

Mendelian Randomisation and Causal Inference in Observational Epidemiology

Nuala Sheehan

Department of Health Sciences

UNIVERSITY of LEICESTER

MRC Collaborative Project Grant G0601625

Vanessa Didelez, Sha Meng, Roger Harbord, Jonathan Sterne, Debbie Lawlor,
Frank Windmeijer, John Thompson, Paul Clarke, Tom Palmer, Paul Burton,
George Davey Smith

Department of Epidemiology and Public Health
Imperial College London March 2009

Causal Inference

Always inference about **interventions** — whatever framework is adopted

Epidemiology is concerned with **finding** and **assessing the size** of the effect of **modifiable** risk factors on diseases so that (public health) interventions can be informed — **always about causality!**

Examples for interventions:

Adding folic acid to flour

Banning smoking in pubs

Dietary advice: “5 portions of fruit & vegetables a day” etc.

Problems

- “Association \neq causation” i.e. might find an association but intervention turns out to be useless,
 - Randomised/controlled experiments are not always possible in Epidemiology \Rightarrow sometimes unavoidable and often desirable to have to work with **observational** data.
 - observational findings often not reproduced in randomised trials when these are possible:
 - reverse causation
 - confounding
 - others?
- \Rightarrow **Instrumental variable** methods as alternative?

Interventions

We need **notation** to formally distinguish between association and causation.

Intervention: **setting** X to a value x denoted by $do(X = x)$.

$p(y|do(X = x))$ not necessarily the same as $p(y|X = x)$.

- $p(y|do(X = x))$ depends on x only if X is causal for Y
⇒ observed in a randomised study.
- $p(y|X = x)$ will also depend on x when there is confounding or reverse causation
⇒ observed in an observational study.

Causal Effect

Some contrast in the effects of different interventions on X on the outcome Y i.e. compare $p(y|do(X = x_1))$ with $p(y|do(X = x_2))$.

Average Causal Effect: $ACE(x_1, x_2) = E(Y|do(x_1)) - E(Y|do(x_2))$

or analogously **Causal Risk Ratio** CRR

$$CRR(x_1, x_2) = \frac{p(Y = 1|do(X = x_1))}{p(Y = 1|do(X = x_2))}$$

or **Causal Odds Ratio** COR .

$$COR(x_1, x_2) = \frac{p(Y = 1|do(X = x_1))p(Y = 0|do(X = x_2))}{p(Y = 0|do(X = x_1))p(Y = 1|do(X = x_2))}$$

Alternative: causal effect on particular subgroups.

Identifiability

We are often interested in estimating the causal effect consistently from observational data.

Mathematically, the causal effect is **identifiable** if we can re-express it **without** $do(X)$ notation using only the distribution of **observable** variables.

If a **sufficient** set C of confounders is measured, the causal effect is identified using

$$p(y|do(X = x)) = \sum_c p(y|X = x, C = c)p(c)$$

→ usual adjustment for **known** confounders.

This is not always possible in the observational regime \Rightarrow need to deal with confounding by other means, e.g. **instrumental variables (IV)**

Identifiability using Instrumental Variables

There are different types of assumption required:

(in)dependencies	}	allow testing for causal effect
structural		
parametric form		for estimation

Core Conditions

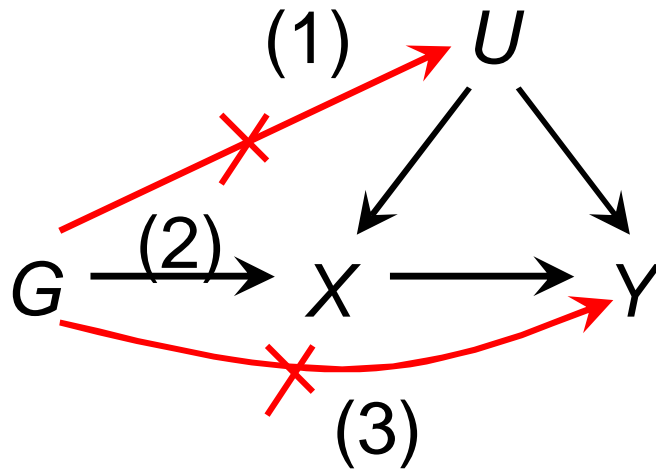
For the effect of X (phenotype/exposure) on Y (disease), a variable G (genotype) is an **instrument** if variable(s) U (unobserved confounders) exist such that

1. $G \perp\!\!\!\perp U$: G and U are **independent**
2. $G \not\perp\!\!\!\perp X$: G and X are (strongly) **associated**
3. $G \perp\!\!\!\perp Y \mid (X, U)$: G and Y conditionally independent given X **and** U .

G is **only** associated with Y **via** its association with X ,

Cannot forget about U!

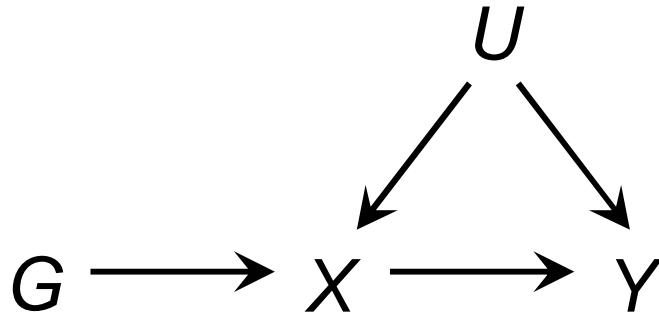
Core Conditions — Graphically



NOTE: In particular, these conditions do not imply that $G \perp\!\!\!\perp Y \mid X$ or $G \perp\!\!\!\perp Y$.

NOTE: Assumptions 1 and 3 cannot be tested from data as U is typically not known/measured \Rightarrow justification must be based on background/subject matter knowledge.

Core Conditions — Graphically



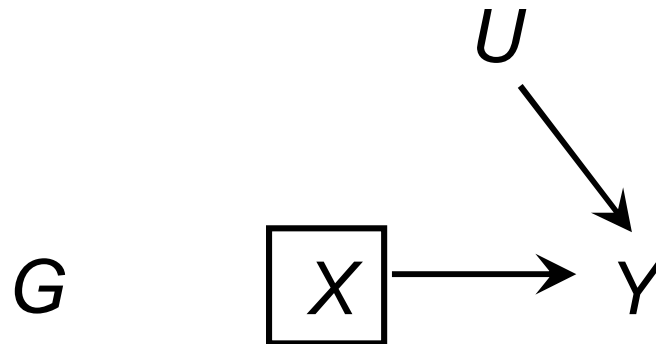
Equivalent to factorisation

$$p(y|x, u)p(x|u, g)p(u)p(g).$$

Structural assumptions:

$p(y|x, u)$, $p(g)$ and $p(u)$ are not changed by intervention in X ,
i.e. when conditioning on $do(X)$.

Core Conditions — Graphically



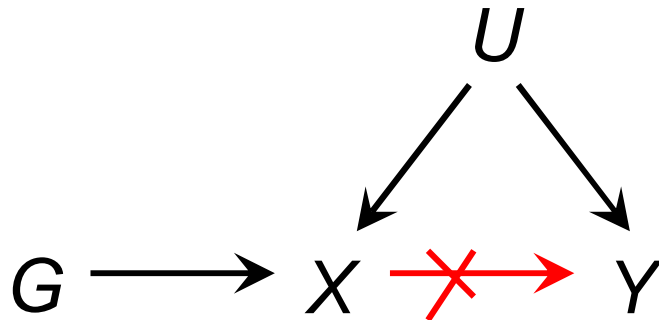
With **structural** assumption: under intervention in X

$$p(y, u, g | do(X = x^*)) = p(y | x^*, u) p(u) p(g)$$

Graphically, the intervention corresponds to removing all arrows leading into X .

Testing for Causal Effect

No causal effect “ \Leftrightarrow ” G independent of Y



From factorisation $p(y, x, u, g) = p(y|u)p(x|u, g)p(u)p(g)$

$$\Rightarrow p(y, g) = \sum_{x, u} p(y|u)p(x|u, g)p(u)p(g) = p(y)p(g).$$

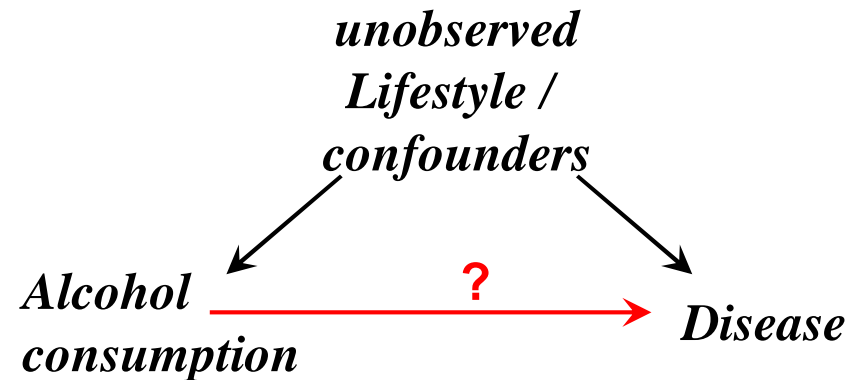
So a test for association between G and Y can be taken as a test for a causal effect of X on Y ([Katan 1986](#))

Mendelian Randomisation

- Consider **risk factors** that are modifiable **behaviours** or **phenotypes** known to be caused by, or strongly related to, certain **genotypes**;
- **Mendel's Second Law** (law of assortment): genotypes can reasonably be assumed to be independent of life style etc.
 - ⇒ Independent of typical confounding factors;
 - ⇒ kind of 'randomised';
- Genes are determined before birth, no reverse causation possible;
- **Conjecture: if and only if** phenotype is causal for disease should find an association between genotype and disease!
- This is known in statistics and econometrics as an **Instrumental Variable (IV)** method — here the genotype is the instrument.

Katan (1986) letter to *Lancet*, Davey Smith & Ebrahim (2003), Lawlor et al. (2008), Greenland (2000), Hernán & Robins (2006), Didelez & Sheehan (2007)

Example: Alcohol Consumption



Chen et al. (2008)

Alcohol consumption has been found in observational studies to have positive 'effects' (coronary heart disease) as well as negative 'effects' (liver cirrhosis, some cancers, mental health problems).

But also strongly associated with all kinds of confounders (lifestyle etc.), as well as subject to self-report bias. Hence doubts in causal meaning of above 'effects'.

Example: Alcohol Consumption

Genetic Instrumental Variable?

Genotype: ALDH2 determines blood acetaldehyde, the principal metabolite for alcohol.

Two alleles/variants: wildtype *1 and “null” variant *2.

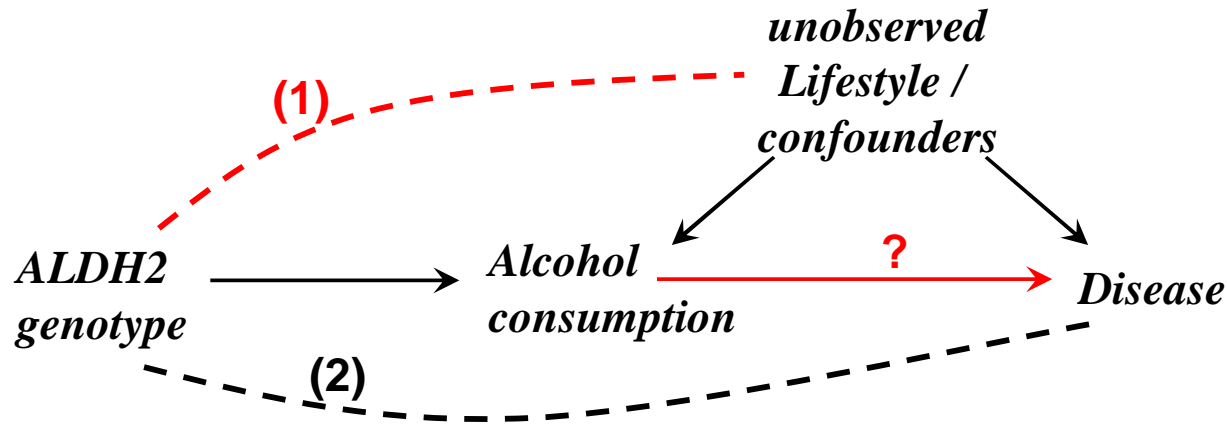
*2*2 homozygous individuals suffer facial flushing, nausea, drowsiness and headache after alcohol consumption.

⇒ *2*2 homozygotes have low alcohol consumption *regardless* of their other lifestyle behaviours

i.e. the gene can be taken as a proxy for alcohol intake.

IV-Idea: check if these individuals have a reduced risk for “alcohol-related” health problems!

Example: Alcohol Consumption

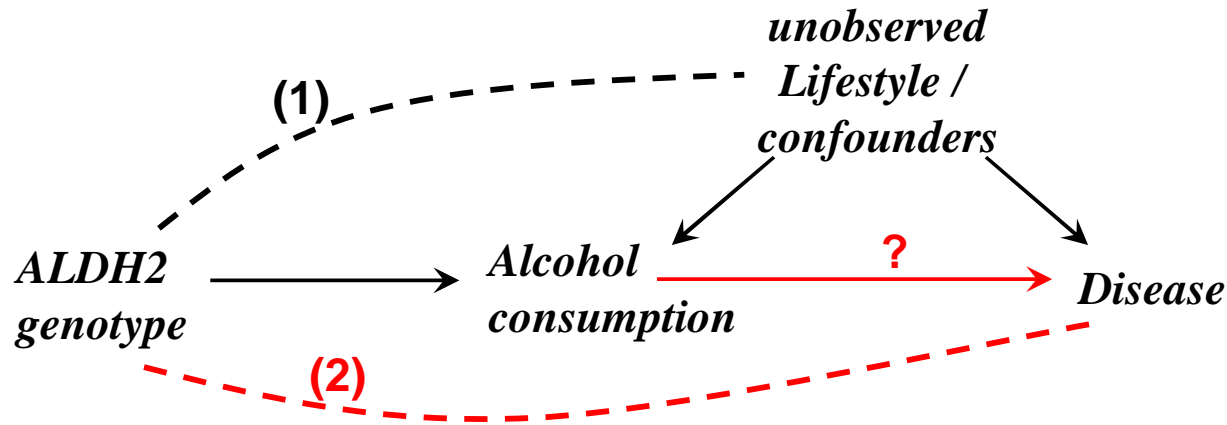


Note 1: due to random allocation of genes at conception, can be fairly confident that genotype is not associated with unobserved confounders.

Further evidence: in extensive studies no evidence for association with *observed* confounders, e.g. age, smoking, BMI, cholesterol.

(see also Davey Smith et al. 2007)

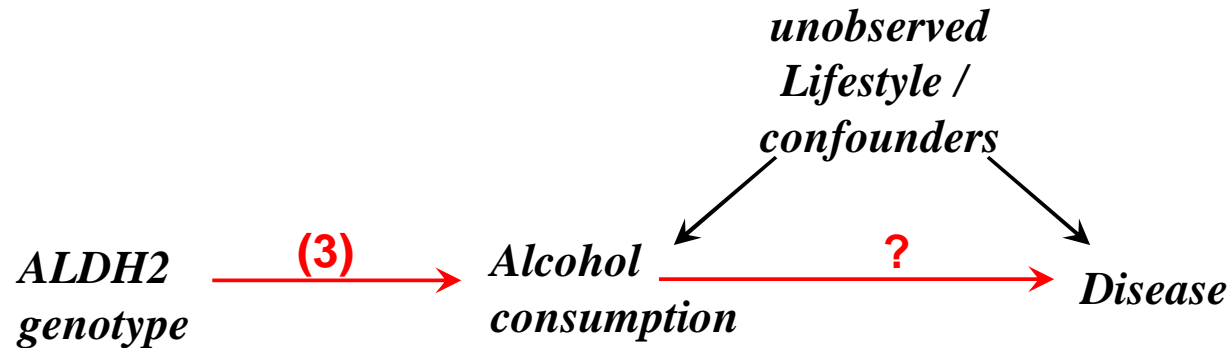
Example: Alcohol Consumption



Note 2: due to known ‘functionality’ of ALDH2 gene, we can exclude that it affects the typical diseases considered by *another* route than through alcohol consumption.

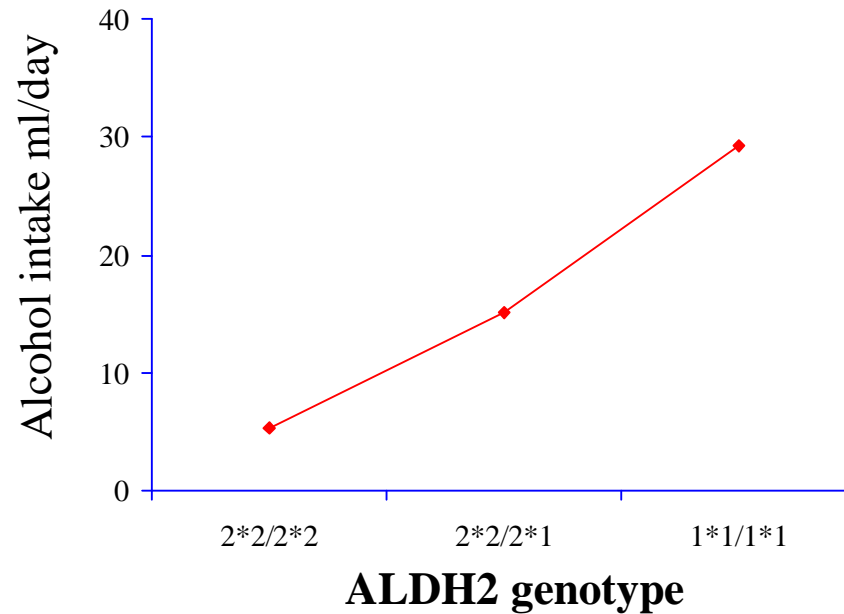
⇒ important to use well studied genes as instruments!

Example: Alcohol Consumption



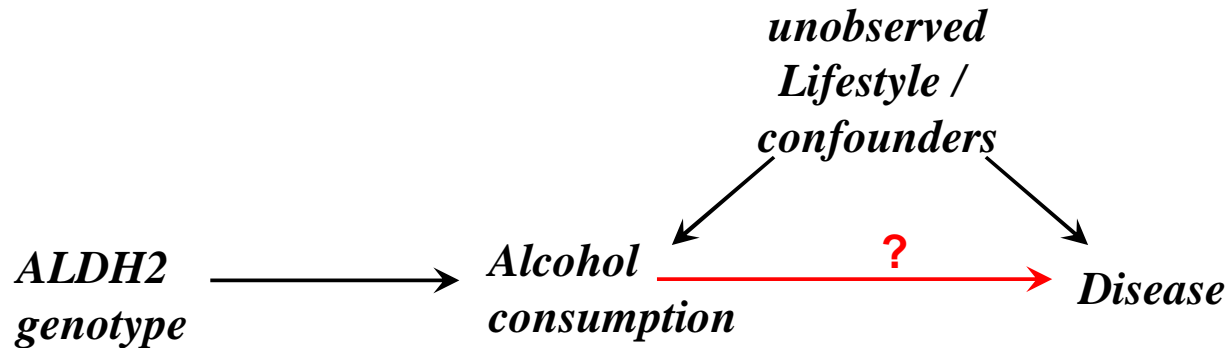
Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biochemistry well understood.

Example: Alcohol Consumption



Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biology well understood.

Example: Alcohol Consumption



Note 4: if the above is our causal graph, then under the null-hypothesis of no causal effect of alcohol consumption, there should be no association between *ALDH2* and disease;
While if alcohol consumption has a causal effect we would expect an association between *ALDH2* and disease.

Example: Alcohol Consumption

Findings:

(Meta-analysis by Chen et al. 2008)

Blood pressure on average 7.44mmHg higher and risk of hypertension 2.5 higher for *1*1 homozygotes than for *2*2 homozygotes (only males).
⇒ mimics the effect of *large versus low* alcohol consumption.

Blood pressure on average 4.24mmHg higher and risk of hypertension 1.7 higher for *1*2 heterozygotes than for *2*2 homozygotes (only males).
⇒ mimics the effect of *moderate versus low* alcohol consumption.
⇒ it seems that **even moderate** alcohol consumption is **harmful**.

Note: studies mostly in Japanese populations (where ALDH2*2*2 is common) and where women drink only little alcohol in general.

Problems with Mendelian Randomisation

Poor inferences may occur due to poor estimates of $G - X$ and $G - Y$ associations

—a genetic epidemiology problem

The core conditions can be violated in many different ways

—an instrumental variable problem

But some situations that ‘look’ like violations are okay.

GRAPHS can be used to check these conditions.

Estimation of Causal Effect

Requires **parametric** assumptions e.g. linearity & no interactions.

Plus: **structural** assumption

$$E(Y|X = x, U = u) = E(Y|do(X = x), U = u) = \mu + \beta x + \delta u$$

Then: consistent estimator for $ACE(x + 1, x) = \beta$ is

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}} \quad \text{and} \quad \text{st.dev}(\hat{\beta}_{IV}) = \frac{\sigma_G \sigma_{Y|X}}{\sigma_{G,X}}$$

where $\hat{\beta}_{Y|G}$ and $\hat{\beta}_{X|G}$ are least squares regression coefficients.

Note: weak instrument ($\sigma_{G,X} \approx 0$) makes $\hat{\beta}_{IV}$ unstable.

However

Typical applications of IV with Mendelian randomisation in epidemiology:

- Y is **binary** (X continuous, G categorical),
- $p(y|x, u)$ usually **non-linear**. Not always clear **how** relevant causal parameter is related to the relevant coefficients from the two separate regressions.
- **case-control** studies: want to use **causal odds ratio** (COR) or **causal risk ratio** (CRR) — not ACE.

IV Methods for Binary Outcome

Various IV estimators for binary outcomes are used in Epidemiology.

They all make **different** additional parametric assumptions (i.e. besides the core conditions and structural assumption).

When assumptions are violated, resulting estimates will be biased estimates of the target causal effect.

Can all behave unreliably in the all-binary situation.

Didelez, Meng & Sheehan (2008) Research Report 08-20 Department of Mathematics, University of Bristol

IV Methods for Binary Outcome

Naïve estimators: ORs and RRs from a regression of Y on X .

Biased if there is unobserved confounding.

Wald-type estimators: Based on ratios of differences e.g. Wald OR:

$$OR_{YG}^{(\mu_1 - \mu_0)^{-1}} \text{ where } \mu_1 - \mu_0 = E(X|G = 1) - E(X|G = 0)$$

and

$$OR_{YG} = \frac{p(Y = 1|G = 1)p(Y = 0|G = 0)}{p(Y = 0|G = 1)p(Y = 1|G = 0)}$$

for binary G .

Heuristic — no model assumptions for theoretical justification.

IV Methods for Binary Outcome

Wald-type estimators: If all relationships are log-linear and conditional variance of X is variation independent of the mean

$$\begin{aligned}\text{WaldRR} &= \left\{ \frac{E(Y|G=1)}{E(Y|G=0)} \right\}^{1/\Delta} \\ &= RR(Y|G)^{1/\Delta}\end{aligned}\tag{1}$$

is consistent for the CRR.

For rare disease, the Wald OR is as good as consistent under the same model assumptions.

Wald-type estimators: useful when we do not have joint data on X, G and Y e.g. meta-analysis.

IV Methods for Binary Outcome

Multiplicative structural mean model:

$$\log \left\{ \frac{E(Y|X = x, G, \text{do}(\tilde{X} = x))}{E(Y|X = x, G, \text{do}(\tilde{X} = 0))} \right\} = \gamma x. \quad (2)$$

where X denote the 'natural' exposure level while \tilde{X} denotes the exposure that is set by an intervention (overruling the 'natural' X) and $\tilde{X} = 0$ stands for an appropriate baseline value.

Y depends **causally** on \tilde{X} but still **associated** with X since X is informative for unobserved confounding that also predicts Y .

Explicit form for binary X \longrightarrow **estimating equations** for continuous X .

IV Methods for Binary Outcome

Multiplicative structural mean model:

- Unlike Wald-type, no distributional assumptions for X (given G).
- Requires **joint information** on all relevant variables.
- Assumes **effect of exposure on exposed** is same for different levels of G — no effect modification by instrument — **local causal risk ratio**.

Reasonable?

Population CRR if no interaction between U and X^* on multiplicative scale.

Can we ever rule out such interactions?

Simulations studies for the all-binary case \Rightarrow SMM performed better in terms of asymptotic bias.

ALSPAC — FTO, BMI and Asthma

About 4,000 seven-year old children.

G — FTO genotype dichotomised TT & AT = 0, AA = 1 (A allele known to be associated with fat mass)

X — Body mass index (kg/m^2)

Y — doctor diagnosed asthma (yes,no) at 91 months (prevalence $\sim 14\%$)

Test for association between G and Y — OR not significantly different from 1 ($p = 0.134$).

ALSPAC — FTO, BMI and Asthma

IV estimates of causal effect of BMI on asthma:

Naïve OR: 1.0806

Wald OR: 2.8776

Wald RR: 2.4584

SMM RR: 0.8799 (GMM RR: 0.7088)

All have wide confidence intervals and we are, perhaps, estimating too close to the null value **BUT** why should we get different directions of causal effect?

SMM and GMM only use information on exposure in the disease group ($Y = 1$): asthmatics heavier on average and go against population trend with genotype.

Conclusion

- Need a **formal causal framework** to disentangle associational and causal concepts in observational epidemiology.
- Causal inference always requires background knowledge to verify that assumptions are met.
- In case of Mendelian randomisation we have background knowledge → genetics.
- IV methods avoid the assumption of **no unobserved confounding** — but make other assumptions instead!
- What do these mean in epidemiological applications?

Discussion Points

IV assumptions that crop up in other contexts:

- No effect modification by instrument.
- No treatment received (i.e. no defiers) in control group — what does that mean here?
- Monotonicity assumption: $X | (G = 1) \geq X | (G = 0)$.
- Local causal effects
 - effect of exposure on the exposed (effect of treatment received) as targeted by SMM. Is this what we want?
 - effect of treatment on the **compliers**??

In order for IV methods to be useful for epidemiological applications, we need assumptions that can be verified in our context.