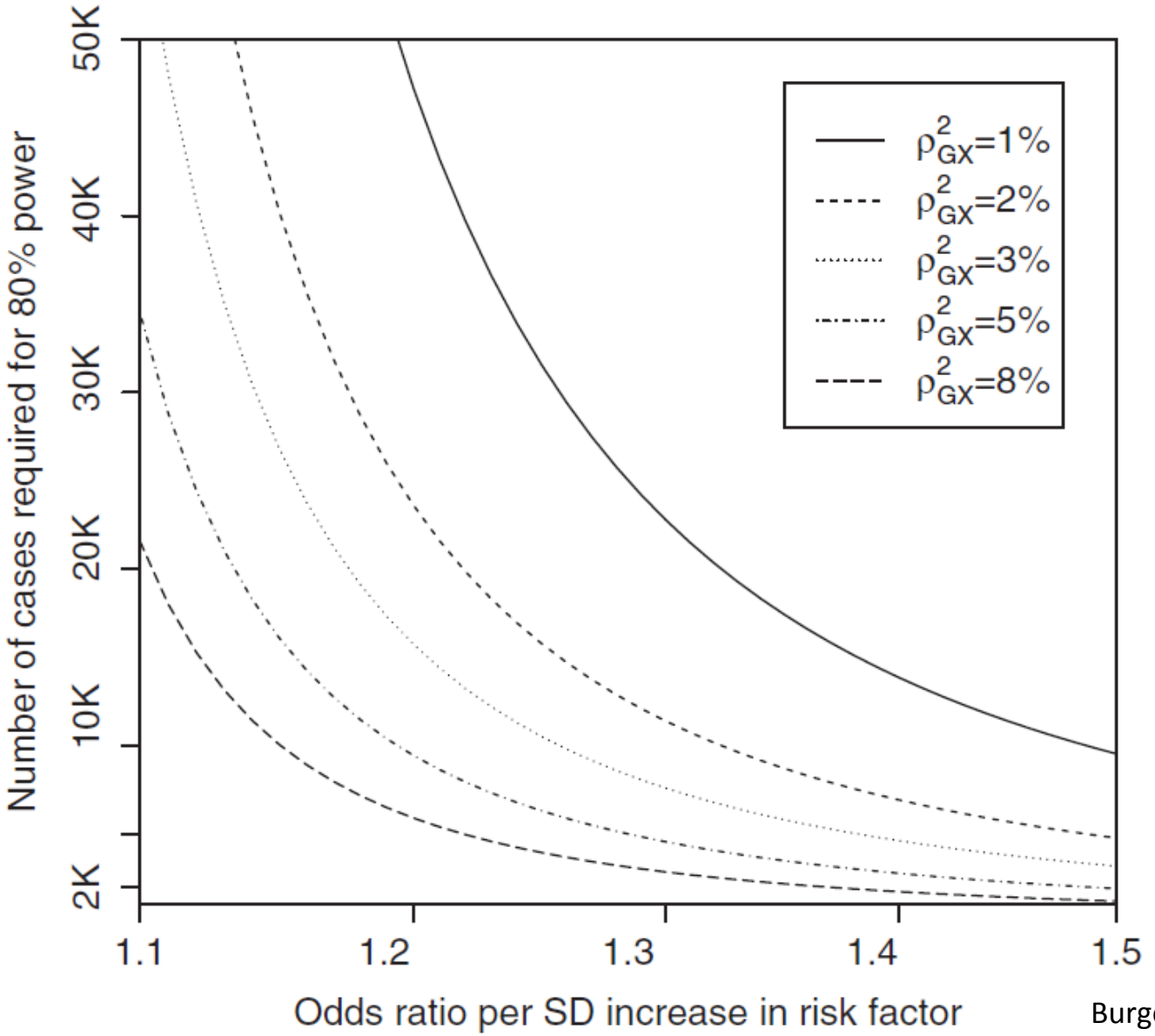
Instrumental variable (IV) estimators

- "Wald" or "ratio": b_{YG} / b_{XG}
- 2-stage least squares
- "Control function"
- Structural mean models
- Generalized method of moments



Palmer TM, et al. Am J Epidemiol 2011;173:1392-1403

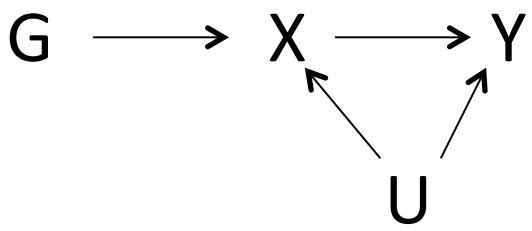
Power and sample size in MR studies



Burgess S. Int J Epidemiol 2014:922-929

Efficient designs of MR studies

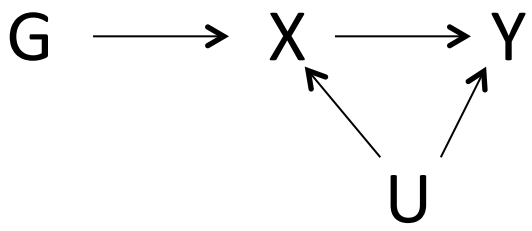
- In the traditional MR setting, data on G, X and Y are available for all participants.
- Subsample IV methods
- Two-sample IV methods
- Use of only summarized literature data



Pierce BL, et al. Am J Epidemiol 2013;178:1177-1184
Burgess S, et al. Genet Epidemiol 2013;37:658-665

Sub-sample and 2-sample IV methods

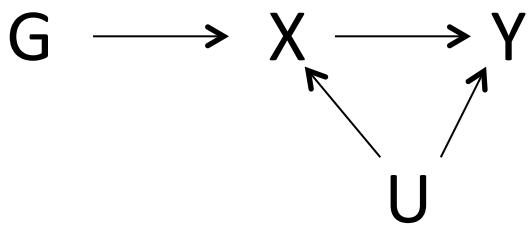
- Exposure data available for a subset (or an independent set) of participants
- >90% of the maximum power can be achieved by obtaining exposure data on only 20% of the sample
- Power for MR is most efficiently increased by increasing the sample size of the outcome-gene association



Pierce BL, et al. Am J Epidemiol 2013;178:1177-1184

Summarized literature data for MR studies

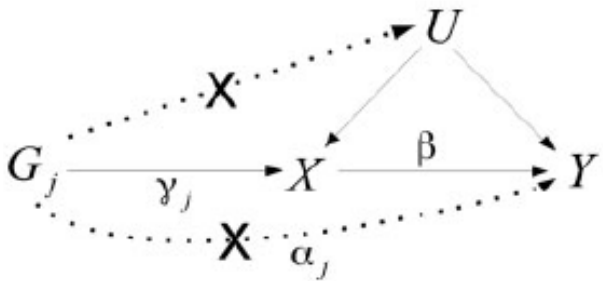
- Inverse-variance weighted combination of ratio estimates
- Likelihood-based method
- Empirical studies and simulations showed that these methods gave similar estimates and precision to the 2-stage least squares method.
- Limitation: the IV assumptions cannot be so fully assessed.



Burgess S, et al. Genet Epidemiol 2013;37:658-665

Further methods to assess MR assumptions

1. Over-identification test
   - H_0: the same causal effect of the risk factor on the outcome is identified by each genetic variant.
2. Goodness-of-fit test
   - H_0: each SNP included in the risk score has an association with the outcome that is proportional to its association with the exposure.
3. Cochran's Q test
   - H_0: homogeneity of MR estimates for each SNP
4. MR-Egger
5. Weighted median

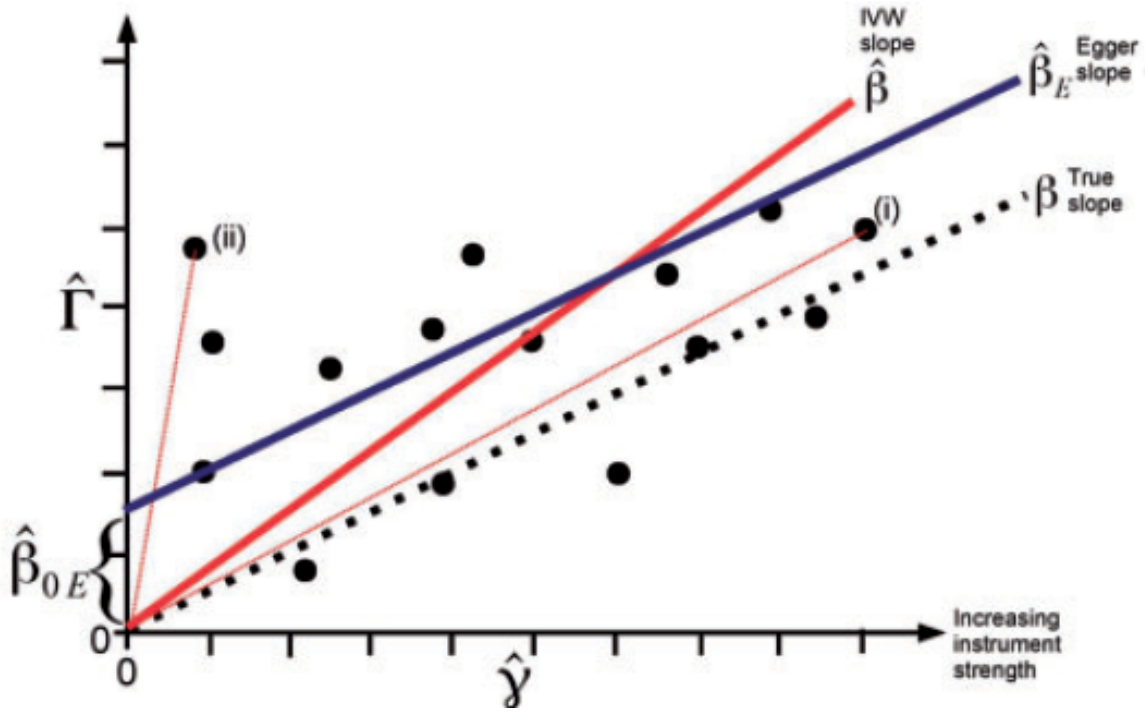


Glymour MM, et al. Am J Epidemiol 2012;175:332-9.
Bowden J, et al. Int J Epidemiol 2015;44:512-25.
Bowden J, et al. Genet Epidemiol 2016;40:304-14.

Further methods to assess MR assumptions

MR-Egger

- Bias caused by pleiotropy can be regarded as analogous to small study bias in meta-analysis
- Limitation: low power with few genetic variants



Slope: MR estimate adjusted for pleiotropy

P-value of intercept: test for existence of pleiotropy

Bowden J, et al. Int J Epidemiol 2015;44:512-25.

Further methods to assess MR assumptions

MR-Egger

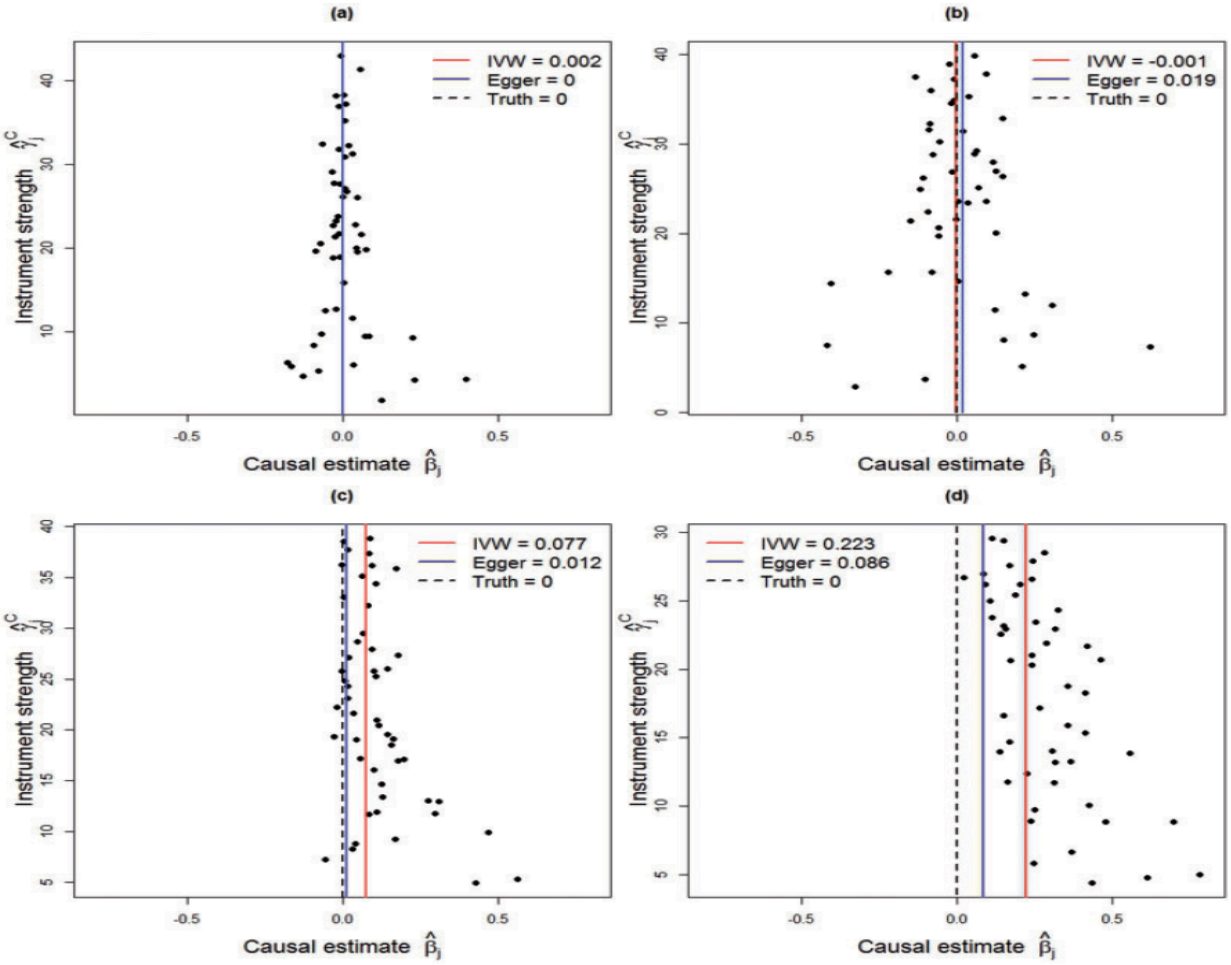


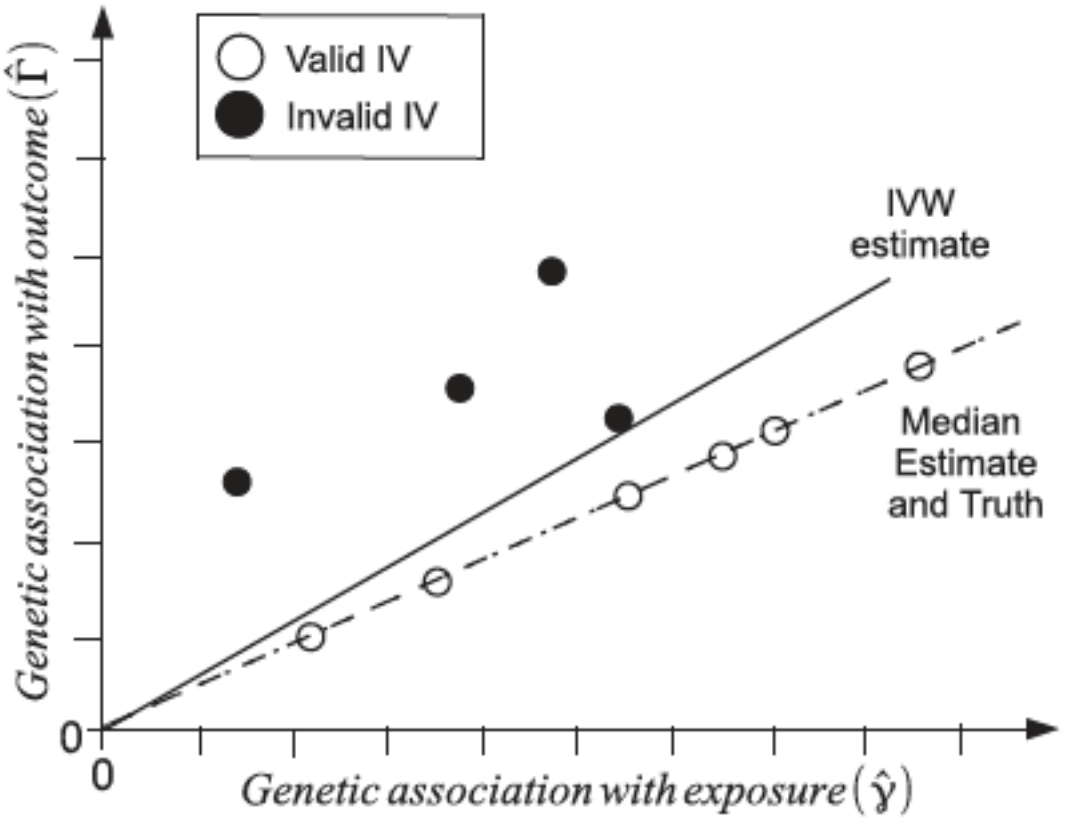
Figure 5. Funnel plots of minor allele frequency corrected genetic associations with exposure ($\hat{\gamma}_j^C$) against causal estimates based on each genetic variant individually ($\hat{\beta}_j$) for 50 IV estimates in four scenarios: (a) no pleiotropy; (b) balanced pleiotropy; (c) directional pleiotropy, InSIDE assumption satisfied; and (d) directional pleiotropy, InSIDE assumption not satisfied. The inverse-variance weighted (IVW, red) and MR-Egger (blue) causal effect estimates are also shown.

Bowden J, et al. Int J Epidemiol 2015;44:512-25.

Further methods to assess MR assumptions

Weighted median approach

- Consistent estimation even when up to 50% of the genetic variants are invalid instruments



Bowden J, et al. Genet Epidemiol 2016;40:304-14.

Further methods to assess MR assumptions

Table 5. Summary of methods considered in this paper

Method	Breakdown	IV2	IV3	Comments
Two-stage least squares	0%	✗	✗	Requires individual-level data. Biased when at least one genetic variant is an invalid IV.
Inverse-variance weighted (IVW)	0%	✗	✗	Equivalent to two-stage least squares method with summary data. Also biased when at least one genetic variant is an invalid IV.
Simple median	50%	√	√	Consistent when 50% of genetic variants are valid IVs. Inefficient compared with IVW and weighted median methods.
Weighted median	50%	√	√	Consistent when 50% of weight contributed by genetic variants is valid. Efficiency is similar to that of IVW method.
Penalized weighted median	50%	√	√	Equivalent to weighted median when there is no causal effect heterogeneity. Downweights the contribution of heterogeneous variants, so may have better finite sample properties, particularly if there is directional pleiotropy.
MR-Egger regression	100%	✗	√	Consistent when 100% of genetic variants are invalid, but requires variants to satisfy a weaker assumption (the InSIDE assumption). This assumption is not automatically violated by an association between a genetic variant and a confounder, but it would be violated if several variants were associated with the same confounder. Substantially less efficient than IVW and median-based methods, and more susceptible to weak instrument bias in a one-sample setting.

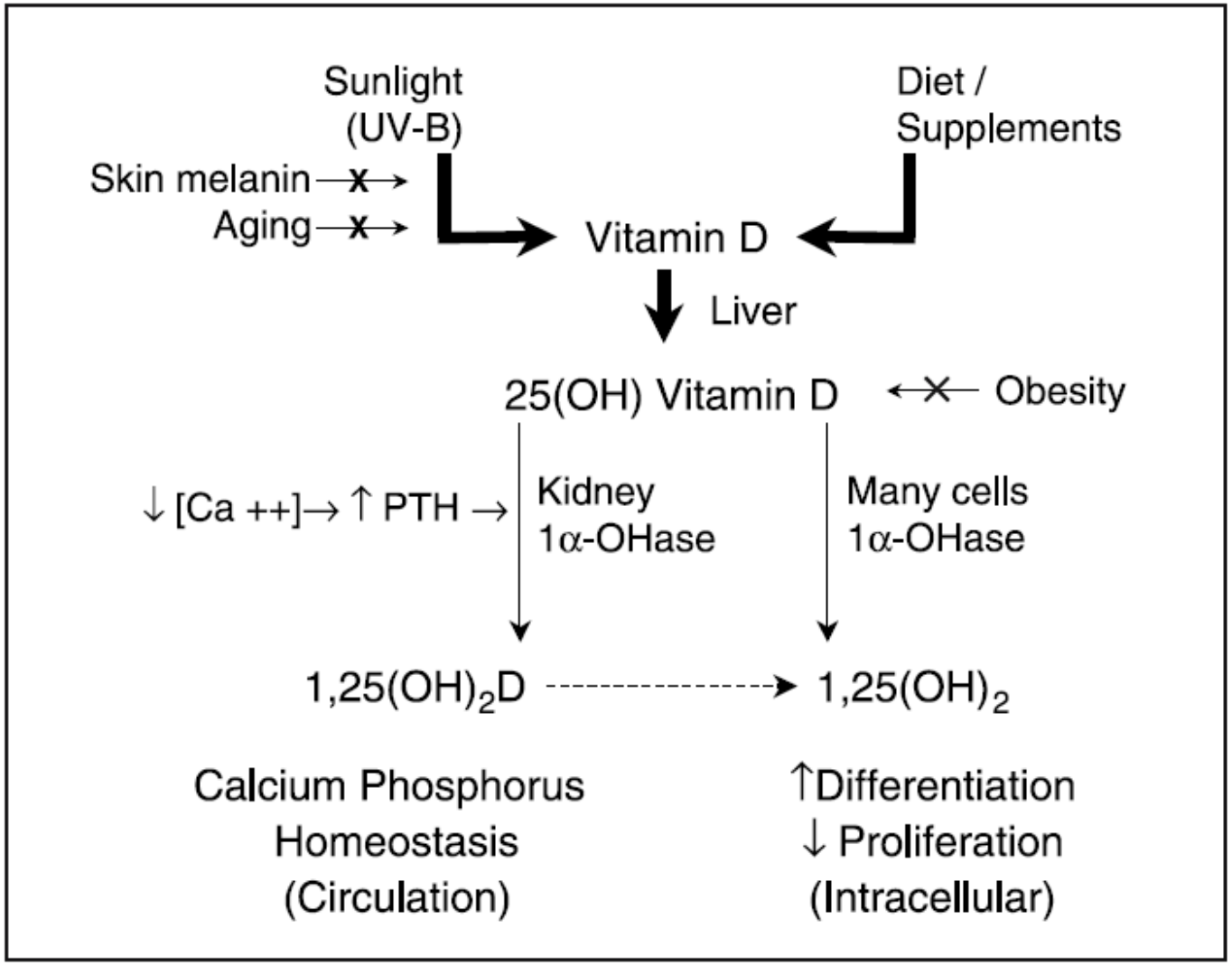
Breakdown refers to the breakdown level, the proportion of information that can come from invalid instrumental variables (IVs) before the method gives biased estimates. IV2 and IV3 refer to whether violations of the second (no association with confounders) and third (no direct effect on the outcome) instrumental variable assumptions are allowed (√) or not allowed (✗).

Bowden J, et al. Genet Epidemiol 2016;40:304-14.

Example: Circulating vitamin D concentrations (25(OH)D) and cancer risk

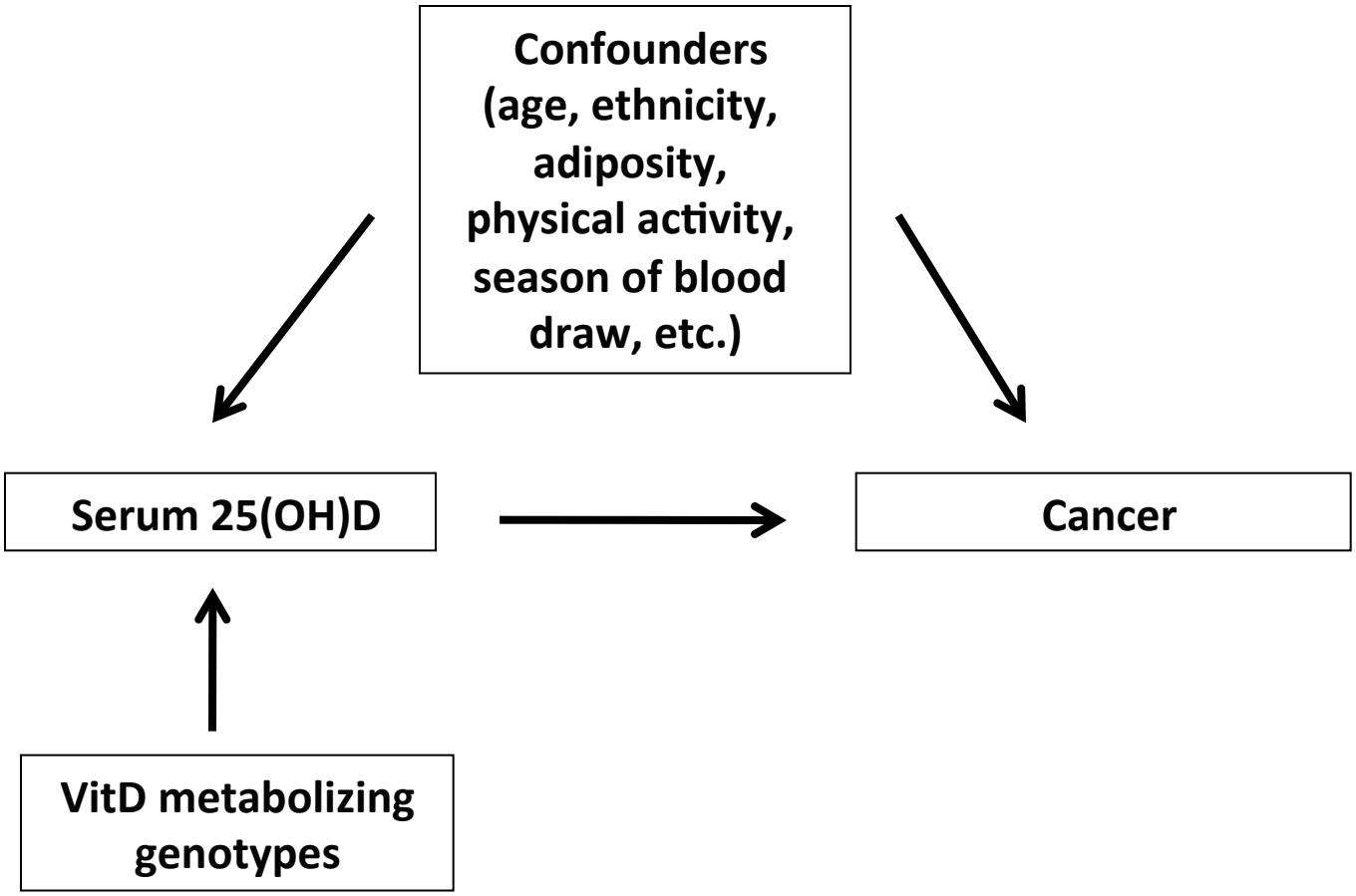
- Ample biological evidence for an anti-cancer role of 25(OH)D
  - Vitamin D metabolites control cellular growth and differentiation.
  - Administration of vitamin D analogues inhibits progression of several cancers in animal models and cell lines.
- Epidemiological studies have been inconclusive.
- Vitamin D supplementation trials currently provide no firm evidence for increase or decrease of cancer occurrence.

Proposed mechanism

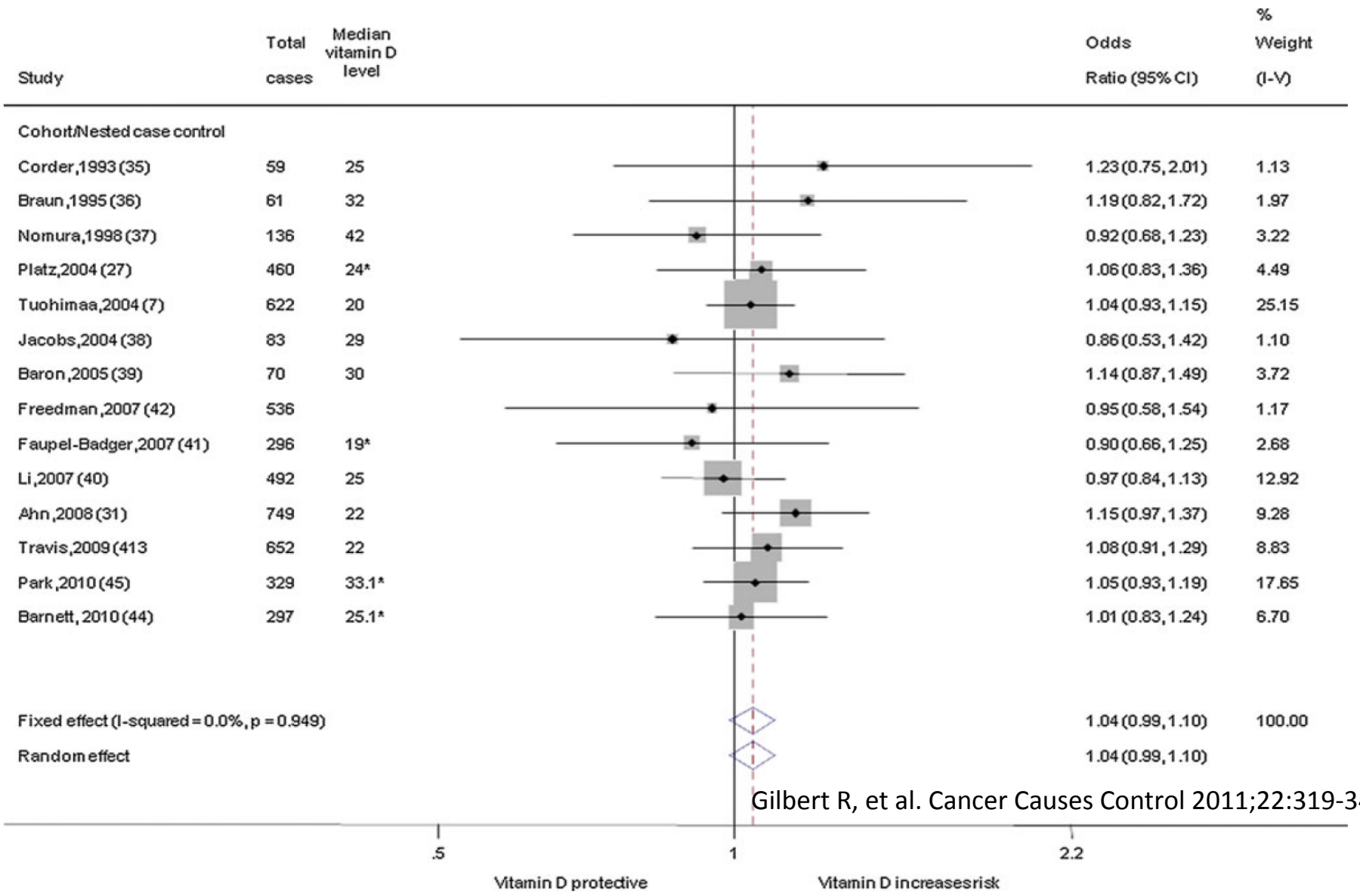


Giovannucci E. Cancer Causes Control 2005;16:83-95

Causal diagram: Vitamin D and cancer

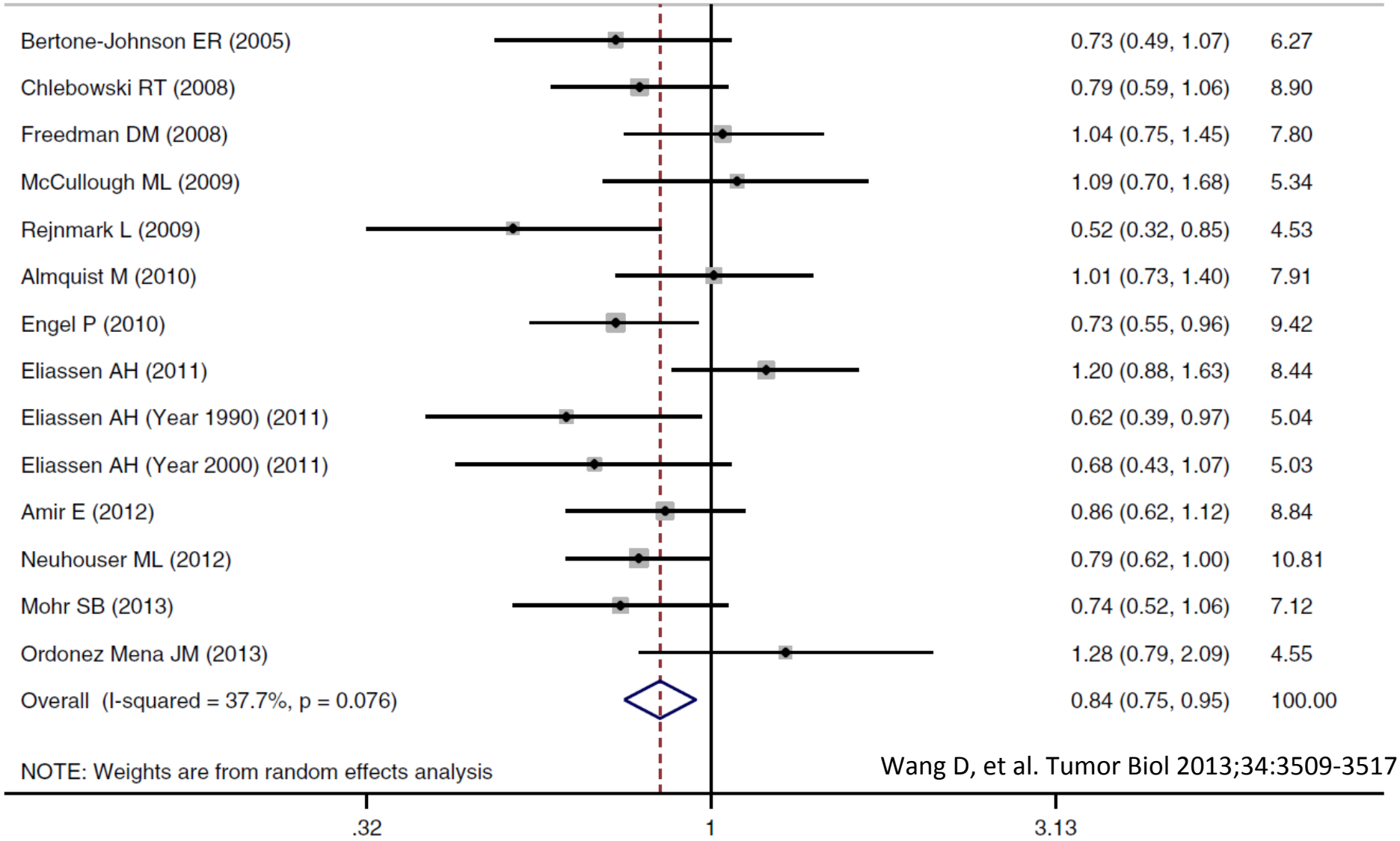


25(OH)D and prostate cancer risk



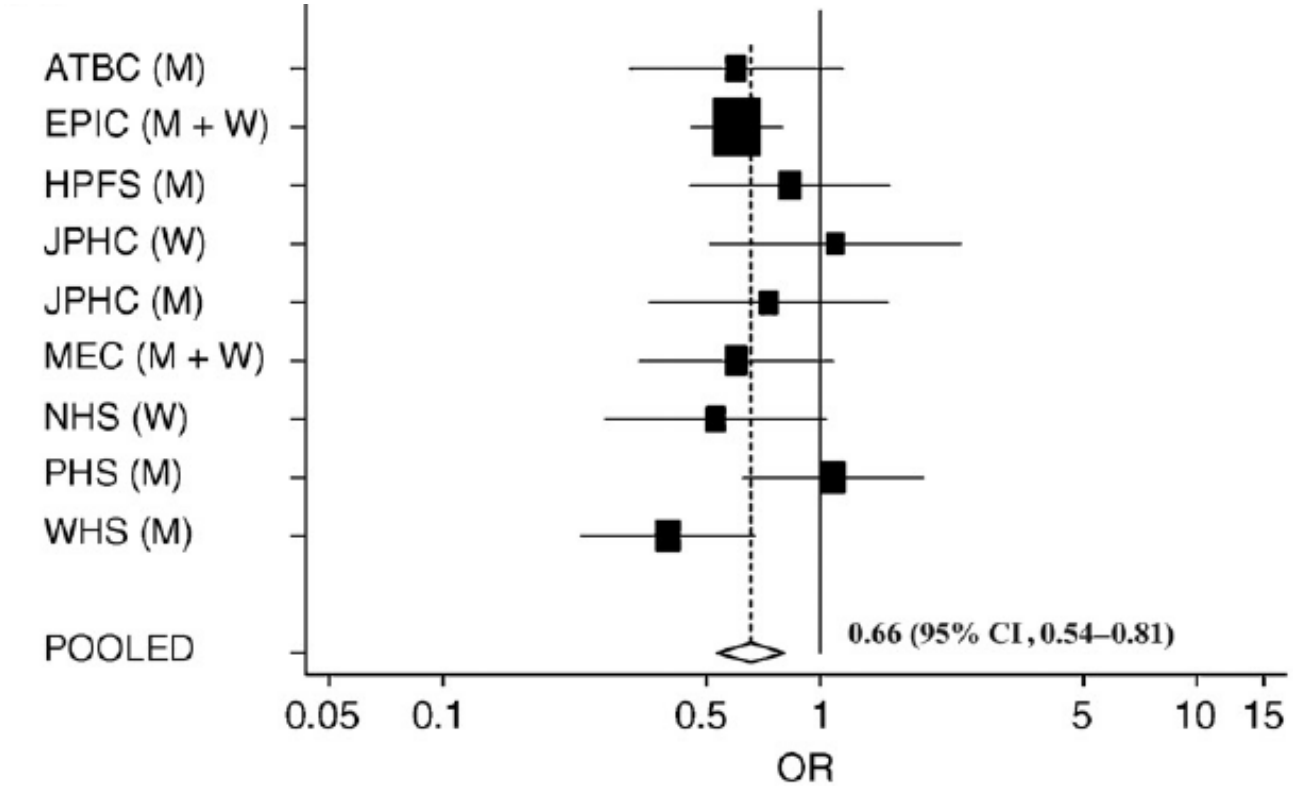
Gilbert R, et al. Cancer Causes Control 2011;22:319-340

25(OH)D and breast cancer risk



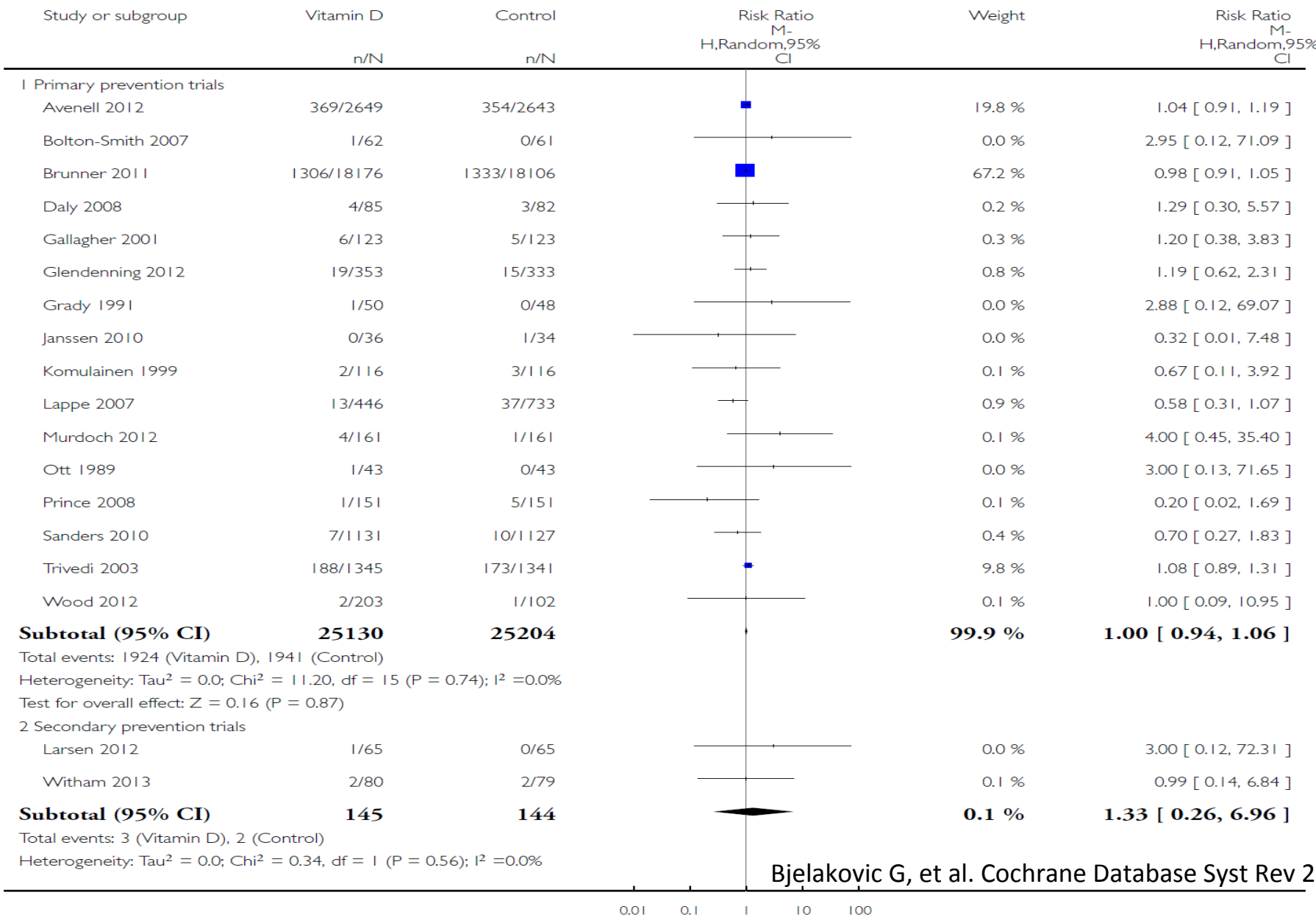
Wang D, et al. Tumor Biol 2013;34:3509-3517

25(OH)D and colorectal cancer risk



Lee JE, et al. Cancer Prev Res 2011;4:735-743

Vitamin D supplements and cancer risk

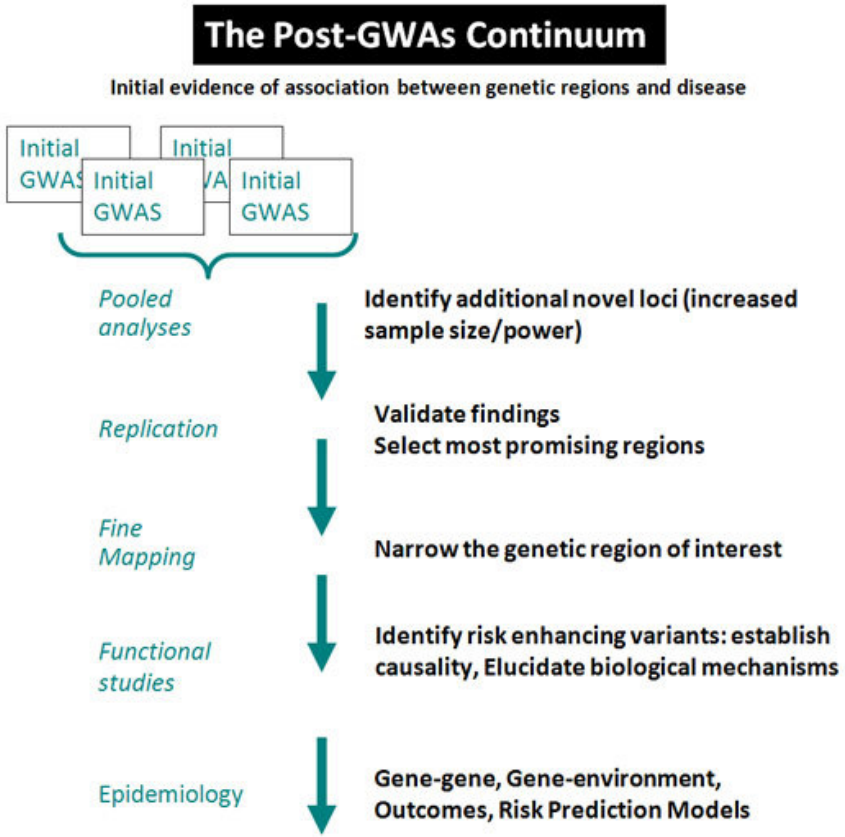


Study or subgroup	Vitamin D n/N	Control n/N	Risk Ratio M-H,Random,95% CI	Weight	Risk Ratio M-H,Random,95% CI
1 Primary prevention trials					
Avenell 2012	369/2649	354/2643		19.8 %	1.04 [0.91, 1.19]
Bolton-Smith 2007	1/62	0/61		0.0 %	2.95 [0.12, 71.09]
Brunner 2011	1306/18176	1333/18106		67.2 %	0.98 [0.91, 1.05]
Daly 2008	4/85	3/82		0.2 %	1.29 [0.30, 5.57]
Gallagher 2001	6/123	5/123		0.3 %	1.20 [0.38, 3.83]
Glendenning 2012	19/353	15/333		0.8 %	1.19 [0.62, 2.31]
Grady 1991	1/50	0/48		0.0 %	2.88 [0.12, 69.07]
Janssen 2010	0/36	1/34		0.0 %	0.32 [0.01, 7.48]
Komulainen 1999	2/116	3/116		0.1 %	0.67 [0.11, 3.92]
Lappe 2007	13/446	37/733		0.9 %	0.58 [0.31, 1.07]
Murdoch 2012	4/161	1/161		0.1 %	4.00 [0.45, 35.40]
Ott 1989	1/43	0/43		0.0 %	3.00 [0.13, 71.65]
Prince 2008	1/151	5/151		0.1 %	0.20 [0.02, 1.69]
Sanders 2010	7/1131	10/1127		0.4 %	0.70 [0.27, 1.83]
Trivedi 2003	188/1345	173/1341		9.8 %	1.08 [0.89, 1.31]
Wood 2012	2/203	1/102		0.1 %	1.00 [0.09, 10.95]
Subtotal (95% CI)	**25130**	**25204**		**99.9 %**	**1.00 [0.94, 1.06]**
Total events: 1924 (Vitamin D), 1941 (Control)					
Heterogeneity: $Tau^2 = 0.0$; $Chi^2 = 11.20$, df = 15 (P = 0.74); $I^2 = 0.0\%$					
Test for overall effect: Z = 0.16 (P = 0.87)					
2 Secondary prevention trials					
Larsen 2012	1/65	0/65		0.0 %	3.00 [0.12, 72.31]
Witham 2013	2/80	2/79		0.1 %	0.99 [0.14, 6.84]
Subtotal (95% CI)	**145**	**144**		**0.1 %**	**1.33 [0.26, 6.96]**
Total events: 3 (Vitamin D), 2 (Control)					
Heterogeneity: $Tau^2 = 0.0$; $Chi^2 = 0.34$, df = 1 (P = 0.56); $I^2 = 0.0\%$					

0.01 0.1 1 10 100

Bjelakovic G, et al. Cochrane Database Syst Rev 2014

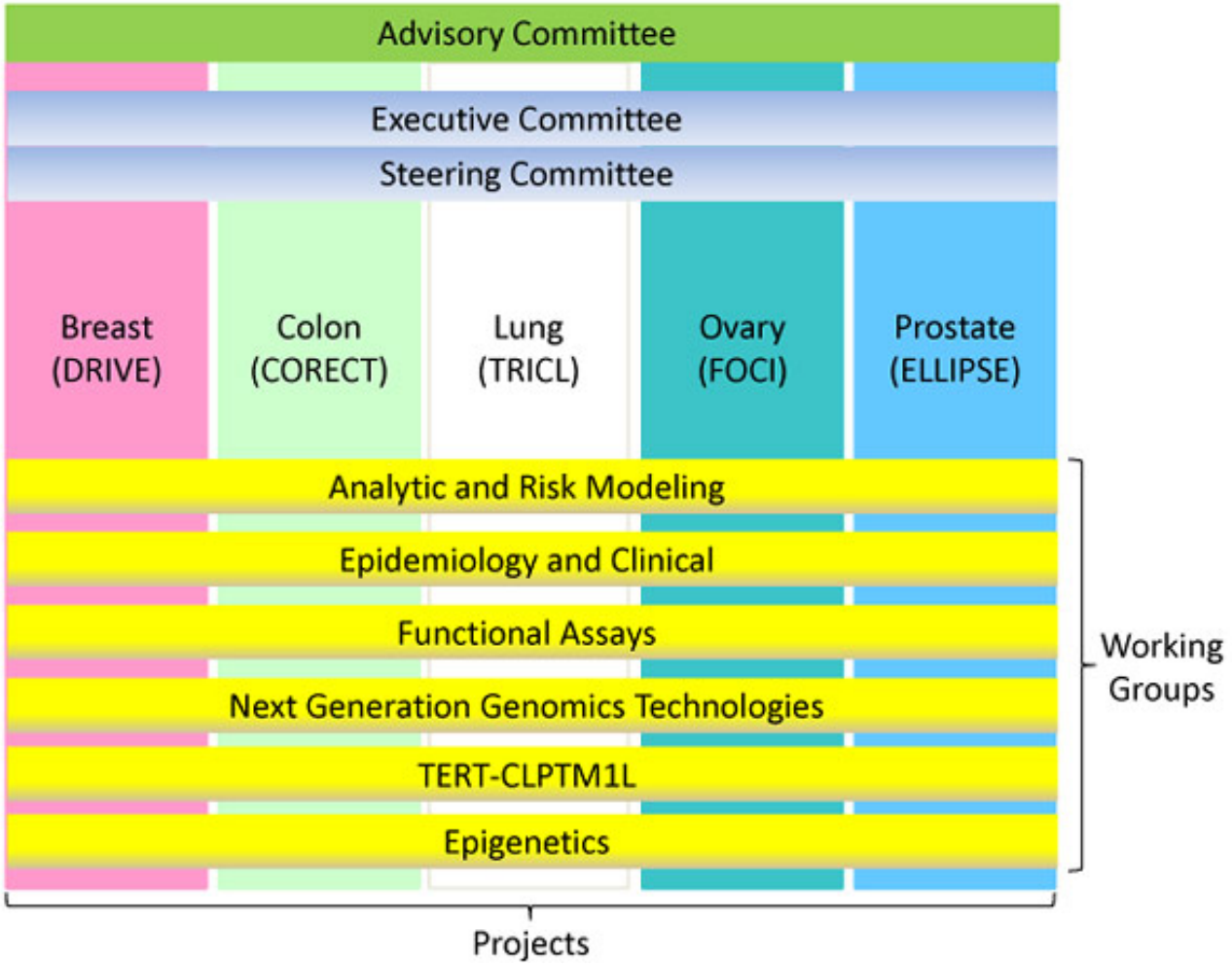
MR study of 25(OH)D and risk of five cancers in GAME-ON



Dimitrakopoulou V & Tsilidis KK, et al. BMJ 2017

http://epi.grants.cancer.gov/gameon/

GAME-ON Organizational Structure



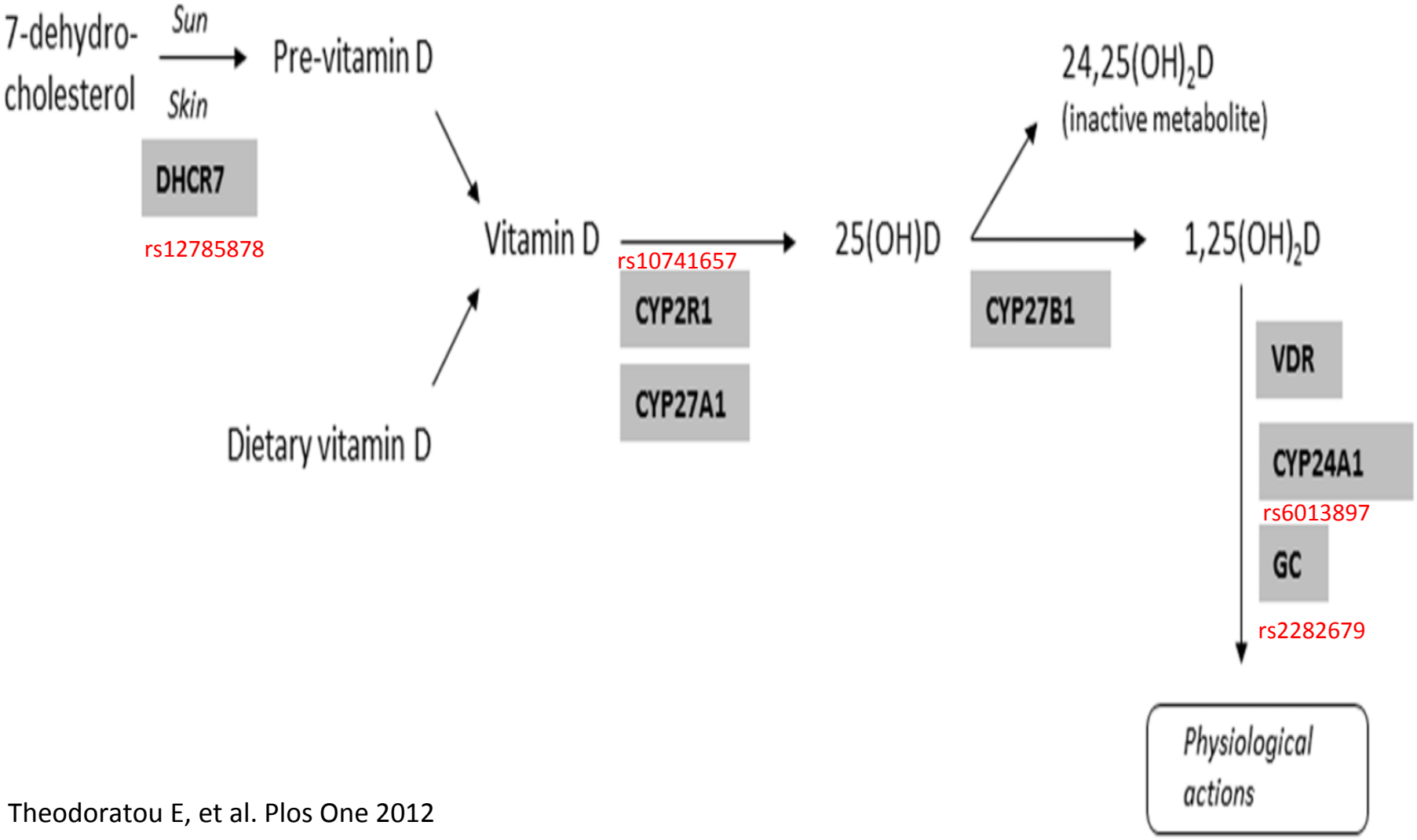
http://epi.grants.cancer.gov/gameon/

Genetic variants associated with vitamin D concentrations

Table 2. Characteristics of genetic variants associated with 25(OH)D concentrations in published GWA studies.

SNP	Chromosome	Locus	Risk Allele	Beta estimate	P-value	Source
25(OH)D: continuous[a]						
rs2282679	4	GC	G	2.25	$<3.4\times10^{-302}$	Vimaleswaran, et al. 2013[22]
rs10741657	11	CYP2R1	G	1.42	6.5×10^{-81}	Vimaleswaran, et al. 2013[22]
rs12785878	11	DHCR7/NADSYN1	G	1.55	6.4×10^{-129}	Vimaleswaran, et al. 2013[22]
rs6013897	20	CYP24A1	A	1.05	3.4×10^{-17}	Vimaleswaran, et al. 2013[22]
25(OH)D: <75nmol/L						
rs2282679	4	GC	C	0.49	3.5×10^{-50}	Wang, et al. 2010[20]
rs10741657	11	CYP2R1	A	0.19	9.4×10^{-11}	Wang, et al. 2010[20]
rs12785878[d]	11	DHCR7/NADSYN1	A	0.19	4.1×10^{-10}	Wang, et al. 2010[20]
25(OH)D: <50nmol/L						
rs2282679	4	GC	C	0.40	7.5×10^{-33}	Wang, et al. 2010[20]
rs12785878[d]	11	DHCR7/NADSYN1	A	0.19	4.7×10^{-9}	Wang, et al. 2010[20]
25(OH)D: <25nmol/L						
rs2282679	4	GC	C	0.60	2.5×10^{-8}	Ahn, et al. 2010[19]

Genetic variants associated with vitamin D concentrations



Theodoratou E, et al. Plos One 2012

Summary statistics of studies included in the GAME-ON consortium

Table 1. Number of cancer cases and controls and statistical power.

Cancer Type	Cases	Controls	Minimum detectable OR[a] ($R^2 = 0.03$)	Minimum detectable OR[a] ($R^2 = 0.05$)	OR (95% CI) in published meta-analyses[b]
Colorectal					
All – GAME-ON	5,100	4,831	0.72/1.39	0.78/1.28	0.74 (0.63, 0.89)[(2)]
All – GECCO	11,488	11,679	0.81/1.23	0.85/1.18	
All (women) – GECCO	6,132	6,380	0.75/1.33	0.80/1.25	NR
All (men) – GECCO	5,356	5,297	0.73/1.37	0.78/1.28	NR
Colon – GECCO	7,678	11,679	0.78/1.28	0.83/1.20	NR
Rectal – GECCO	2,783	11,679	0.68/1.47	0.75/1.33	NR
Distal Colon – GECCO	3,354	11,679	0.70/1.43	0.77/1.30	NR
Proximal Colon – GECCO	4,185	11,679	0.73/1.37	0.79/1.27	NR
Breast (DRIVE)					
All	15,748	18,084	0.84/1.19	0.87/1.15	0.89 (0.81, 0.98)[(40)]
ER-negative	4,939	13,128	0.75/1.29	0.80/1.22	NR
Prostate					
All – PRACTICAL	22,898	23,054	0.86/1.16	0.89/1.12	1.04 (0.99, 1.10)[(46)]
All – GAME-ON	14,159	12,712	0.82/1.22	0.86/1.17	
Aggressive – GAME-ON	4,445	12,724	0.74/1.30	0.79/1.23	0.98 (0.84, 1.15)[(46)]
Ovarian (FOCI)					
All	4,369	9,123	0.73/1.33	0.79/1.25	0.91 (0.79, 1.04)[(54)]
Clear-cell	356	9,123	0.19/1.86	0.36/1.67	NR
Endometrioid	715	9,123	0.43/1.62	0.55/1.48	NR
Serous	2,556	9,123	0.67/1.39	0.74/1.30	NR
Lung (TRICL-ILLCO)					
All	12,537	17,285	0.82/1.20	0.86/1.16	0.98 (0.96, 0.99)[(58)]
Adenocarcinoma	3,804	16,289	0.73/1.30	0.78/1.23	NR
Squamous	3,546	16,434	0.72/1.31	0.78/1.24	NR

Statistical analysis

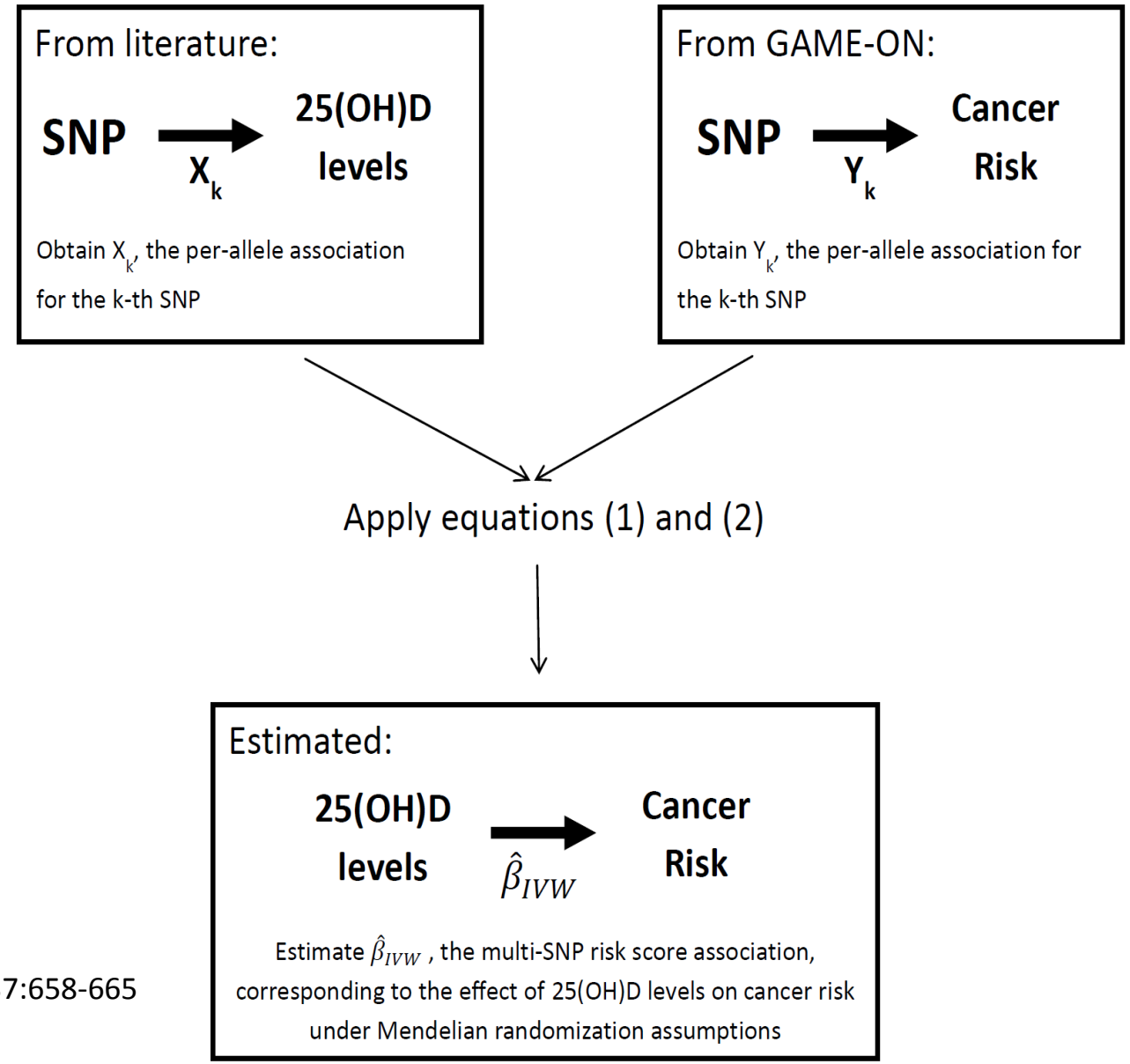
- SNPs and Cancer association
  - Standard quality control steps
  - Imputation using the 1K Genomes Project Phase 1 version 3
  - Logistic regression adjusted for age, sex, smoking, top PC
  - By cancer histology, aggressiveness, hormone receptor status
  - Fixed-effects meta-analysis to combine results by GWA studies
- Instrumental variable (IV) analysis
  - Multi-SNP 25(OH)D score
  - Using only summary association estimates
  - Inverse-variance weighted average of betas of the three SNPs
  - Likelihood-based method
  - Checking IV assumptions

Burgess S, et al. Genet Epidemiol 2013;37:658-665

Schematic of IV analysis

$$\hat{\beta}_{IVW} = \frac{\sum_k X_k Y_k \sigma_{Y_k}^{-2}}{\sum_k X_k^2 \sigma_{Y_k}^{-2}} \quad (1)$$

$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_k X_k^2 \sigma_{Y_k}^{-2}}} \quad (2)$$



Burgess S, et al. Genet Epidemiol 2013;37:658-665

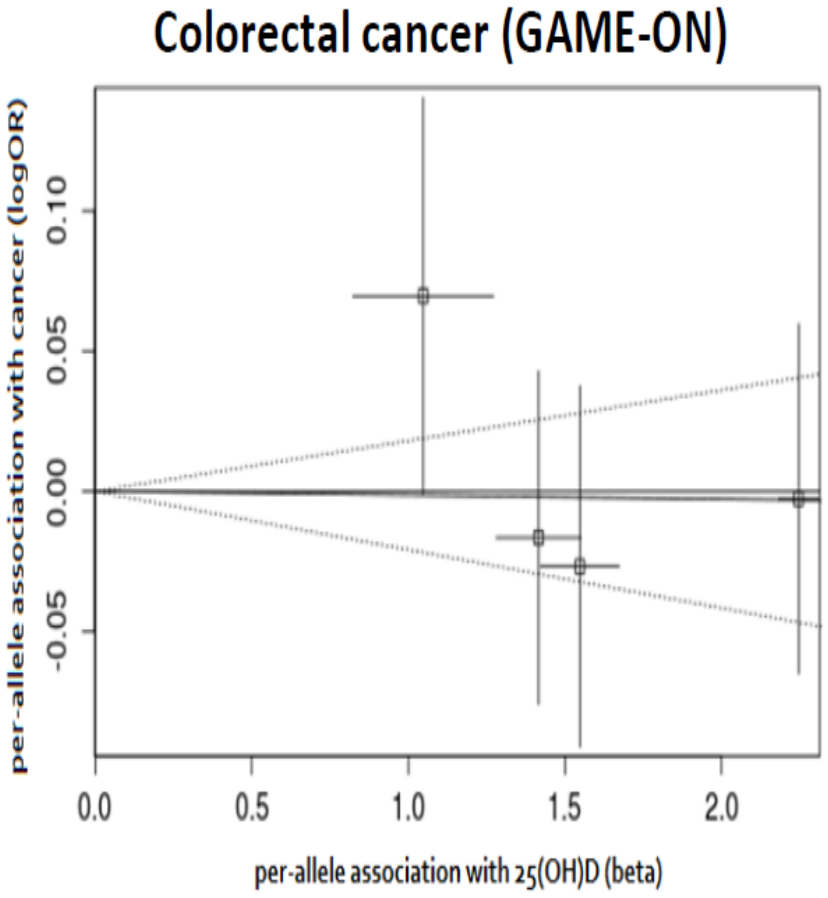
Instrumental variable corrected odds ratios of cancer risk per cont. 25(OH)D

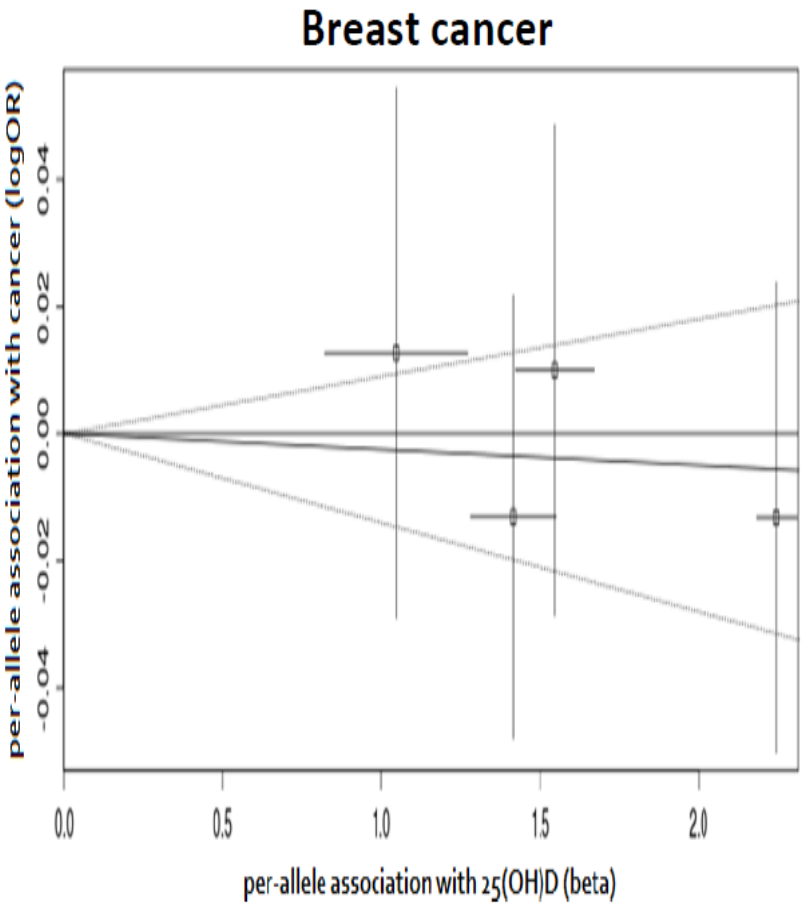
Table 3. Mendelian randomization estimates between multi-SNP risk scores of continuous 25(OH)D and cance risk calculated using the inverse-variance weighted method (left) and the likelihood method (right).

	Inverse-Variance Weighted Method			Likelihood Method		
Cancer Type	**OR[a]**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Colorectal						
All – GAME-ON	1.00	(0.98, 1.02)	0.90	1.00	(0.98, 1.02)	0.90
All – GECCO	1.00	(0.99, 1.02)	0.61	1.00	(0.99, 1.02)	0.61
All (women) – GECCO	1.00	(0.99, 1.02)	0.65	1.00	(0.99, 1.02)	0.66
All (men) – GECCO	1.00	(0.98, 1.02)	0.81	1.00	(0.98, 1.02)	0.81
Colon – GECCO	1.00	(0.99, 1.02)	0.57	1.00	(0.99, 1.02)	0.57
Rectal – GECCO	1.00	(0.98, 1.02)	0.95	1.00	(0.98, 1.02)	0.95
Distal Colon – GECCO	1.00	(0.98, 1.02)	0.92	1.00	(0.98, 1.02)	0.92
Proximal Colon – GECCO	1.01	(0.99, 1.03)	0.31	1.01	(0.99, 1.03)	0.31
Breast						
All	1.00	(0.99, 1.01)	0.68	1.00	(0.99, 1.01)	0.68
ER-negative	0.99	(0.98, 1.01)	0.45	0.99	(0.98, 1.01)	0.45
Prostate						
All – PRACTICAL	1.01	(0.99, 1.02)	0.09	1.01	(1.00, 1.02)	0.09
All – GAME-ON	1.00	(0.98, 1.01)	0.80	1.00	(0.98, 1.01)	0.81
Aggressive – GAME-ON	0.99	(0.97, 1.01)	0.53	0.99	(0.97, 1.01)	0.53
Ovarian						
All	0.99	(0.97, 1.01)	0.44	0.99	(0.97, 1.01)	0.44
Clear-cell	1.00	(0.95, 1.06)	0.91	1.00	(0.95, 1.06)	0.91
Endometrioid	1.01	(0.98, 1.05)	0.54	1.01	(0.98, 1.05)	0.54
Serous	0.98	(0.96, 1.01)	0.18	0.98	(0.96, 1.01)	0.18
Lung						
All	1.00	(0.99, 1.01)	0.97	1.00	(0.99, 1.01)	0.97
Adenocarcinoma	1.00	(0.98, 1.02)	0.94	1.00	(0.98, 1.02)	0.94
Squamous	1.00	(0.99, 1.02)	0.63	1.00	(0.99, 1.02)	0.63

[a]The odds ratios (ORs) represent increase/decrease of risk per unit increase in nmol/L in the log scale of

Per-allele association with cancer risk against per-allele association with cont. 25(OH)D





Evaluating IV assumptions

- Excluding SNPs with lack of consistency or biological explanation in the association with 25(OH)D (F>500)
- Over-identification test
  - H_0: the same causal effect of the risk factor on the outcome is identified by each genetic variant.
  - All p-values>0.05
- Goodness-of-fit test
  - H_0: each SNP included in the risk score has an association with cancer risk that is proportional to its association with 25(OH)D.
  - All p-values>0.05
- MR-Egger and weighted median approaches did not show evidence of assumption violation
- No evidence in published GWAS that the four single nucleotide polymorphisms associated with vitamin D were genome-wide significantly associated with any other phenotype.

Discussion

- A multi-SNP score for 25(OH)D was not associated with risk of five cancers.
- Results were consistent using two different analytic approaches.
- IV assumptions don't seem violated, although we cannot prove their validity.
- Limitations due to use of summary data
  - Cannot perform stratified analyses
  - Assumed linear associations
  - Cannot more fully assess IV assumptions
- Other limitations
  - Small fraction of 25(OH)D variation explained by the three used SNPs
  - Potential pleiotropy of used SNPs?

Limitations and promise of MR studies

- Lack of suitable polymorphisms for studying modifiable exposures
- Failure to establish reliable associations between genotype-exposure and genotype-disease due to limited sample sizes
- Confounding due to linkage disequilibrium and population stratification
- Pleiotropy and multi-functionality of genes
- Need large sample sizes (because gene variants typically yield only small changes in exposure variable)
- Need replication!
- **But, great potential of MR to assist causal inference in the future given large samples from genetic consortia, new efficient study design methods and new methods for testing MR assumptions (e.g., MR Egger, weighted median, etc.)**