Mendelian Randomisation: an Instrumental Variable Approach to Inferring Causality in Observational Epidemiology

Nuala Sheehan

Departments of Health Sciences and Genetics

UNIVERSITY of LEICESTER

ESRC Research Methods Festival

Oxford July 2010

Causal Inference

Will assume that this is always inference about interventions.

Epidemiology is often 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 with Inferring Causality

- Epidemiology mainly based on observational studies
- "Association $\neq$ causation" i.e. might find an association but intervention turns out to be useless
- Randomised trials are the ideal "gold standard" but not always possible for ethical or practical reasons
- observational findings often not reproduced in randomised trials.
- Possible reasons:
  - reverse causation
  - confounding
  - selection effect etc.

Interventions

Causal vocabulary is often used carelessly in the literature.

We must formally distinguish between association and causation and for this, we need special notation.

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
  $\Rightarrow$ observed in a randomised study.

- $p(y|X = x)$ also depends on x with confounding/reverse causation
  $\Rightarrow$ observed in an observational study.

e.g. $X =$ yellow fingers, $Y =$ lung cancer.

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

Risk Ratio: $CRR(x_1, x_2) = \dfrac{p(Y = 1|do(X = x_1))}{p(Y = 1|do(X = x_2))}$

Odds Ratio: $COR(x_1, x_2) = \dfrac{p(Y = 1|do(X = x_1))p(Y = 0|do(X = x_2))}{p(Y = 0|do(X = x_1))p(Y = 1|do(X = x_2))}$

Mathematically, the causal effect is identifiable (hence estimable) if we can re-express it purely in observational terms i.e. without $do(X)$.

Identifiability using Instrumental Variables

Standard Approach "No unobserved confounding": Assumes all confounders (or a sufficient set) measured $\Rightarrow$ adjust for them in regression models in the usual way.

Can not always assume this $\longrightarrow$ need to deal with confounding by other means, e.g. **instrumental variables (IVs)**

There are different types of assumption required:

(in)dependencies
structural
} allow testing for causal effect

parametric form for estimation

Core Conditions

For the effect of X (phenotype/exposure) on Y (disease) in the presence of unobserved confounding, U, a third observable variable G qualifies as an **instrument** if

1. $G \perp\!\!\!\perp U$: G independent of unobserved confounders

2. $G \not\perp\!\!\!\perp X$: G associated with phenotype/exposure

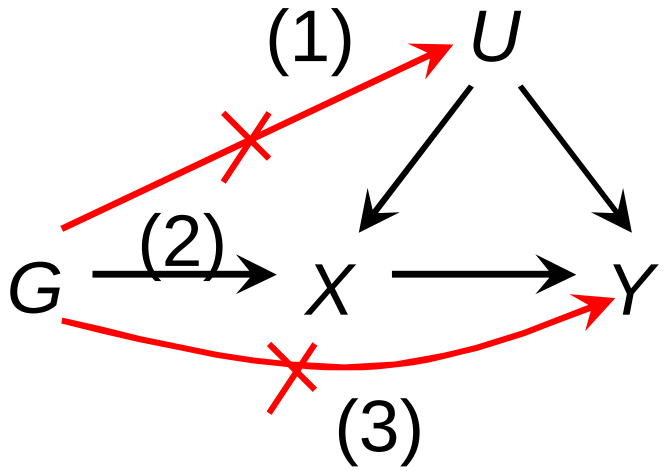
3. $G \perp\!\!\!\perp Y \mid (X, U)$: G and Y conditionally independent given X and U.

G is only associated with disease via its effect on the phenotype/exposure with X,

Cannot forget about U!

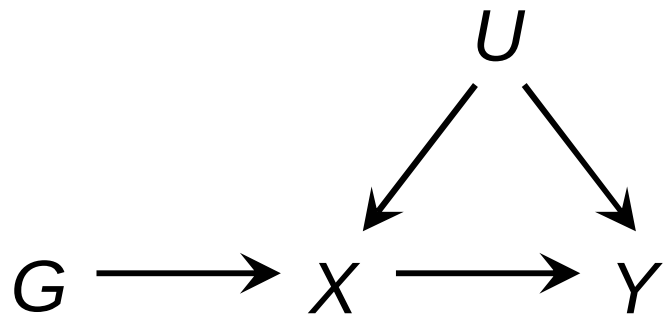
Core Conditions — Graphically

DAG — shorthand way to encode conditional independence restrictions.



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

Core Conditions — Graphically



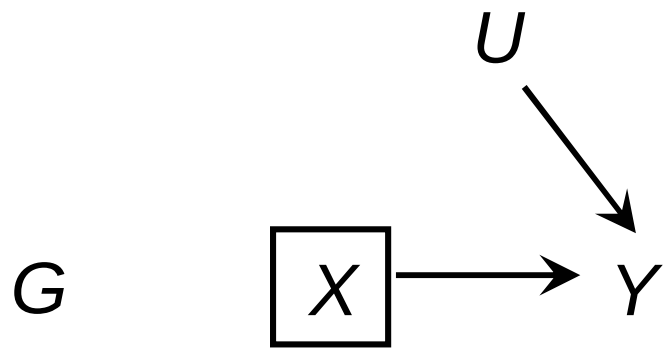
Equivalent to factorisation

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

Also need structural assumption for causal inference:

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

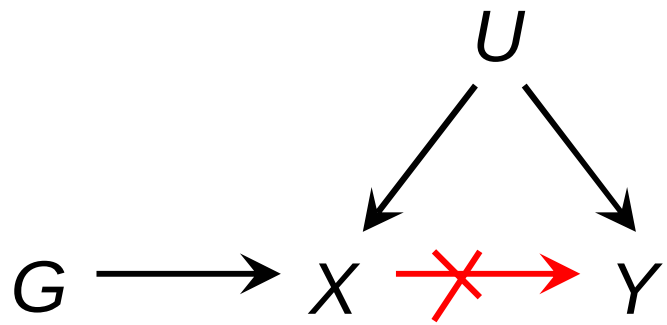
$$p(y, u, g|do(X = x^*)) = p(y|x^*, u)p(u)p(g)$$

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

Testing for Causal Effect

With these conditions alone, we have that there is

no causal effect of X on Y iff G independent of Y.



So any test for association between G and Y can be taken as a test for a causal effect of X on Y — regardless of the distributions of G, X and Y. (Katan 1986)

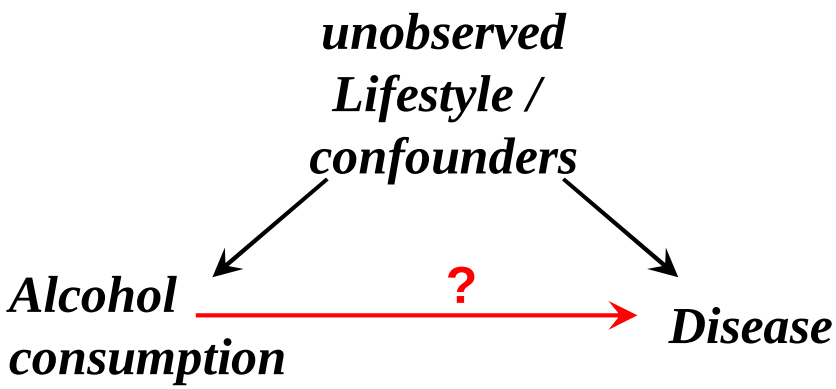
Mendelian Randomisation

An Instrumental Variable (IV) method — with genotype as instrument.

- 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. — typical confounding factors $\Rightarrow$ kind of 'randomised';
- Genes are determined before birth, no reverse causation possible;
- Conjecture: if and only if phenotype is causal for disease should we find an association between genotype and disease.

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.

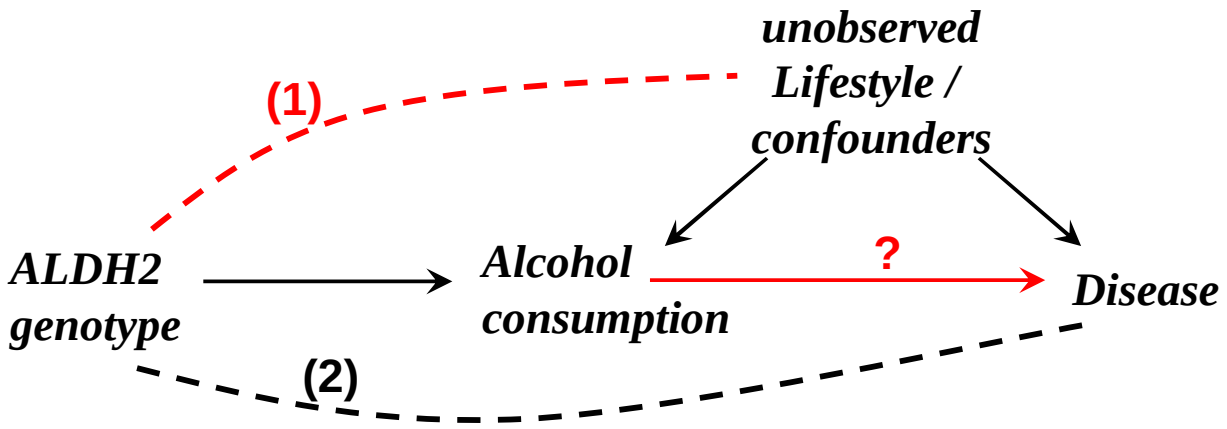
Two alleles/variants: wild type *1 and "null" variant *2.

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

$\Rightarrow$ *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

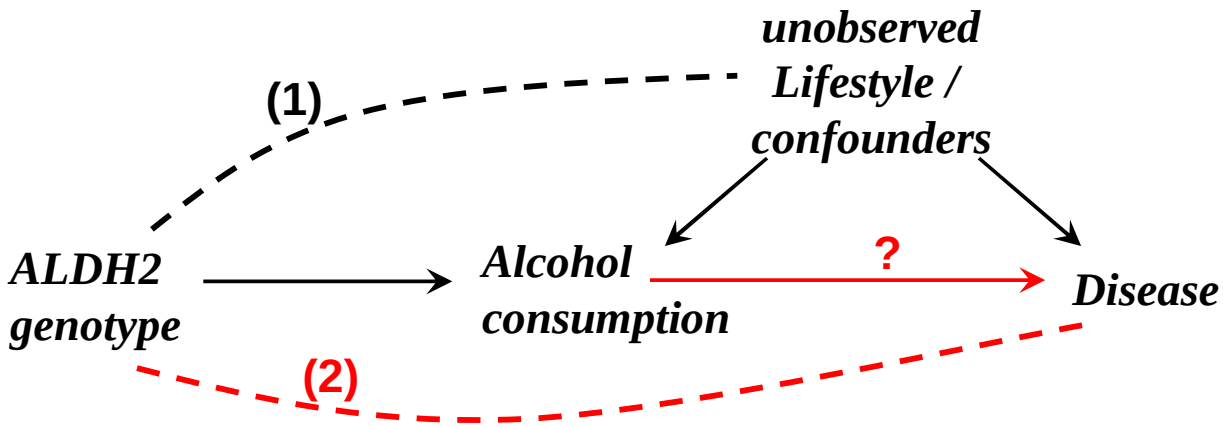


Note 1: due to random allocation of genes at conception, can be fairly confident that genotype is not associated with unobserved confounders.

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)

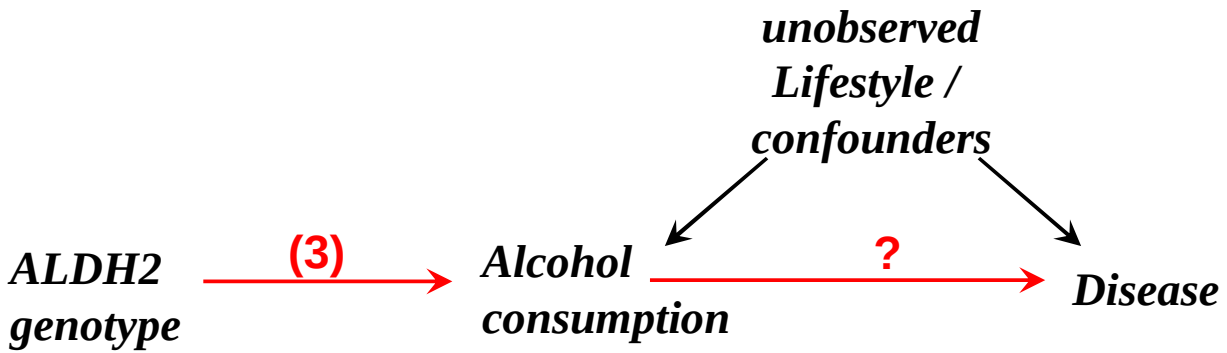
Example: Alcohol Consumption



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.

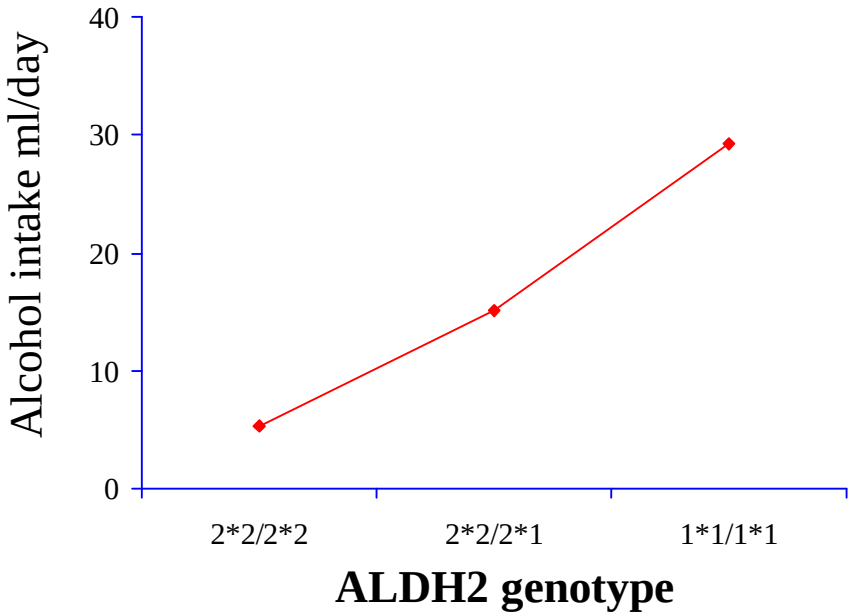
$\Rightarrow$ important to use well studied genes as instruments!

Example: Alcohol Consumption



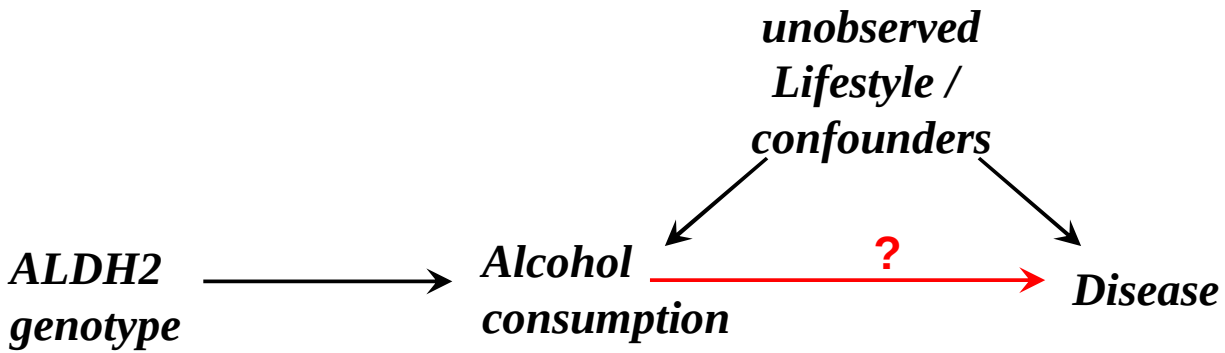
Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biochemistry well understood.

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).

$\Rightarrow$ 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).

$\Rightarrow$ mimics the effect of *moderate versus low* alcohol consumption.

$\Rightarrow$ 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 $\longrightarrow$

No association between variant and BP/hypertension in women.

Problems with Mendelian Randomisation

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

—a genetic epidemiology problem. May need very large studies.

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: (2SLS) consistent estimator for $ACE(x + 1, x) = \beta$ is

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

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.

Typical Mendelian Randomisation IV Applications

- Y is binary (X continuous, G categorical),

- $p(y|x,u)$ hence non-linear. Not always clear how target causal parameter is related to relevant coefficients from the two regressions

  —involves marginalising over U and result typically dependent on (unknown) distribution of U e.g. logistic case

$$E(Y \mid do(X=x)) = \int \frac{exp(\alpha+\beta_1 x+\beta_2 u)}{1+exp(\alpha+\beta_1 x+\beta_2 u)} p(u)du \neq \frac{exp(\alpha^*+\beta_1 x)}{1+exp(\alpha^*+\beta_1 x)}$$

  even if U normally distributed — non-collapsibility of logistic regression model (Greenland et al. 1999).

- typically want COR or CRR — not ACE.

IV Methods for Binary Outcome

Various IV estimators for binary outcomes are used in Epidemiology.

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

They may target different causal parameters depending on what is assumed (local versus population effects).

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

Can be quite sensitive to these assumptions and have all been shown to behave unreliably in a small numerical study.

Didelez, Meng & Sheehan (2010) Statistical Science. In Press

Issues

- All measurements in a Mendelian randomisation study are prone to measurement error. Need to check core conditions apply to observed values rather than underlying values
- Weak instrument: Many gene–phenotype associations are weak possibly due to population stratification / LD / genetic heterogeneity / measurement errors or when behaviour (e.g. under social pressure) 'overrules' genetic predisposition.
- Finding good genetic instruments: functionality of genes not well understood if only based on association studies.
- Case-control data: selection on disease status violates core IV condition.
- Sampling versus asymptotic behaviour of these estimators?

Conclusion

- Despite historical reluctance, we need to be able to use causal terminology in epidemiology.
- Need a formal causal framework to disentangle associational and causal concepts.
- IV methods avoid the assumption of no unobserved confounding — but make other assumptions instead!
- What do these mean in epidemiological applications? Can we live with them for any particular application?
- Causal inference always requires background knowledge to verify that assumptions are met $\longrightarrow$ genetics for Mendelian randomisation.
- Must pay attention to details as not all IV methods target the same causal parameters. "Sometimes, we get what we need". (Angrist & Pischke 2009)