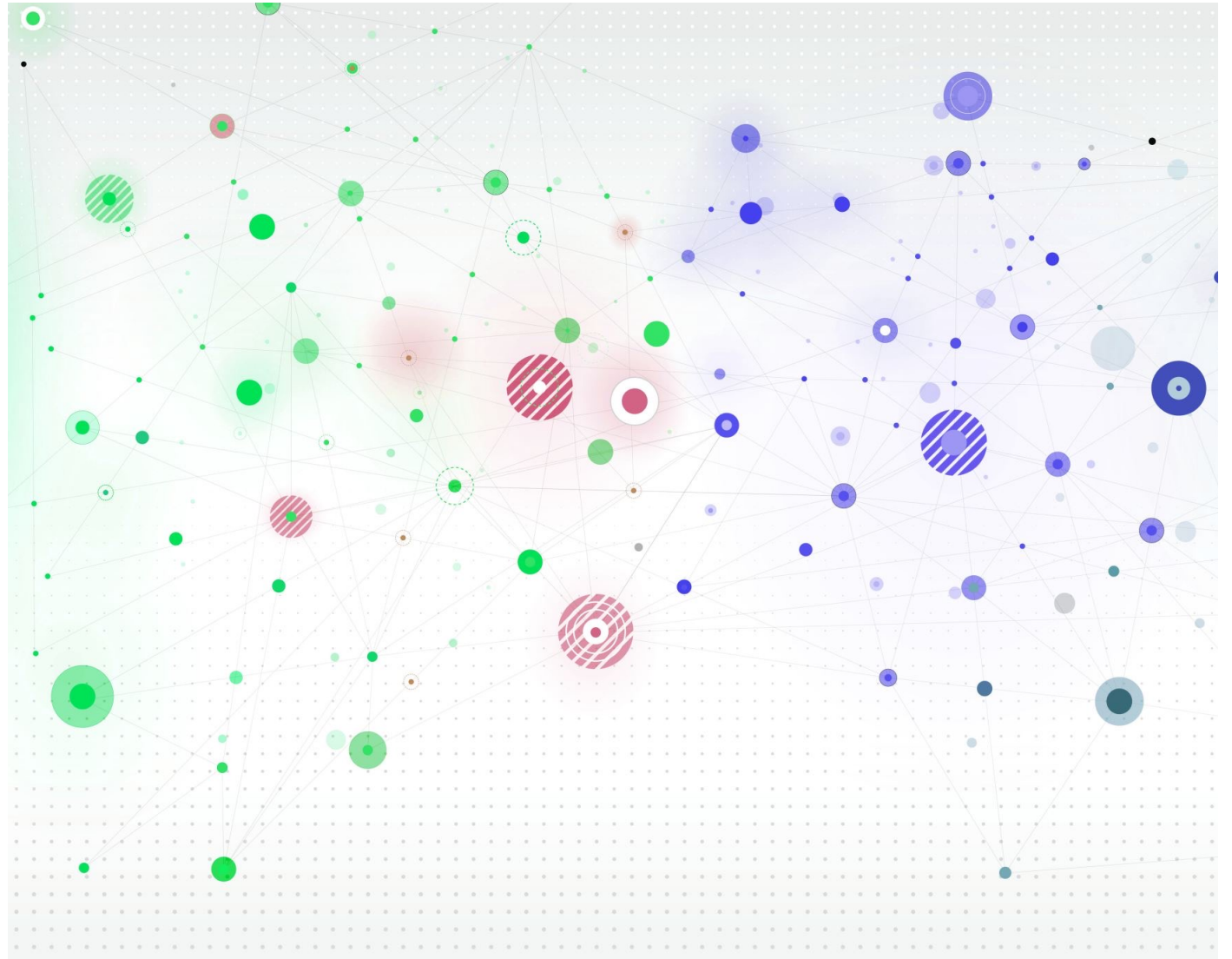

LEVERAGING BIG DATA TO STUDY CAUSAL MECHANISMS IN COMPLEX DISEASES

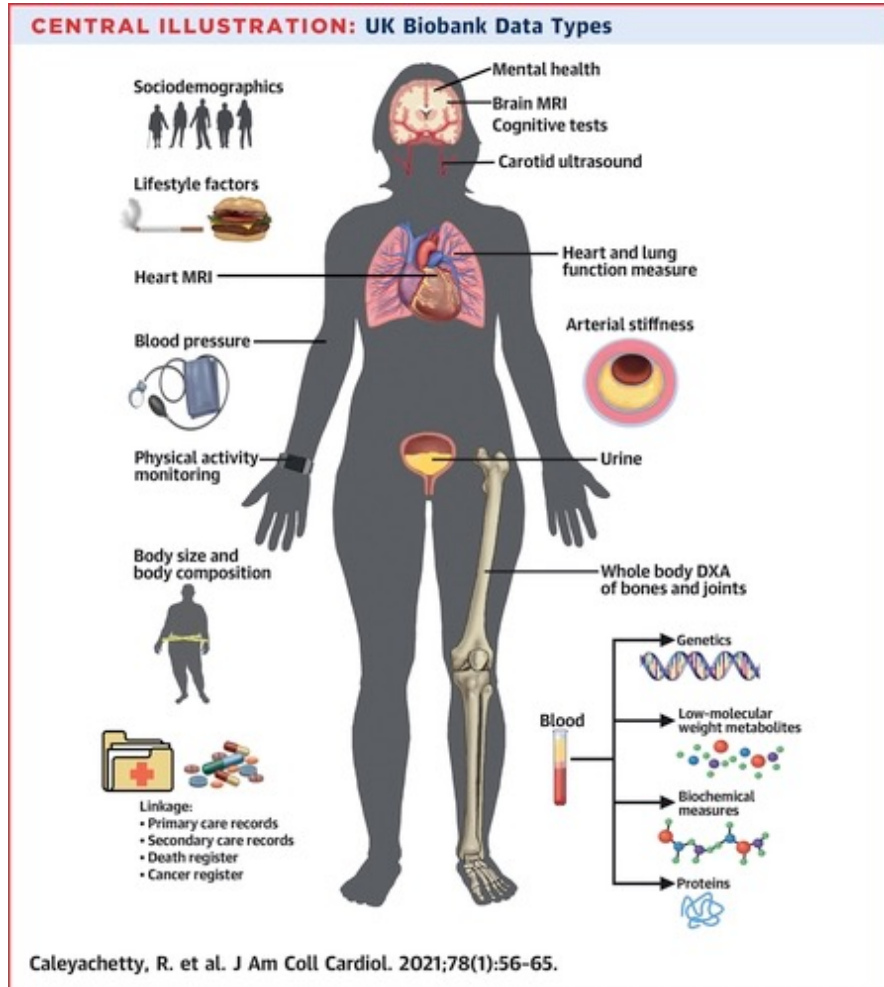
Ioanna Tzoulaki

**Research Director, Biomedical Research
Foundation Academy of Athens**

Professor in Chronic Disease
Epidemiology, School of Public Health,
Imperial College London



Large biobank with large N and deep phenotyping



Linked Electronic Health Record Data



Improved (novel) Methodologies

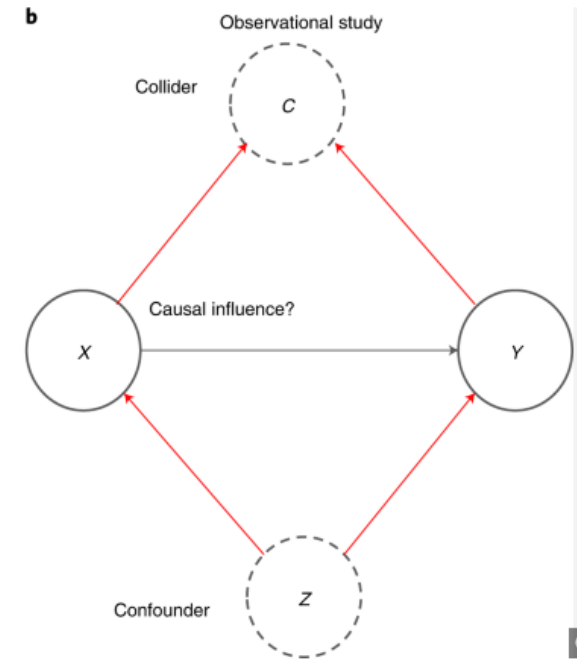




PREDICTION VS CAUSALITY

CAUSALITY IN EPIDEMIOLOGY

- Observational studies vs Randomized Control Trials
 - Confounding
 - Reverse causation

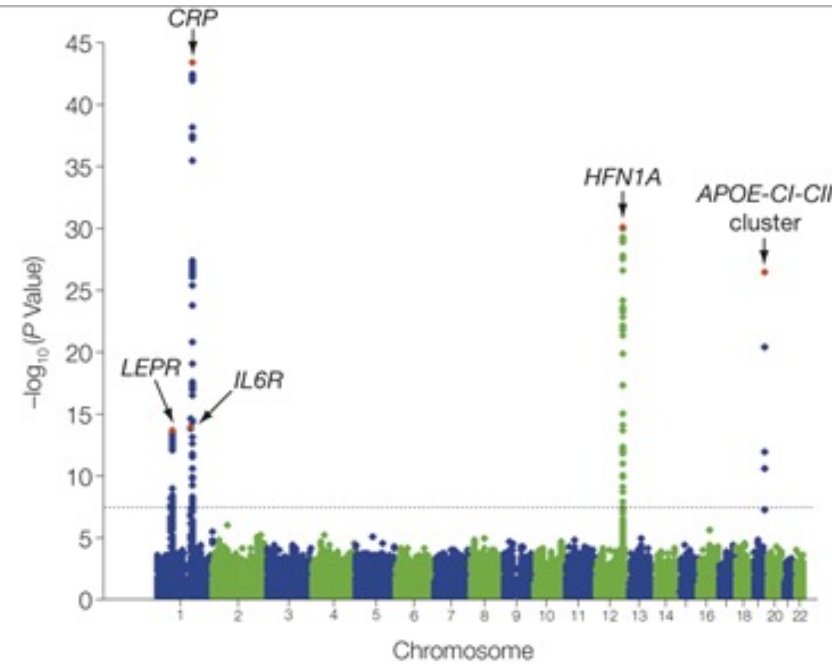


- Challenge → use observational (big) data to infer causality (or choose the trials more likely to be successful)
 - Mendelian Randomization
 - Drug target emulation
-

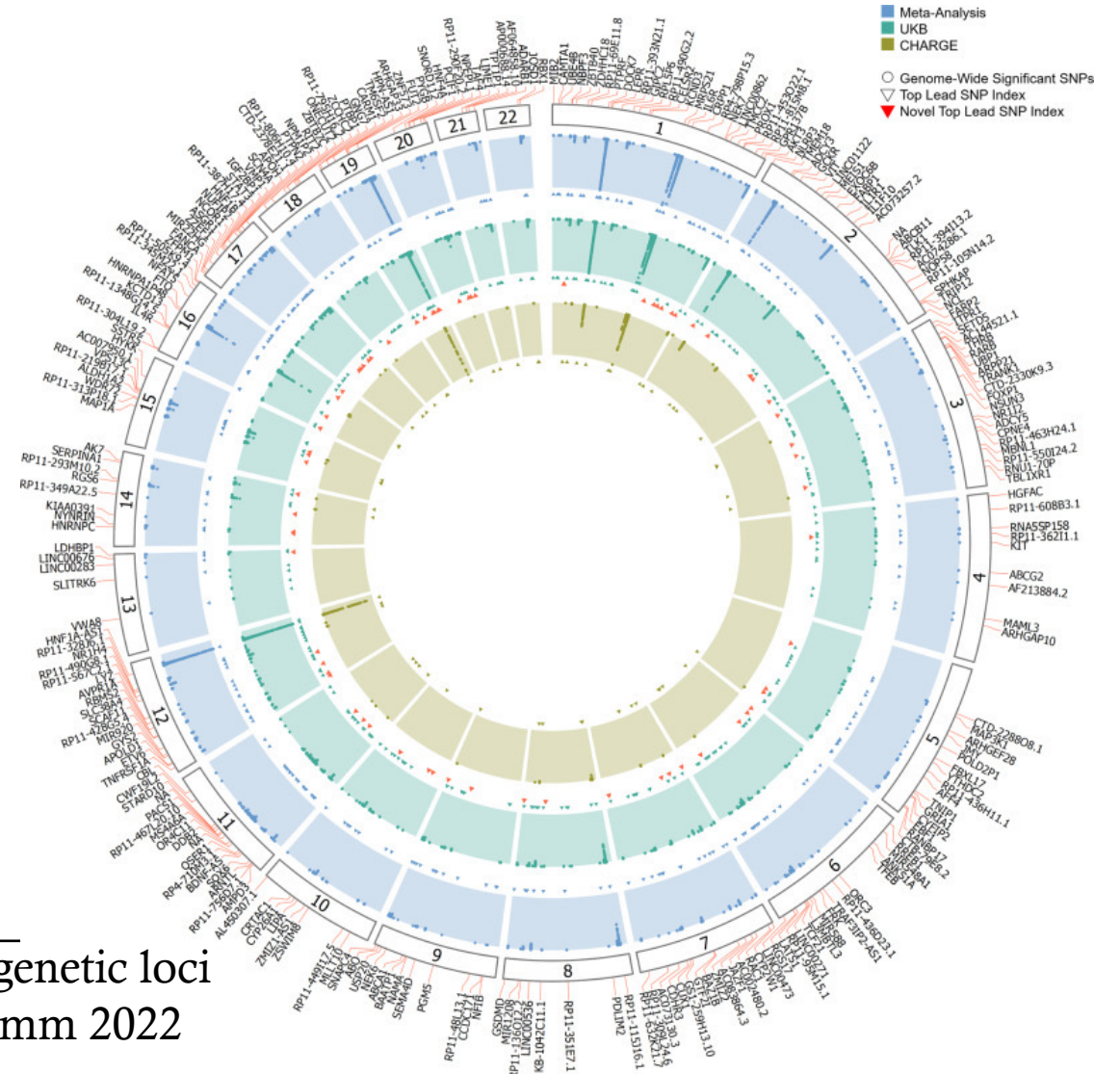
GENOME WIDE ASSOCTION STUDIES

- Improved genotyping technologies (cost and high throughput) & large consortia and biobanks → **genetic determinants of large number of traits (no reverse causality, no confounding)**
 - Most traits are polygenic – large number of common genetic variants of small effect
 - With higher sample size → large number of variants
-

GENETIC DETERMINANTS OF C-REACTIVE PROTEIN

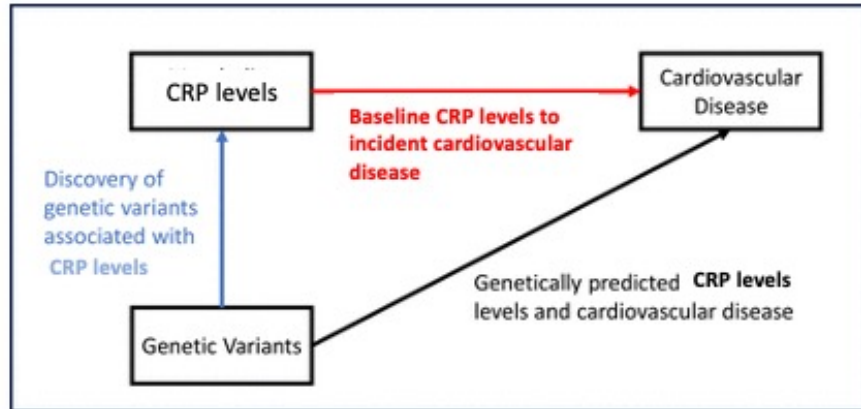


N~21,000, 5 genetic loci
Elliott et al. JAMA 2009

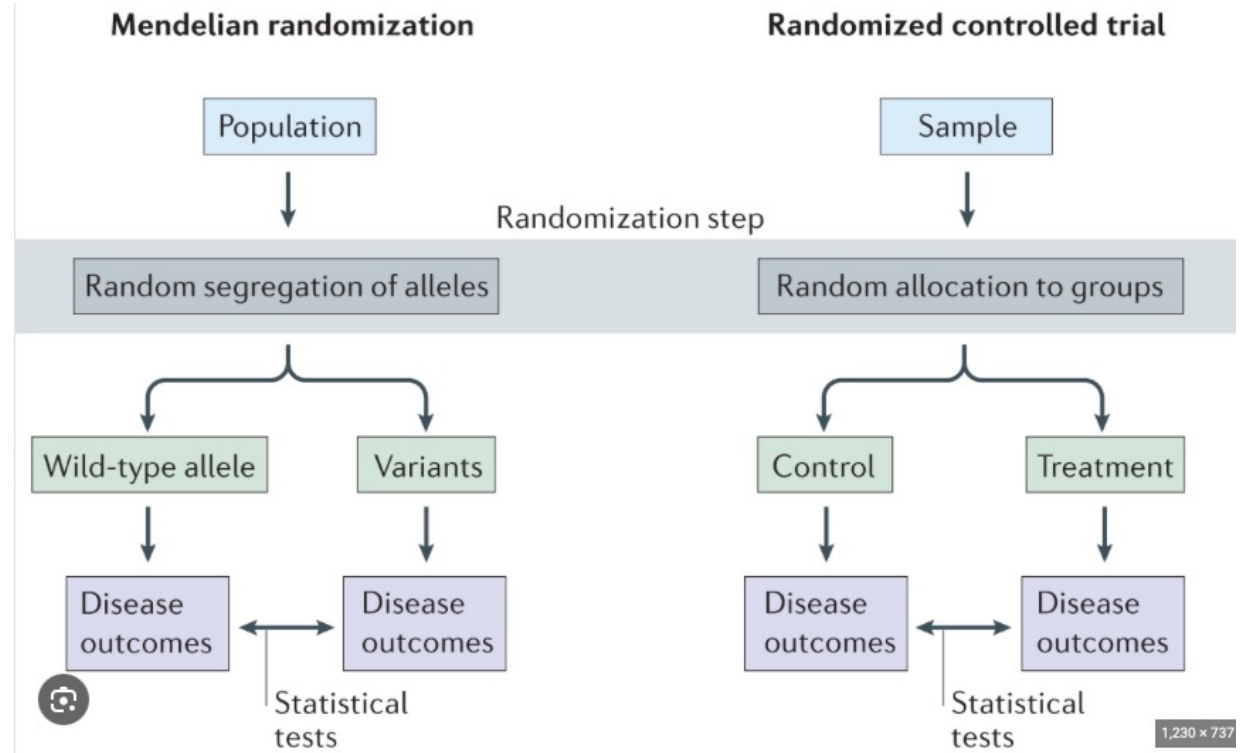


N~500,000, 266 genetic loci
Said et al Nat Comm 2022

MENDELIAN RANDOMIZATION



Instrumental variable analysis

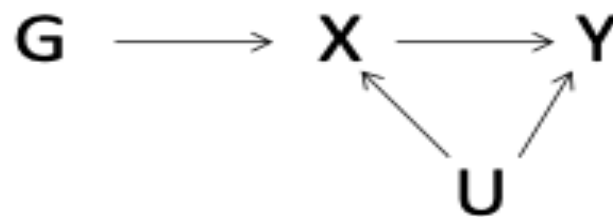


ATTRIBUTES OF MENDELIAN RANDOMIZATION

- 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
 - Genetic risk score of systolic blood pressure
- Because of random assortment of alleles, MR reduces bias due to confounding
- MR also largely avoids exposure measurement error and reverse causation bias

Assumptions of MR

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



Analogy to an RCT

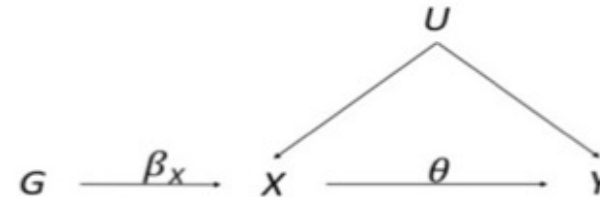


MR STUDY DESIGNS

- One sample MR
- In the traditional MR setting (one-sample MR), data on G , X and Y are available for all participants.

➤ “Wald” or “ratio” method (can accommodate one IV)

- ▶ β_X : The association of G with the risk factor X
- ▶ β_Y : The association of G with the outcome Y



If the exclusion restriction assumptions holds, the effect of G on Y can be decomposed into

$$G \xrightarrow{\beta_Y = \beta_X \theta} Y$$

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

$$\hat{\theta}_{ratio} = \frac{\beta_Y}{\beta_X}$$

MR STUDY DESIGNS

- One sample MR
- In case of multiple instruments
→ two stage least squares
methods

1. Stage: Predict the risk factor X based on the genotype G

- ▶ Linear regression of G on X

$$X = \beta_X G + \epsilon \quad (1)$$

- ▶ Predicted values \hat{X} based on genetic model

$$\hat{X} = \beta_X G \quad (2)$$

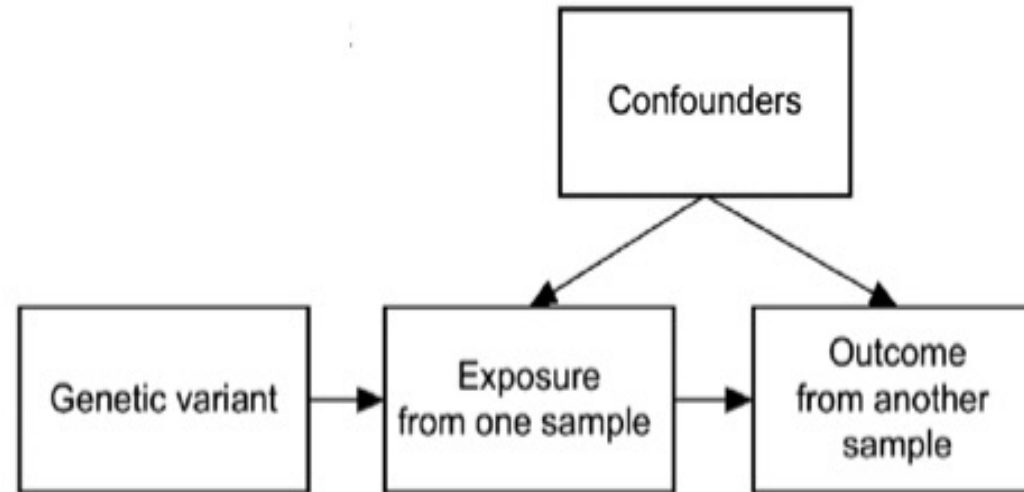
2. Stage: Linear regression of \hat{X} on Y

$$Y = \hat{\theta}_{2SLS} \hat{X} + \epsilon \quad (3)$$

$\hat{\theta}_{2SLS}$ is the two-stage least squares causal effect estimate for X on Y .

TWO SAMPLE MR

- Summary level data from GWAS

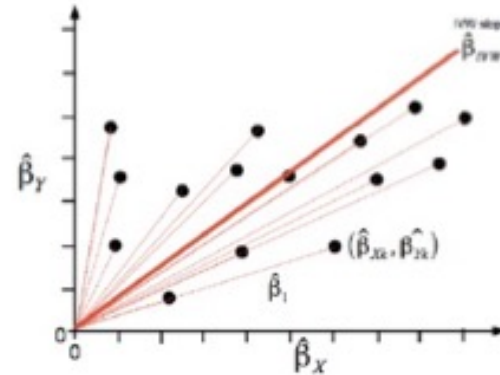


Zheng J, et al. *Curr Epidemiol Rep* 2017;4:330-345

INSTRUMENTAL VARIABLE (IV) ESTIMATORS IN TWO-SAMPLE MR

- Inverse variance weighting of individual IV effects (can accommodate multiple IVs)

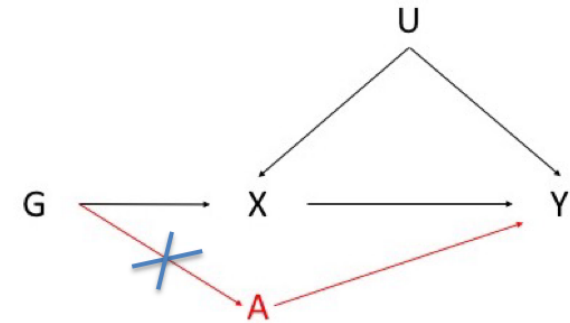
$$\hat{\theta}_{IVW} = \frac{\sum_{j=1}^n \hat{\theta}_j / \text{var}(\hat{\theta}_j)}{\sum_{j=1}^n 1 / \text{var}(\hat{\theta}_j)}$$



- Assumption: The genetic variants are independent and thus the ratio estimates are independent

METHODS TO ASSESS MR ASSUMPTIONS

- Horizontal pleiotropy is the most problematic assumption
 - Function of genetic variants is often unknown, particularly if these have been identified in GWAS
- We do have methods available to deal with pleiotropy



METHODS TO ASSESS MR ASSUMPTIONS

- **In one-sample MR**, very few methods available to detect (Sargan test) and correct for pleiotropy
- **In two-sample MR**, can detect pleiotropy with:
 - Cochran's Q test
 - I^2 statistic
 - Diagnostic plots (e.g. forest, funnel plots)
 - MR-Egger intercept test
- **In two-sample MR**, can correct for pleiotropy with:
 - MR-Egger slope
 - Weighted median
 - MR-PRESSO
 - Several more...

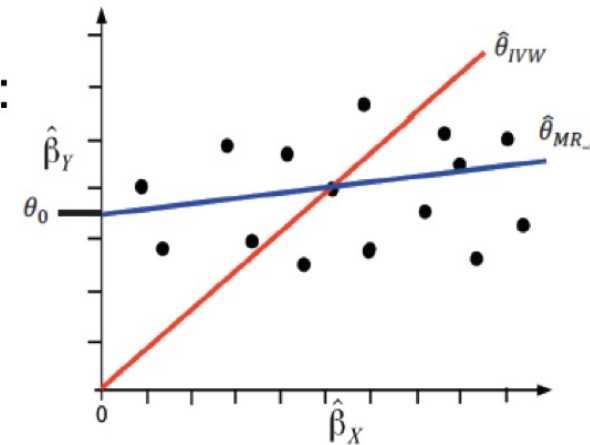
Glymour MM, et al. Am J Epidemiol 2012;175:332-9.
Bowden J, et al. Int J Epidemiol 2015;44:512-25.
Bowden J, et al. Genet Epidemiol 2016;40:304-14.
Greco et al. Stat Med 2017

MR-Egger regression

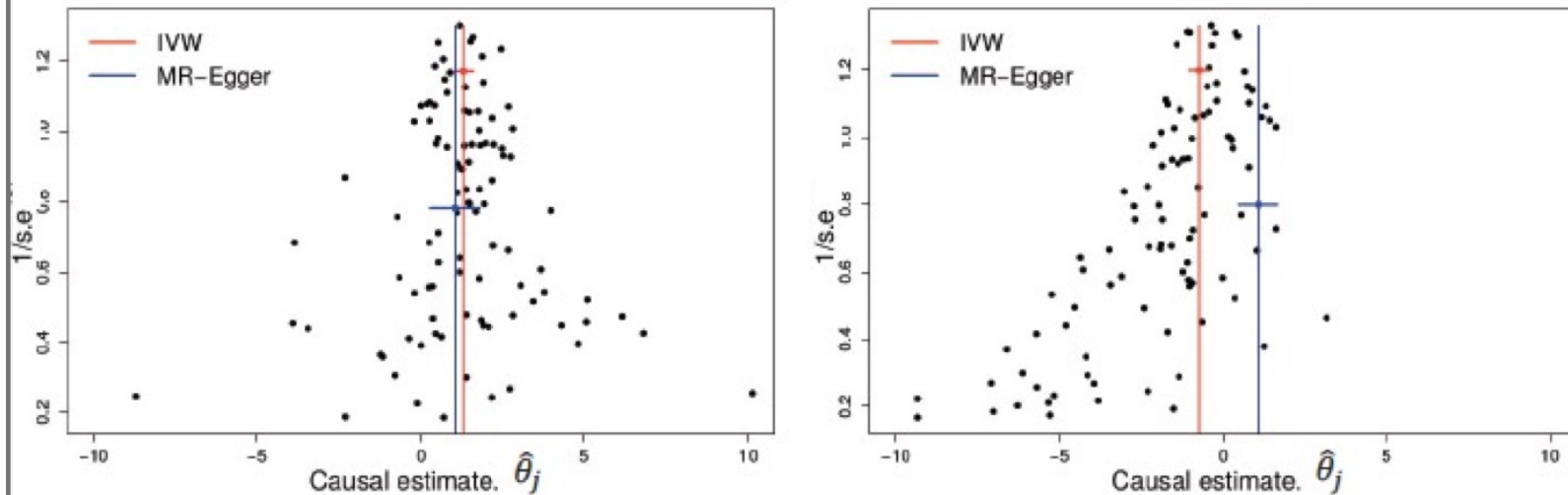
- MR-Egger relaxes the assumption that the average pleiotropic effect is zero (to allow for directional pleiotropy)
- This is achieved by introducing an intercept θ_0 in the regression model:

$$\hat{\beta}_{Yj} = \theta_0 + \theta \hat{\beta}_{Xj} + \varepsilon_j$$

- The intercept should be zero if all variants are valid IVs
- θ is the MR-Egger causal effect



MR-Egger regression: examples

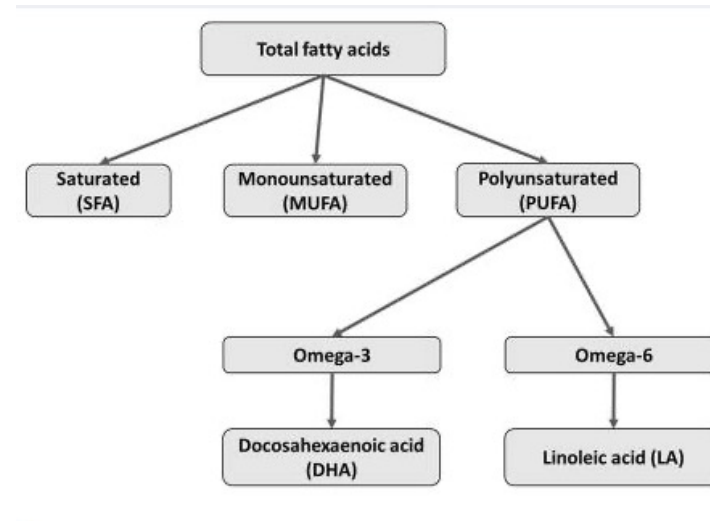
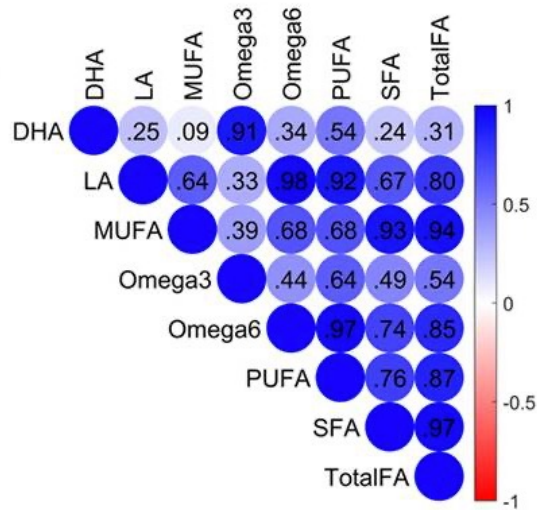


- Left: funnel appears symmetric, i.e. balanced pleiotropy appears valid, and IVW suitable
- Right: funnel is asymmetric, i.e. directional pleiotropy probable, and IVW not suitable

Median based estimation

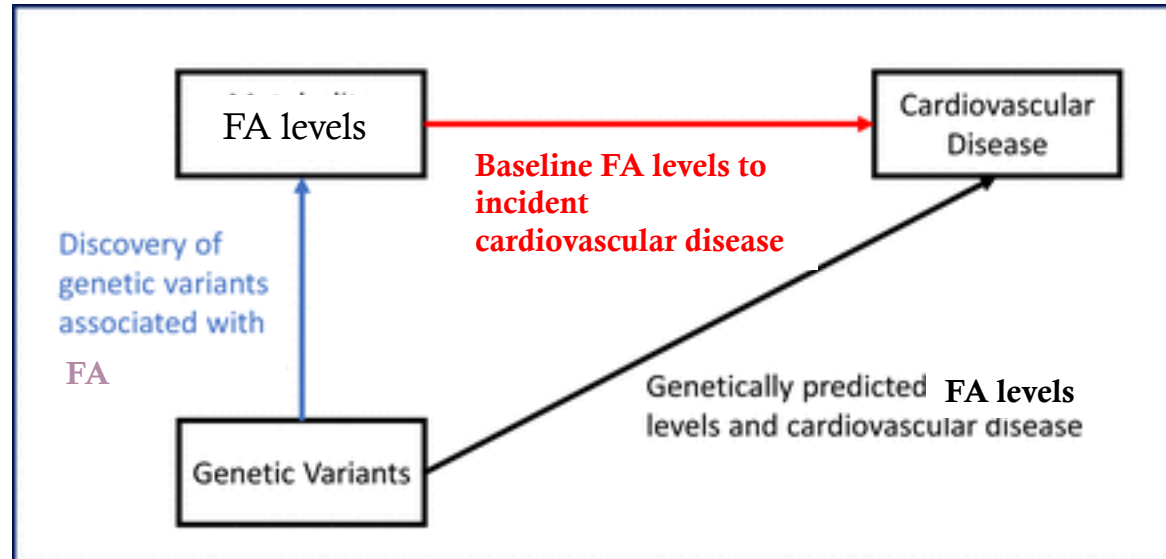
- Suppose that the majority of variants, i.e. >50%, are valid IVs
- In a large sample size, the variant-specific ratio estimates based on the valid instruments will all estimate the true causal effect
- So the median of the ratio estimates can be used as an estimate of θ
- The median estimate will be less influenced by outlying variants than the IVW estimate (which is a weighted mean of the variant-specific ratio estimates)
- No assumption necessary for invalid variants

PHENOME WIDE EFFECTS: FATTY ACIDS



- Epidemiological studies have investigated the influence of dietary fatty acids on several chronic diseases
- Evidence not always supported by randomised controlled trials (RCTs) on fatty acids supplementation.
- Human fatty acid metabolome partly reflects the fatty acid intake

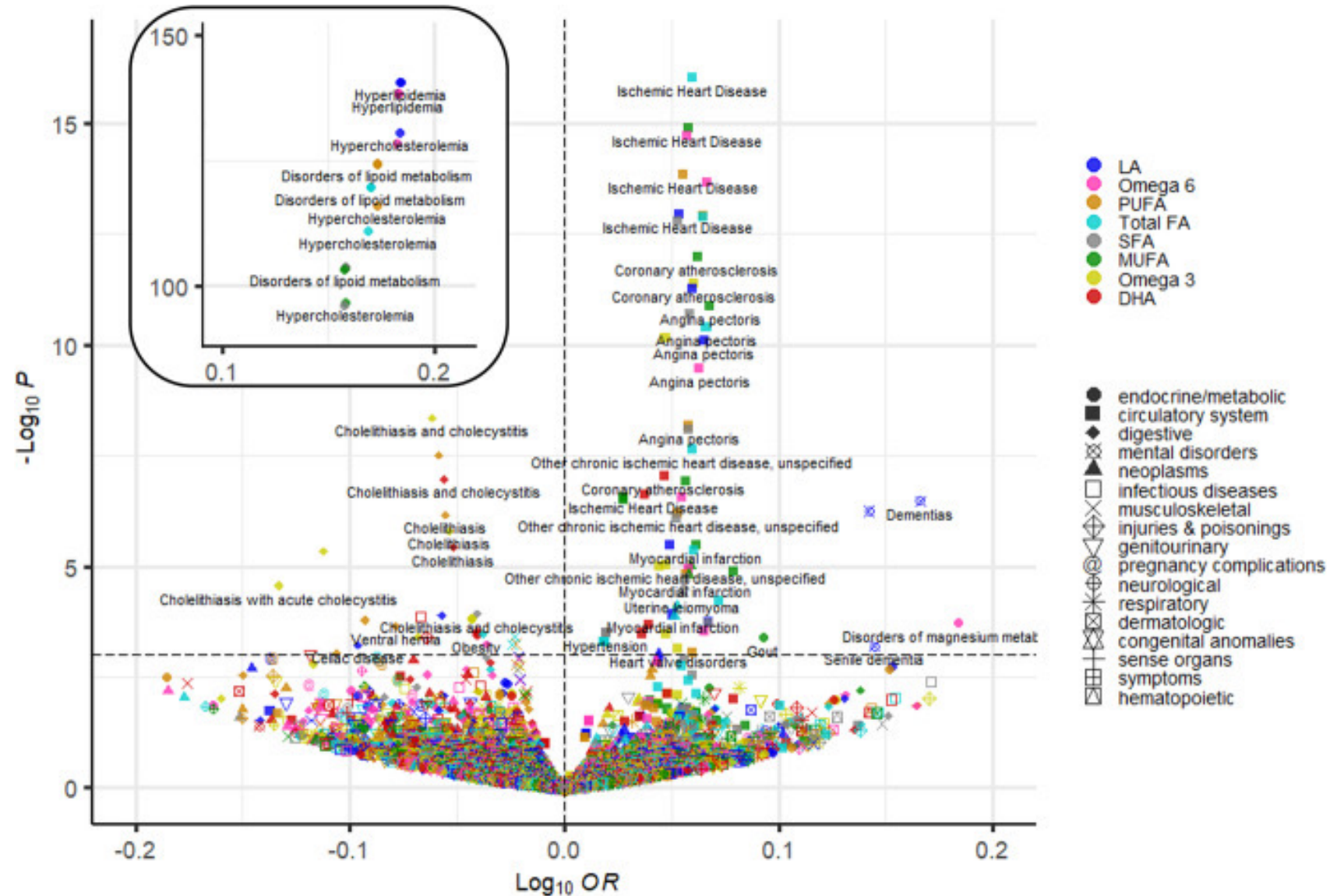
PHENOME WIDE EFFECTS: FATTY ACIDS



- Epidemiological studies have investigated the influence of dietary fatty acids on several chronic diseases
- Evidence not always supported by randomised controlled trials (RCTs) on fatty acids supplementation.

PHENOME WIDE EFFECTS: FATTY ACIDS

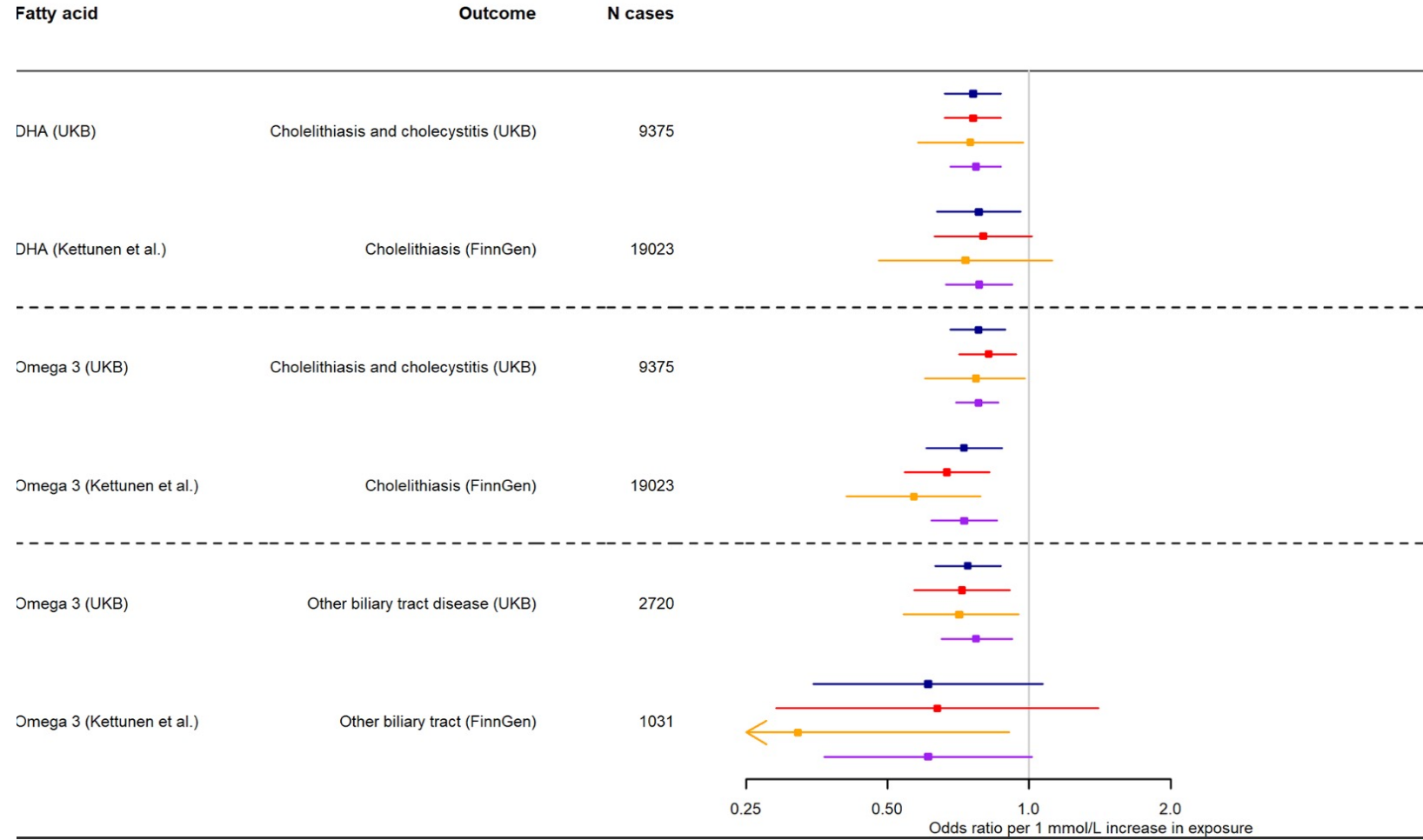
- Potentially protective effects of DHA and omega-3 on cholelithiasis and cholecystitis and obesity
- Supplementation of unsaturated fatty acids for cardiovascular disease prevention not supported by these data



Genetically predicted fatty acid levels

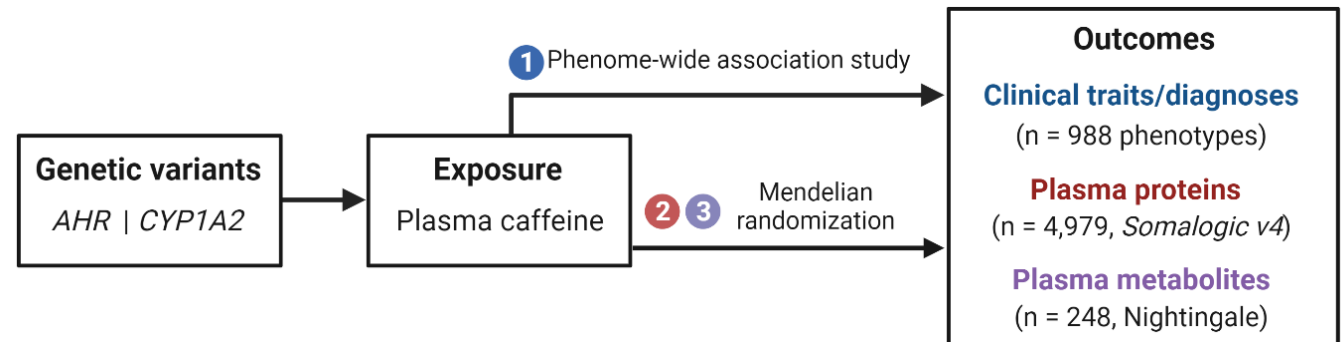
Cholelithiasis, cholecystitis and 'other biliary tract disease'

■ MR-IVW ■ MR-WM ■ MR-Egger ■ MR-PRESSO



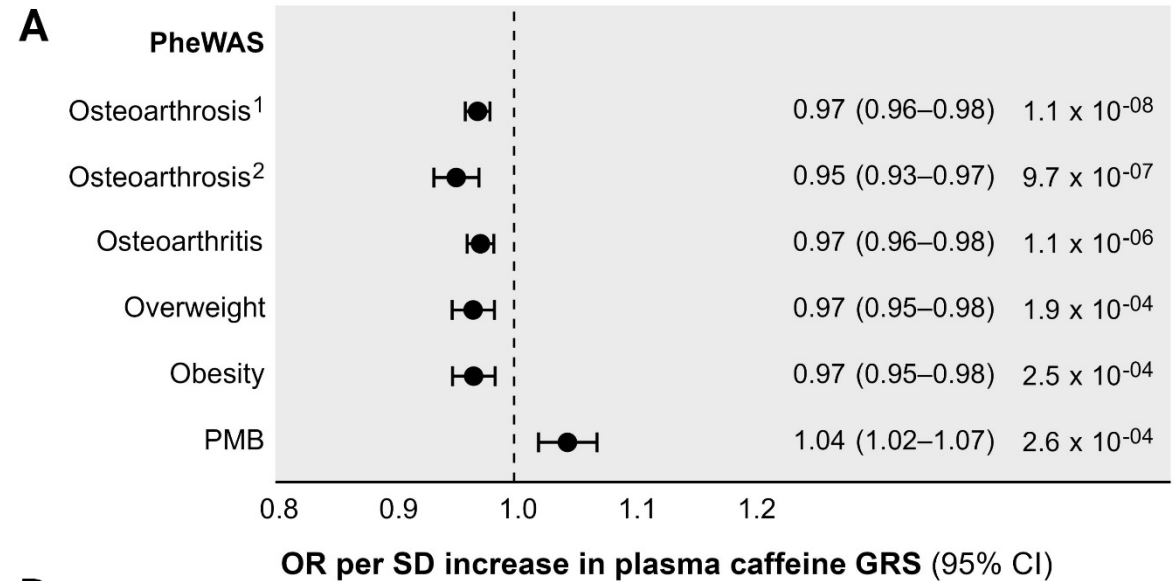
PHENOME-WIDE EFFECTS: COFFEE

- Caffeine is one of the most utilized drugs in the world, yet its clinical effects are not fully understood.
 - widespread availability of caffeine, effects on alertness, endurance, concentration, and productivity
- Observational consumption data limited by the inability to disentangle the effect of caffeine from co-occurring bioactive compounds in caffeinated foods and beverages
- Interindividual differences in caffeine metabolism (*CYP1A2*)



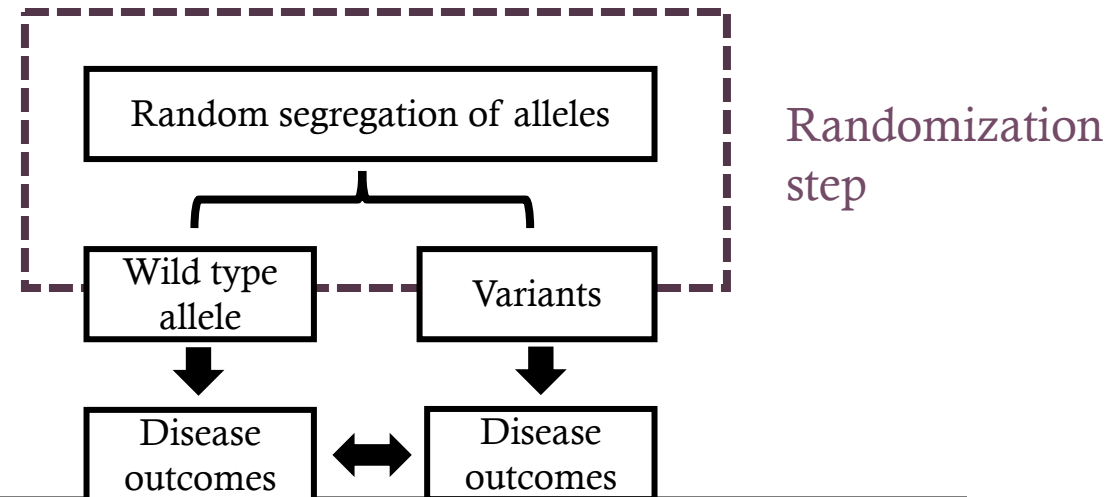
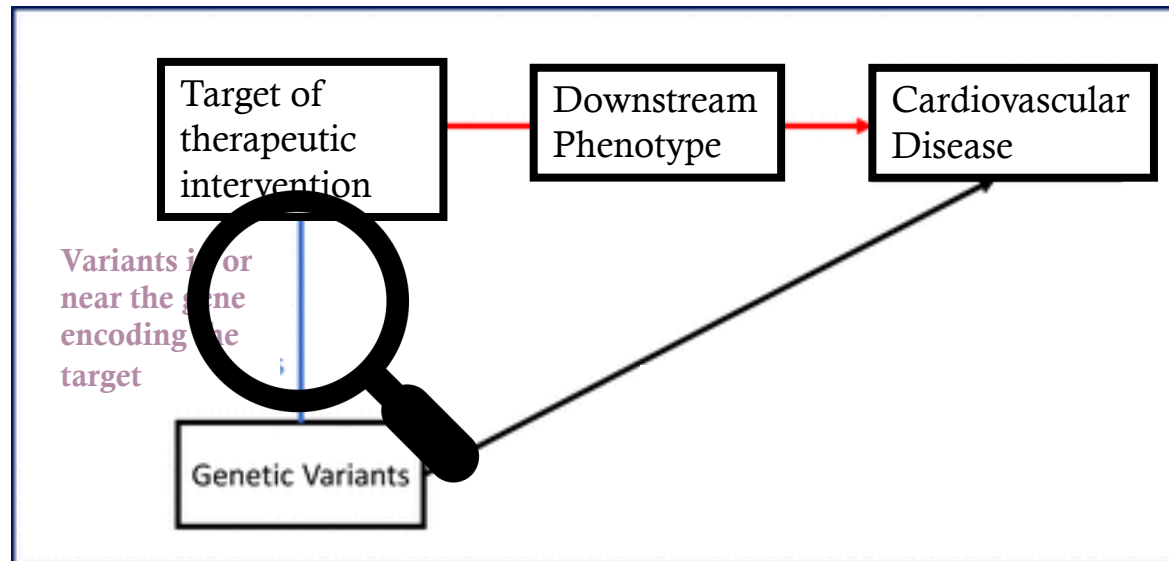
PHENOME-WIDE EFFECTS: COFFEE

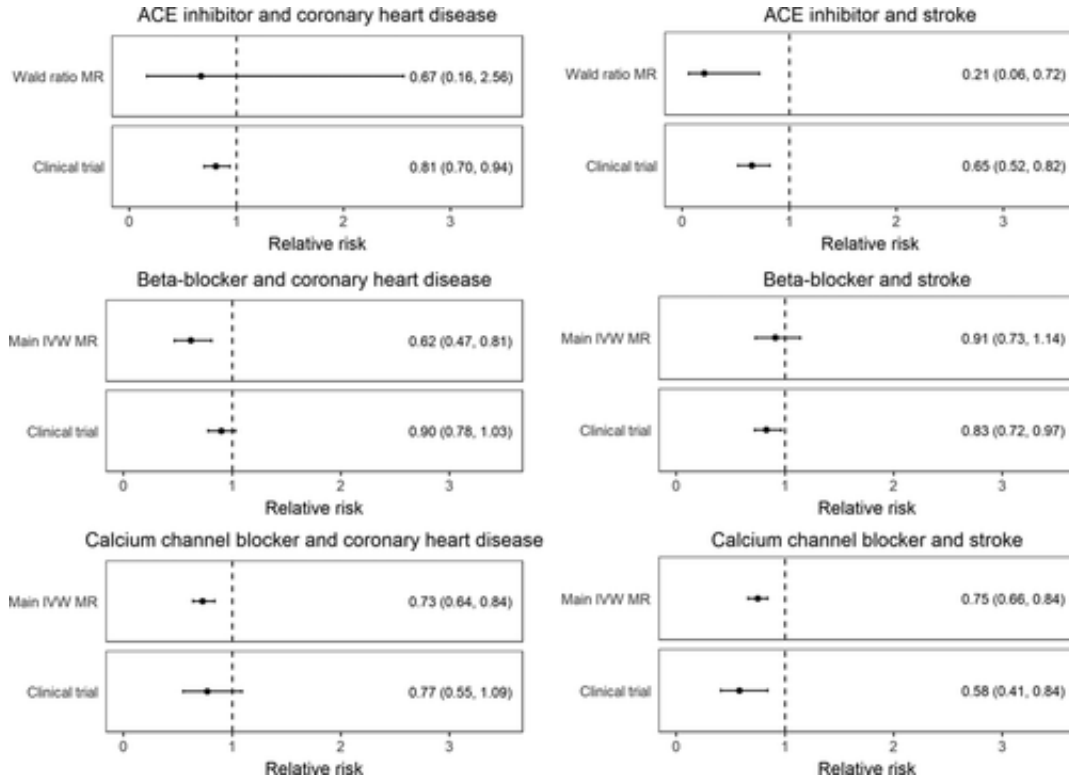
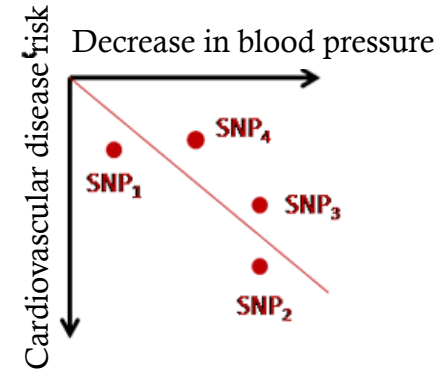
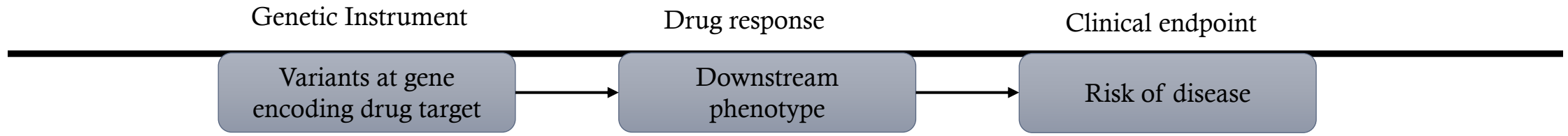
- Higher levels of genetically predicted circulating caffeine among caffeine consumers → lower risk of obesity and osteoarthritis.
- 1/3 of the protective effect of plasma caffeine on osteoarthritis risk mediated through lower bodyweight.
- Proteomic and metabolomic perturbations indicated lower chronic inflammation, improved lipid profiles, and altered protein and glycogen metabolism as potential biological mechanisms underlying these effects.



B

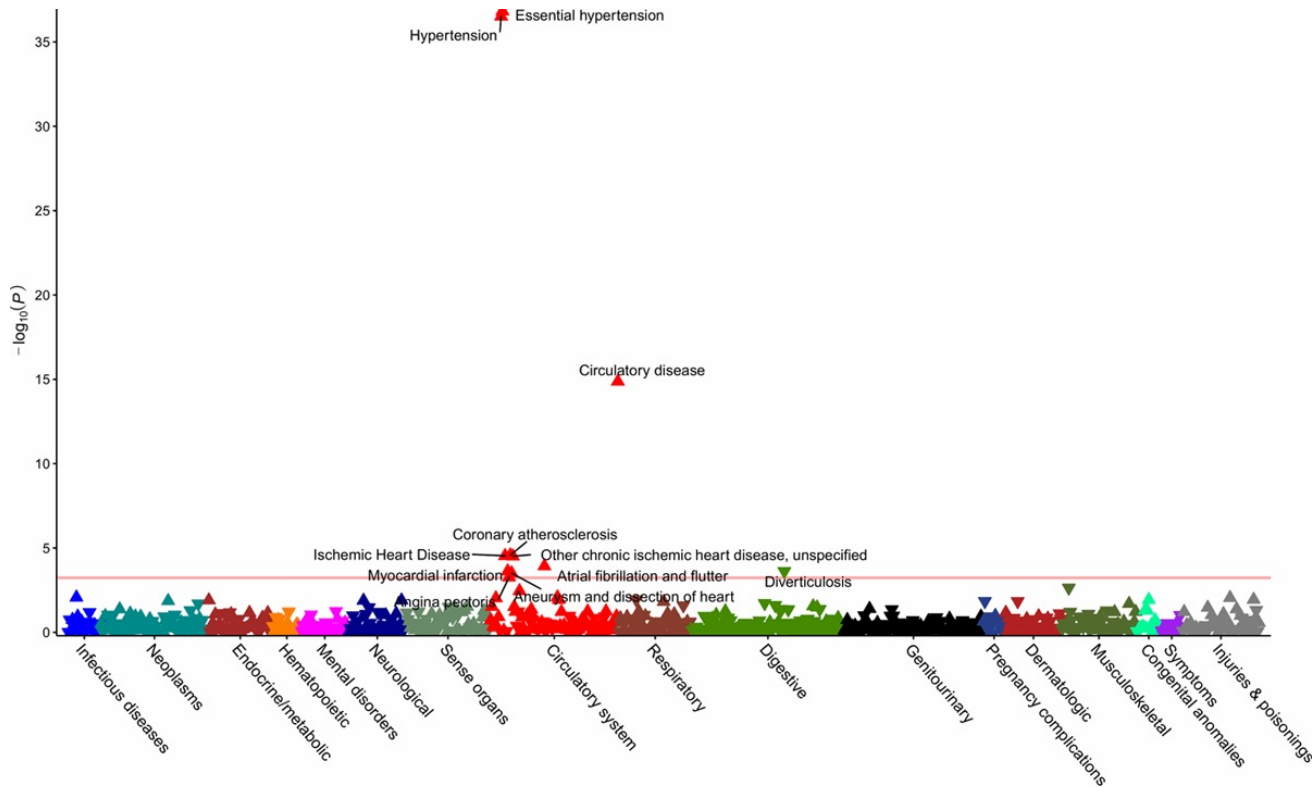
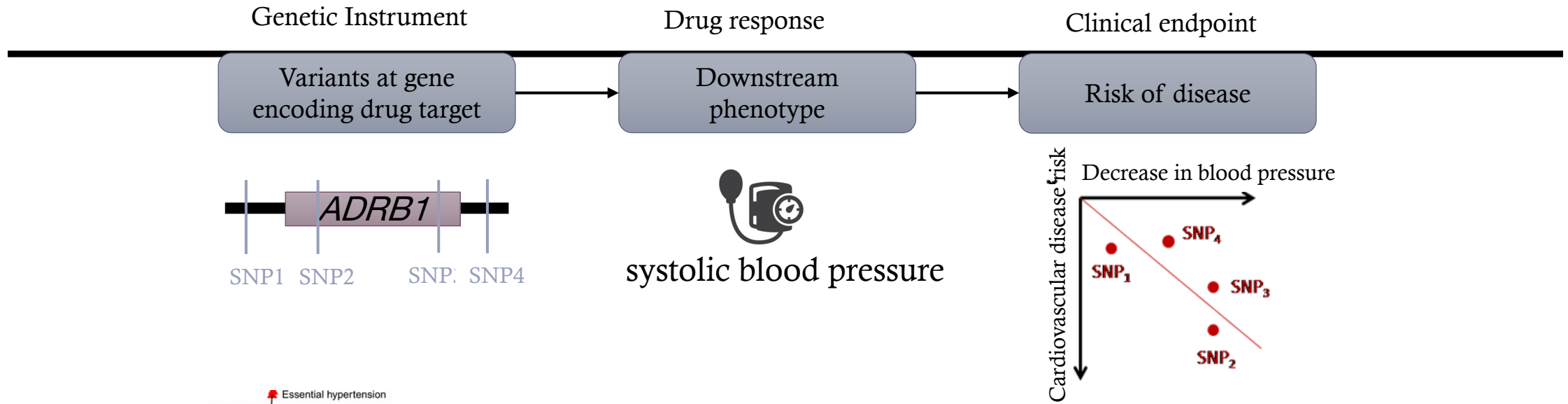
A STEP FURTHER : MENDELIAN RANDOMIZATION FOR TREATMENTS





MR estimates for the effect of genetically lower systolic blood pressure through the ACE inhibitor, β -blocker, and calcium channel blocker variants

Gill...Tzoulaki *Circulation*. 2019;140:270–279



MR estimates for the effect of genetically lower systolic blood pressure through the ACE inhibitor, β -blocker, and calcium channel blocker variants


Gill...Tzoulaki Circulation. 2019;140:270–279

RESEARCH ARTICLE

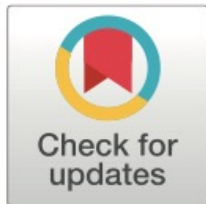
Long-term cost-effectiveness of interventions for obesity: A mendelian randomisation study

Sean Harrison ^{1,2*}, Padraig Dixon ^{1,2}, Hayley E. Jones ², Alisha R. Davies³, Laura D. Howe ^{1,2}, Neil M. Davies ^{1,2,4}

1 MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **2** Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, **3** Research and Evaluation Division, Public Health Wales NHS Trust, Cardiff, United Kingdom, **4** K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

 These authors contributed equally to this work.

* sean.harrison@bristol.ac.uk



Abstract

CONCLUSIONS (MR)

- 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)
 - Exponential use in the literature – a plethora of new MR methodologies (clustering, non-linear effects, multivariate etc.)
 - Assumptions needs to be thoroughly checked, bias still exist (eg survival bias), not causality but evidence for causality.
-

EMULATE RANDOMIZED CLINICAL TRIALS USING ELECTRONIC HEALTH DATA

- Availability of electronic health records has increased the potential for conducting analyses for different treatments in real word settings
 - comparison of outcomes under different courses of action
 - Most important challenges include
 - Confounding by indication
 - Missing data not at random
 - Time zero and treatment assignment
-

EMULATE TARGET TRIALS (STEP 1)

- Causal question in a form of a RCT protocol (treatment assignment, end and start date, contrasts)
 - Treatment assignment at time zero without using later information (intention to treat principle)



Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia

Received: 13 June 2021

Accepted: 21 November 2022

Published online: 10 December 2022

Check for updates

Marie-Laure Charpignon^{1,16}, Bella Vakulenko-Lagun^{2,16}, Bang Zheng^{3,16}, Colin Magdamo⁴, Bowen Su⁵, Kyle Evans^{4,6}, Steve Rodriguez^{4,6}, Artem Sokolov⁶, Sarah Boswell⁶, Yi-Han Sheu⁷, Melek Somai⁸, Lefkos Middleton^{3,9}, Bradley T. Hyman⁴, Rebecca A. Betensky¹⁰, Stan N. Finkelstein^{1,11}, Roy E. Welsch^{1,12}, Ioanna Tzoulaki^{5,13,14,17}, Deborah Blacker^{7,15,17}, Sudeshna Das^{4,17} & Mark W. Albers^{4,6,17} ✉

Table 1 | Specification and emulation of a target trial of antidiabetic drug metformin vs. sulfonylureas on the risk of death and dementia, using observational data from Electronic Health Records of the US RPDR and the UK CPRD

Target trial specification	Emulation (US RPDR)	Emulation (UK CPRD)
Eligibility criteria		
Age ≥ 50	Same	
No hypoglycemics	No recorded prior exposure to any hypoglycemic agents	
No MCI*, dementia, or prescription of dementia drugs; normal cognitive testing	No recorded diagnosis of dementia or MCI*, or use of dementia-specific drugs (see Extended Data Tables 10–11)	No recorded diagnosis of dementia (MCI* diagnoses not available in CPRD) or use of dementia-specific drugs (see Extended Data Tables 12–13)
No chronic kidney disease (metformin contraindication)	No ICD*-9/10 code for chronic kidney disease or eGFR* <45 (Extended Data Table 1)	No diagnosis of chronic kidney disease at or prior to baseline (Extended Data Table 2)
Trial with 1-year run in period conducted for a specified duration with history obtained at baseline and ongoing monitoring of outcomes	<ul style="list-style-type: none"> • PCP* within Mass General Brigham Health Care system EHR* system • At least one visit during the 18 months preceding baseline • At least 1 year of follow-up • No dementia or death in first year (1 year washout period) 	<ul style="list-style-type: none"> • At least 1-year registration in CPRD practices before the first prescription • At least 1 year of follow-up • No dementia or death in first year (1-year washout period)
Treatment strategies		
Treatment arm: metformin monotherapy Control arm: sulfonylurea monotherapy	Initiation of metformin or sulfonylurea from 1/2007-9/2017 (see Extended Data Fig. 8 for the number of new prescriptions per year)	Initiation of metformin or sulfonylurea from 1/2001-5/2017, with ≥2 monotherapy prescriptions for first 12 months (see Extended Data Fig. 9 for the number of new prescriptions per year)
Treatment assignment		
Double-blind, randomized treatment assignment	Emulated randomization by balancing baseline confounders using IPTW* for treatment choice	
Outcomes		
Diagnosis of MCI* or dementia	Diagnosis of MCI/Dementia by: ICD*-9/10 codes (Extended Data Table 10) OR at least one dementia-specific drug prescription (Extended Data Table 11)	Diagnosis of dementia by: Medcodes in CPRD or ICD*-9/10 codes in linked HES* or ONS* database (Extended Data Table 12) OR at least one dementia-specific drug prescription (Extended Data Table 13)
Time to death	Time to death recorded in EHR*	
Follow-up		
From baseline and ends at dementia onset, death, lost to follow-up, or end of study	From the date of initial prescription of drug until the date of dementia incidence, death, last encounter date, 9/2018 (US RPDR) or 5/2018 (UK CPRD), whichever occurred first	
Causal contrast		
Intention-to-treat effect	Observational analog of intention-to-treat effect	
Statistical analysis		
Intention-to-treat analysis of primary outcomes (dementia and death) using Cox PH	Intention-to-treat analysis using Cox Proportional Hazards (PH) regression model and a competing risks framework accounting for death prior to dementia Subgroup analyses by age, sex, and BMI* level at baseline	

* BMI body mass index, eGFR estimated glomerular filtration rate, EHR Electronic Health Records, HES Hospital Episode Statistics, ICD International Classification of Diseases, IPTW inverse propensity score of treatment weighting, MCI mild cognitive impairment, ONS Office for National Statistics, PCP primary care physician.

EMULATE TARGET TRIAL (STEP 2)

- Same analysis as the corresponding target trial, but with adjustment for baseline confounders to emulate random treatment assignment
 - Multivariable regression (including confounders as covariates) and inverse probability of treatment
 - Inverse probability weighting (or propensity score)
 - Instrumental variables (IV)
 - Regression discontinuity
 - Interrupted time series (ITS)
 - Difference in differences
- Methods depend on data availability

Table 2 Definitions of different types of treatment effect

Effect	Potential outcome notation	Description
Average treatment effect (ATE)	$E(\gamma^{a=1} - \gamma^{a=0})$	The difference between the average outcome when everyone is exposed, and the average outcome when nobody is.
Average treatment effect in the treated (ATT)	$E(\gamma^{a=1} - \gamma^{a=0} A = 1)$	The ATE in the subpopulation of individuals who were actually exposed.
Average treatment effect in the untreated (ATU/ATUT)	$E(\gamma^{a=1} - \gamma^{a=0} A = 0)$	The ATE in the subpopulation of individuals who were actually unexposed.
Intention-to-treat effect (ITT)	$E(\gamma^{z=1} - \gamma^{z=0})$	Average effect of being assigned to (but not necessarily receiving) the exposure.
Complier average causal effect (CACE) or local average treatment effect (LATE)	$E(\gamma^{a=1} - \gamma^{a=0} \begin{matrix} A^{z=0} = 0, \\ A^{z=1} = 1 \end{matrix})$	The ATE among the 'compliers', that is, the subpopulation whose exposure status was affected by the assignment mechanism.

A denotes actual exposure status ($a=1$ for exposed, $a=0$ for unexposed). Z denotes assignment to the exposure, which may or may not have been adhered to ($z=1$ for assignment to the exposure, $z=0$ for assignment away from the exposure).



Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia

Received: 13 June 2021

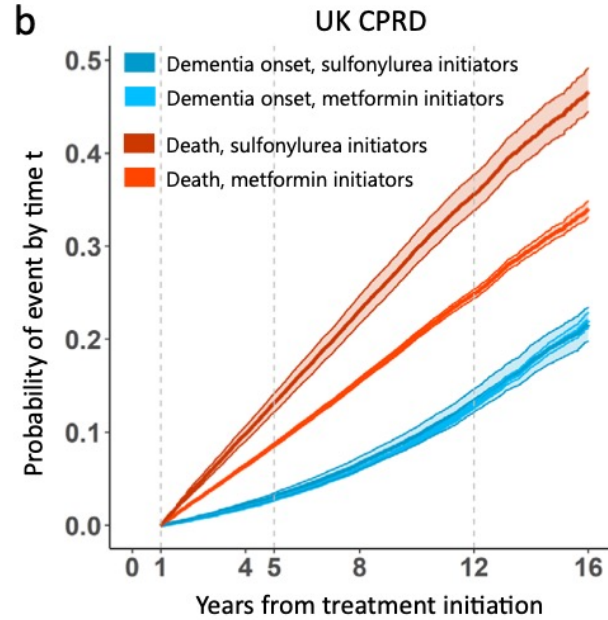
Accepted: 21 November 2022

Published online: 10 December 2022

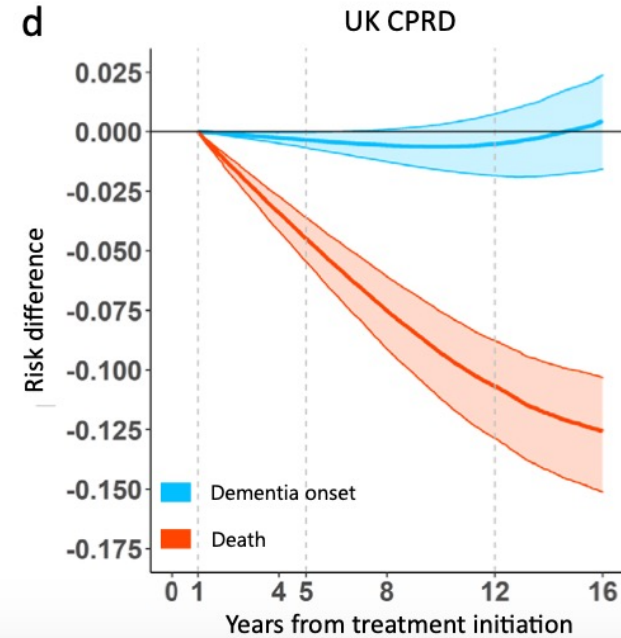
Check for updates

Marie-Laure Charpignon^{1,16}, Bella Vakulenko-Lagun^{2,16}, Bang Zheng^{3,16}, Colin Magdamo⁴, Bowen Su⁵, Kyle Evans^{4,6}, Steve Rodriguez^{4,6}, Artem Sokolov⁶, Sarah Boswell⁶, Yi-Han Sheu⁷, Melek Somai⁸, Lefkos Middleton^{3,9}, Bradley T. Hyman⁴, Rebecca A. Betensky¹⁰, Stan N. Finkelstein^{1,11}, Roy E. Welsch^{1,12}, Ioanna Tzoulaki^{5,13,14,17}, Deborah Blacker^{7,15,17}, Sudeshna Das^{4,17} & Mark W. Albers^{4,6,17}

b



d



CONCLUSIONS (TARGET TRIALS)

- RCTs are top of the evidence hierarchy
 - Triangulation of evidence through other methods in observational data can prioritise and inform the RCTs most likely to be successful
 - Credible sources of real-world evidence to support regulatory decisions in the absence of RCTs
-

THANK YOU

University of Ioannina
MR methodology slides:
Kostantinos Tsilidis
Christos Chalitsios
Fotios Koskeridis

IIBEAA
Maria Manou
Giorgos Ntritsos



Imperial College London
Abbas Dehghan
Loukas Zagkos
Dipender Gill



ΕΛΙΔΕΚ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας



National Institutes of Health
Turning Discovery Into Health

ALZHEIMER'S  ASSOCIATION®