Graphical Models for Causal Inference using Mendelian Randomisation

Vanessa Didelez

Department of Mathematics

University of Bristol

(Thanks to many collaborators)

Warwick, April 2009

Overview

- Motivation: causal inference in observational epidemiology
- A formal framework for causality
- The idea of instrumental variables
- The idea of Mendelian randomisation
- Using graphs to represent assumptions / check violations
- Problems with finding genetic variants as IVs
- Conclusions

Motivation

Epidemiology interested in effect of interventions (‘drink less alcohol’, ‘eat folic acid’ etc.)

Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded.

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

Motivation

Epidemiology interested in effect of interventions (‘drink less alcohol’, ‘eat folic acid’ etc.)

Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded.

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

Motivation

Epidemiology interested in effect of interventions (‘drink less alcohol’, ‘eat folic acid’ etc.)

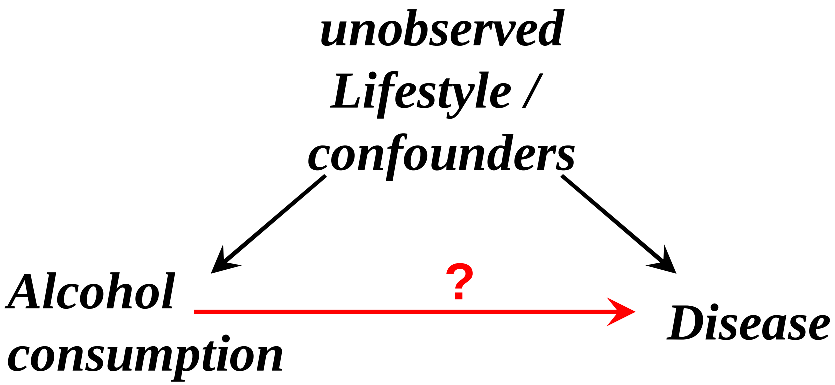
Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded. Example

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

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'.

Motivation ctd.

Epidemiology interested in effect of interventions (‘drink less alcohol’, ‘eat folic acid’ etc.)

Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded.

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

Motivation ctd.

Epidemiology interested in effect of interventions (‘drink less alcohol’, ‘eat folic acid’ etc.)

Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded.

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

Effect of Interventions — Formally

Want notation to 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
  $\Rightarrow$ typically observed in a randomised study.

- $p(y|X = x)$ will also depend on x when there is confounding, reverse causation etc.
  $\Rightarrow$ typically observed in an observational study.

Instrumental Variables

A variable G **is an instrument** for the effect of manipulating X on Y if variable(s) $U (=$ unobserved confounders) exist such that

1. $G \perp\!\!\!\perp U$
2. $G \not\perp\!\!\!\perp X$
3. $G \perp\!\!\!\perp Y \mid (X, U)$.

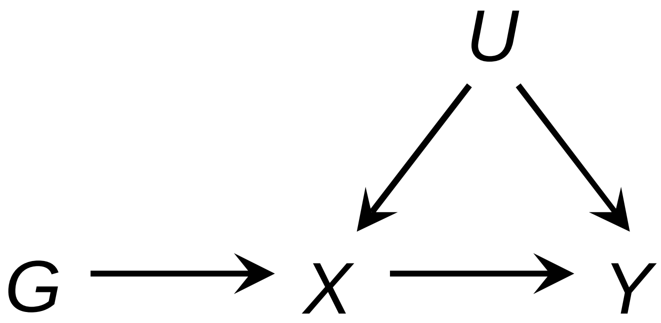
And if the following **structural** assumptions are valid:

$$p(y|x,u), \quad p(g) \quad \text{and} \quad p(u)$$

are not changed by intervention in X, i.e. under conditioning on $do(X)$.

Instrumental Variables — Graphically

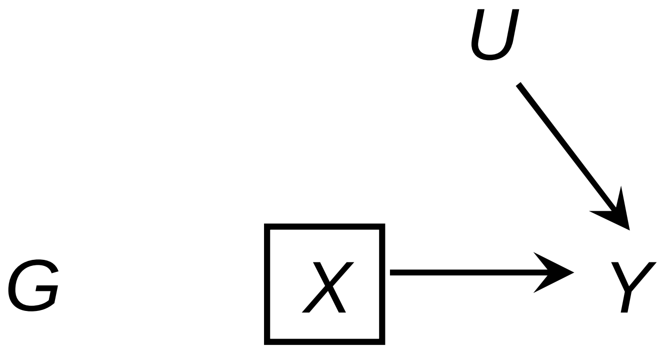
(This is a **conditional independence** graph, not a *causal* graph)



Equivalent to factorisation

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

Instrumental Variables — Graphically



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

$$p(y, u, g | do(X = \tilde{x})) = p(y | \tilde{x}, u) p(u) p(g)$$

Note: implies $Y \perp\!\!\!\perp G | do(X)$ — also known as **exclusion restriction**.

Why does this Help with Causal Inference?

Testing:

check if $Y \perp\!\!\!\perp G$ — this is (roughly) testing whether there is a causal effect at all.

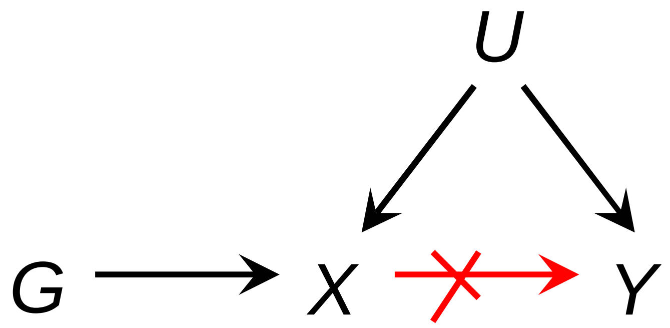
Estimation:

(1) when all observable variables are discrete, we can obtain bounds on causal effects without further assumptions.

(2) for point estimates need some (semi–)parametric / structural assumptions, as well as clear definition of target causal parameter.

Testing for Causal Effect

No causal effect "$\Leftrightarrow$" G independent of Y



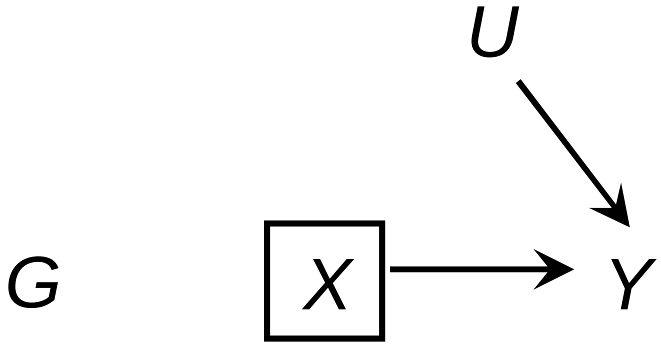
from factorisation $p(y,x,u,g) = p(y|u)p(x|u,g)p(u)p(g)$

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

Note: need *faithfulness* assumption here.

Estimating the Causal Effect

Implicit / or explicit in many approaches:
exclusion restriction $Y \perp\!\!\!\perp G | do(X)$.



Set up model for $p(y|do(X=x))$ and fit it so it is independent of G

Cf. (generalised) **method of moments** / **structural mean models**;
plug–in estimators use G to predict ‘unconfounded’ X ($\approx do(X)$).

‘Untestable’ Assumptions

The assumptions

1. $G \perp\!\!\!\perp U$

3. $G \perp\!\!\!\perp Y \mid (X, U)$.

cannot (easily) be tested from data!
In particular they do not imply that $G \perp\!\!\!\perp Y | X$ or $G \perp\!\!\!\perp Y$.

Also: structural assumptions cannot be tested and may even depend on the particular intervention you have in mind.

⇒ Need to be justified based on subject matter background knowledge.

⇒ **Mendelian randomisation**

Example: Alcohol Consumption

Genetic Instrumental Variable?

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

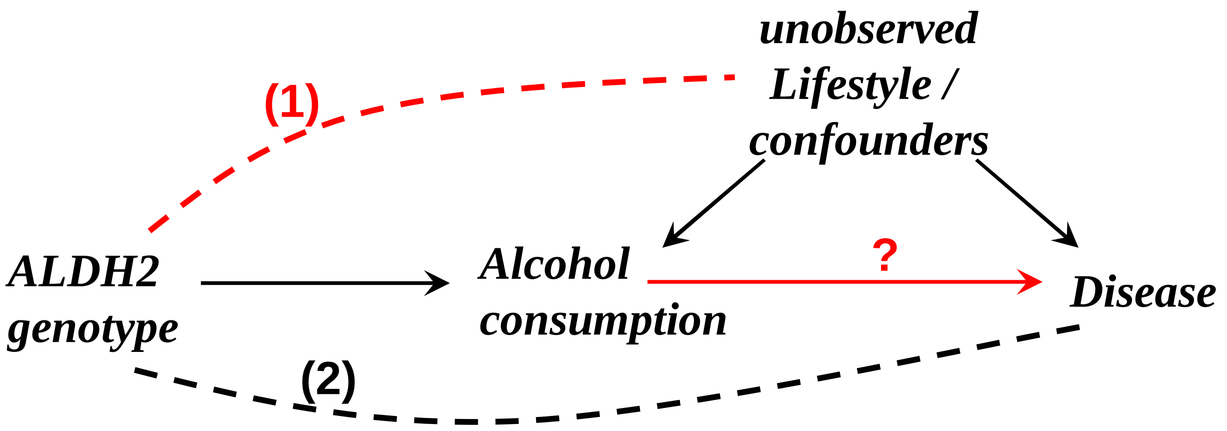
Two alleles/variants: wildetype *1 and “null” variant *2.

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

$\Rightarrow$ *2*2 homozygous individuals have low alcohol consumption *regardless* of their other lifestyle behaviours

IV–Idea: check if these individuals have a different risk than others for alcohol related health problems!

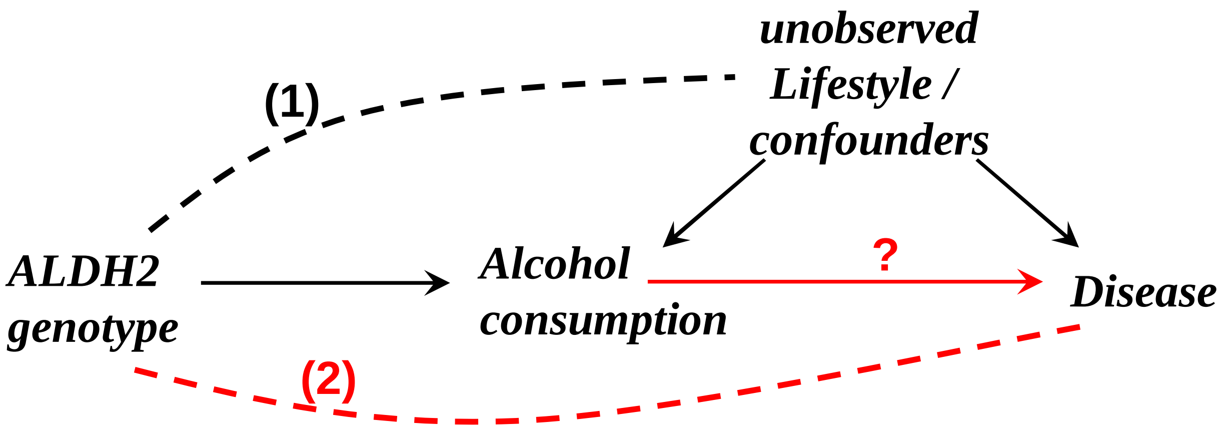
Example: Alcohol Consumption



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

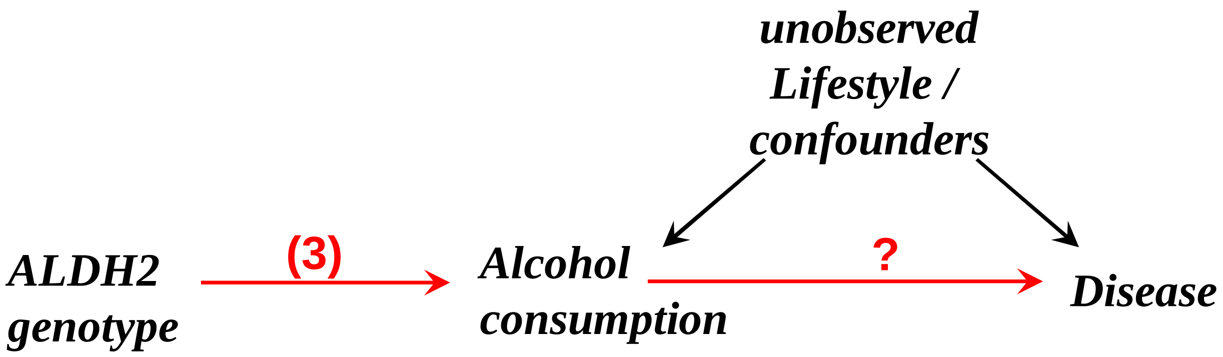
Example: Alcohol Consumption



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.

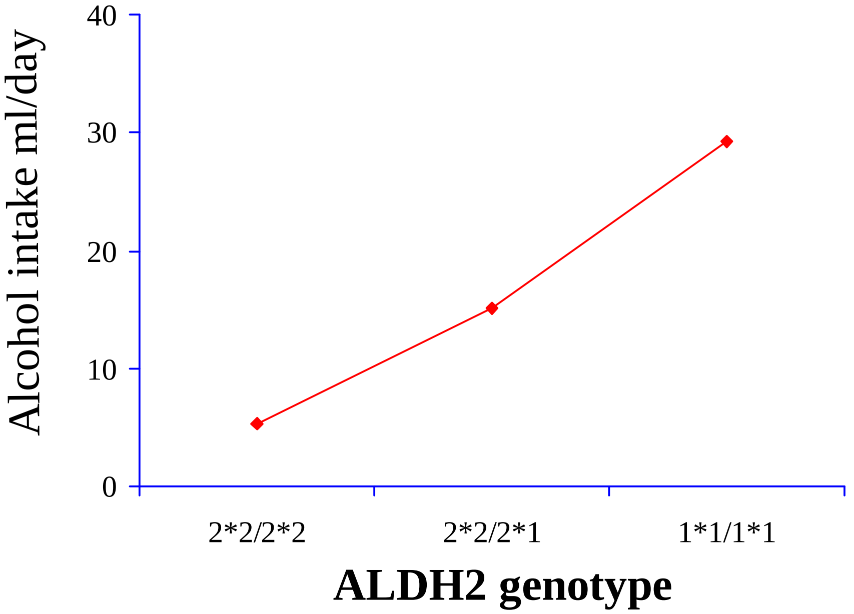
⇒ important to use well studied genes as instruments!

Example: Alcohol Consumption



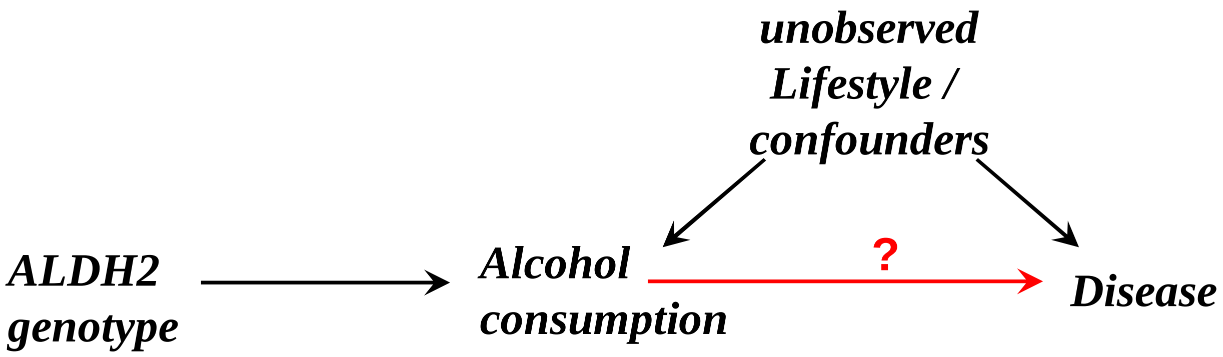
Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biochemistry well understood.

Example: Alcohol Consumption



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

Example: Alcohol Consumption



Causal Effect? under IV assumptions, the null–hypothesis of no causal effect of alcohol consumption, should imply 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 ALDH2*1*1/*1*1 than for ALDH2*2*2/*2*2 carriers (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 ALDH2*1*1/*2*2 than for ALDH2*2*2/*2*2 carriers (only males).
⇒ mimics the effect of *moderate versus low* alcohol consumption.

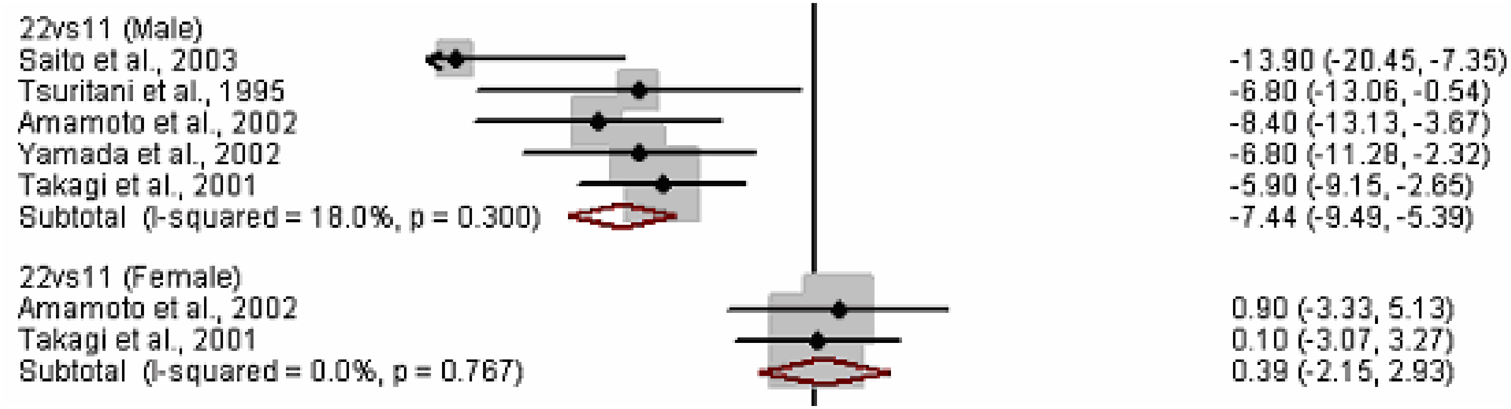
⇒ it seems that **even moderate** alcohol consumption is **harmful**.

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

Example: Alcohol Consumption

(Chen et al., 2008)

Is condition $Y \perp\!\!\!\perp G | (X, U)$ satisfied?



Some indication

Women in Japanese study population do not drink. ALDH2 genotype in women not associated with blood pressure ⇒ there does not seem to be another pathway creating a G–Y association here.

Violations of Core Conditions

So far have argued that in Mendelian randomisation studies we can reasonably believe that core conditions are *satisfied*.

Now will discuss situations and examples when they are typically *violated*.

(Davey Smith & Ebrahim, 2003)

Use **graphs** to represent assumptions / background knowledge to check if core conditions are violated or not.

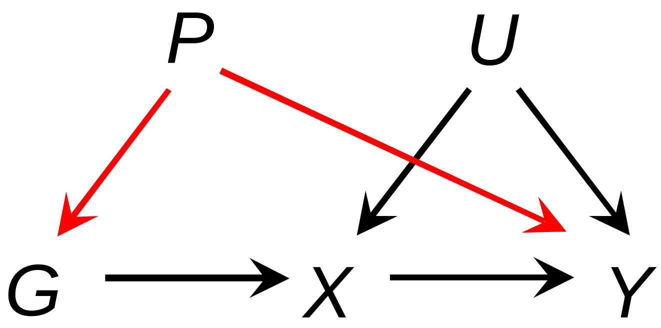
Population Stratification

Population stratification occurs when there exist population subgroups that experience both, different disease rates (or different distributions of phenotypes) and have different frequencies of alleles of interest.

$\Rightarrow$ might violate condition $Y \perp\!\!\!\perp G | (X, U)$.

Population Stratification (1)

Graphical representation:
variable P = population predicts G as well as Y



Here $Y \not\!\perp\!\!\!\perp G | (X, U)$!

$\Rightarrow$ can be avoided by sensible study design.

Population Stratification (1)

Example: Study of native Americans from Pima and Papago tribe.

(Knowler et al, 1988)

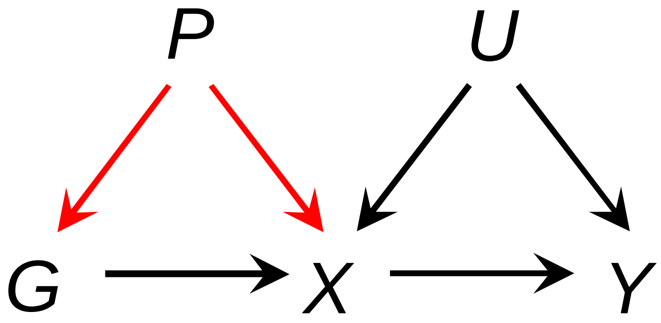
- Strong inverse association between a HLA–haplotype and type 2 diabetes;
- Individuals with full American Indian heritage: haplotype prevalence 1% and type 2 diabetes prevalence 40%;
- In Caucasian population: haplotype prevalence 66% and type 2 diabetes prevalence 15%.

⇒ Solution: carry out analyses *within population strata*!

- **Note:** need to be aware of existence of such population subgroups.

Population Stratification (2)

Also possible: phenotype distribution different in subgroups



Core conditions still satisfied.

Strength of IV could be affected in positive or negative way.

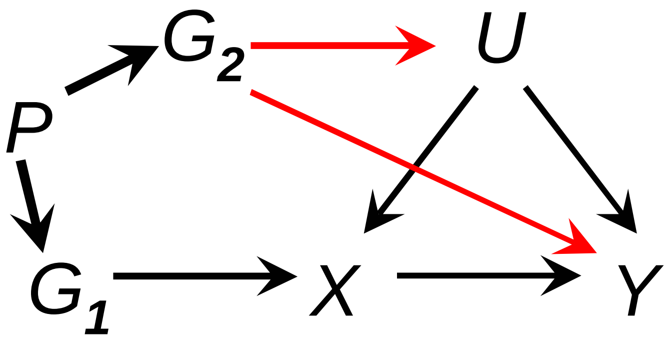
Linkage Disequilibrium (1)

Linkage disequilibrium (LD): is the correlation between allelic states at different loci, traditionally regarded as stemming from close proximity to each other on chromosome.

If genetic variant chosen as instrument is in LD with another gene that is in turn associated with some of the unobserved confounders or even predicts the disease, then condition $Y \perp\!\!\!\perp G | (X, U)$ might again be violated.

Linkage Disequilibrium (1)

Graphical representation: G_1 = chosen instrument, in LD with other gene G_2 through parental genes P.



Here $Y \not\perp\!\!\!\perp G | (X, U)$ or $G \not\perp\!\!\!\perp U$!

Linkage Disequilibrium (2)

‘Right’ gene?

Often it is plausible that the gene chosen as instrument is not *the* causal gene for the phenotype of interest, but instead it is in LD with the causal gene.

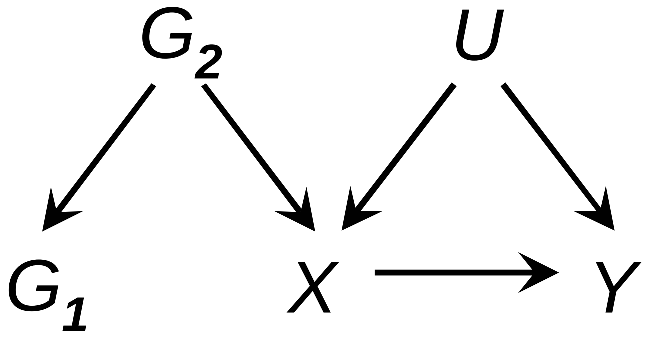
We could also regard this as **measurement error** when assessing the gene.

This does not necessarily imply any violations of the core IV conditions.

$\Rightarrow$ we do not need to use the causal gene for Mendelian randomisation, just one that is correlated with the phenotype, as long as core conditions satisfied.

Linkage Disequilibrium (2)

Chosen gene G_1 is not ‘causal’



Core conditions still satisfied.

Again, strength of IV could be affected in positive or negative way.

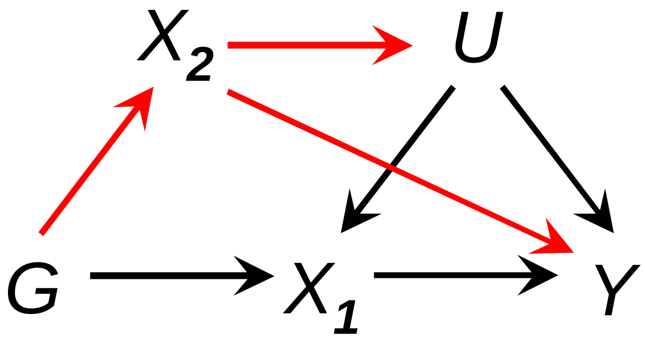
Pleiotropy

Peiotropy refers to a genetic variant having multiple functions, i.e. the chosen gene / instrument might not only affect the phenotype of interest but also other traits.

If the pleiotropic effects influence or predict the outcome through other pathways, then the IV core conditions might again be violated.

Pleiotropy

Graphical representation:
G affects another phenotype X_2 (phenotype of interest is X_1)



Here, again, $Y \not\perp\!\!\!\perp G | (X, U)$ or $G \not\perp\!\!\!\perp U$.

Pleiotropy

Example: use of APOE genotype as IV for causal effect of both types of cholesterol (HDLc and LDLc) on myocardial infarction risk.

- Causal effect of HDLc and LDLc well known from RCTs.
- APOE strongly associated with HDLc and LDLc, but surprisingly APOE not associated with myocardial infarction risk.
- Explanation: the $\epsilon 2$ variant of APOE is also related to less efficient transfer of very low-density lipoproteins and chylomicrons from the blood to the liver, greater postprandial lipaemia, and an increased risk of type III hyperlipoproteinaemia $\rightarrow$ all of which increase myocardial infarction risk.

$\Rightarrow$ Due to multiple effects of APOE, all affecting myocardial infarction risk, this is an unsuitable IV for a Mendelian randomisation study.

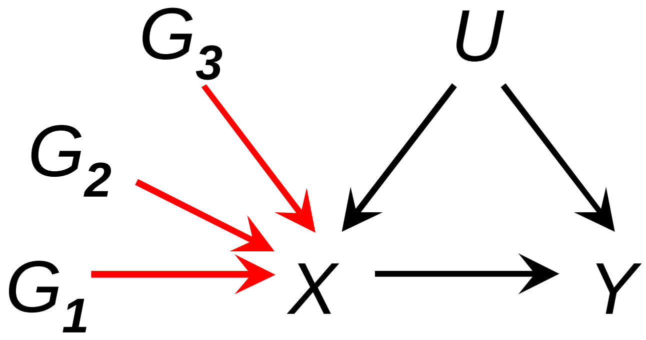
Genetic Heterogeneity

Genetic heterogeneity means that more than one genotype affect or determine the phenotype of interest, possibly via different biochemical routes.

To decide if this is a problem we need to establish whether these are in LD and whether any of them have pleiotropic effects.

Genetic Heterogeneity

Graphical representation:
Independent genes affecting the phenotype.



All core IV conditions still satisfied for each (or any subset of) genes.

Genetic Heterogeneity

Potential **Benefits**:

- Multiple instruments can be used to examine if the core conditions are violated. If different genes affect the phenotype via different pathways, it is unlikely that both are affected in the same way by population stratification, LD or pleiotropy.

$\Rightarrow$ compare results obtained for each genotype separately.

- CRP example: circulating CRP is related to CRP–gene but also to other locus, $IL6$, via different pathways.
- Can also combine multiple instruments for *one* IV analysis (common in econometrics).

Measurement Errors

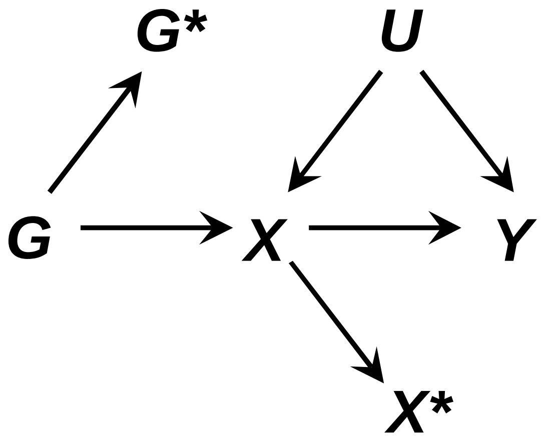
All measurements in a Mendelian randomisation study are prone to measurement error.

In particular G: we often do not have the 'right' causal gene, but just a strongly linked locus (cf. LD).
But G might also be mis-measured for other reasons $\Rightarrow$ "in principle" not a problem if measurement error not differential.
But might be a problem if G–X association is obtained from a different study than G–Y association (meta analysis).

In practice, X is also typically measured with error. The 'true' phenotype through which G acts is e.g. lifelong exposure to high or low CRP levels, while what is measured is only the value at a particular time.

Measurement Errors

Graphical representation:
G^* and X^* are the actual measurements of G and X with possible measurement errors.



Can use G^* instead of G but not X^* instead of X as $G^* \not\perp\!\!\!\perp Y|(X^*, U)$.

Example: CRP and Insulin Resistance

(Lawlor et al., 2008)

Is condition $Y \perp\!\!\!\perp G|(X, U)$ satisfied?

- Must believe: no other pathway / association than via X=circulating CRP present.
- Variants used to generate CRP–haplotypes are in close LD with another locus on the CRP gene, but unlikely to be involved in pleiotropic effect due to their known functional role.
- No published evidence that CRP has been influenced by population selection; also study used white British women from areas with little migration.
- Genetic heterogeneity is present and planned to be exploited using multiple instruments.
- Measurement error: CRP–haplotypes and CRP association in same direction and magnitude as in previous studies.

Finding Genetic Instrumental Variable

Main limitation:

finding genetic variant that is suitable as instrumental variable!
Not many are known yet for typical exposures of interest in epidemiology.
Optimism: rapid expansion of knowledge in functional genomics!

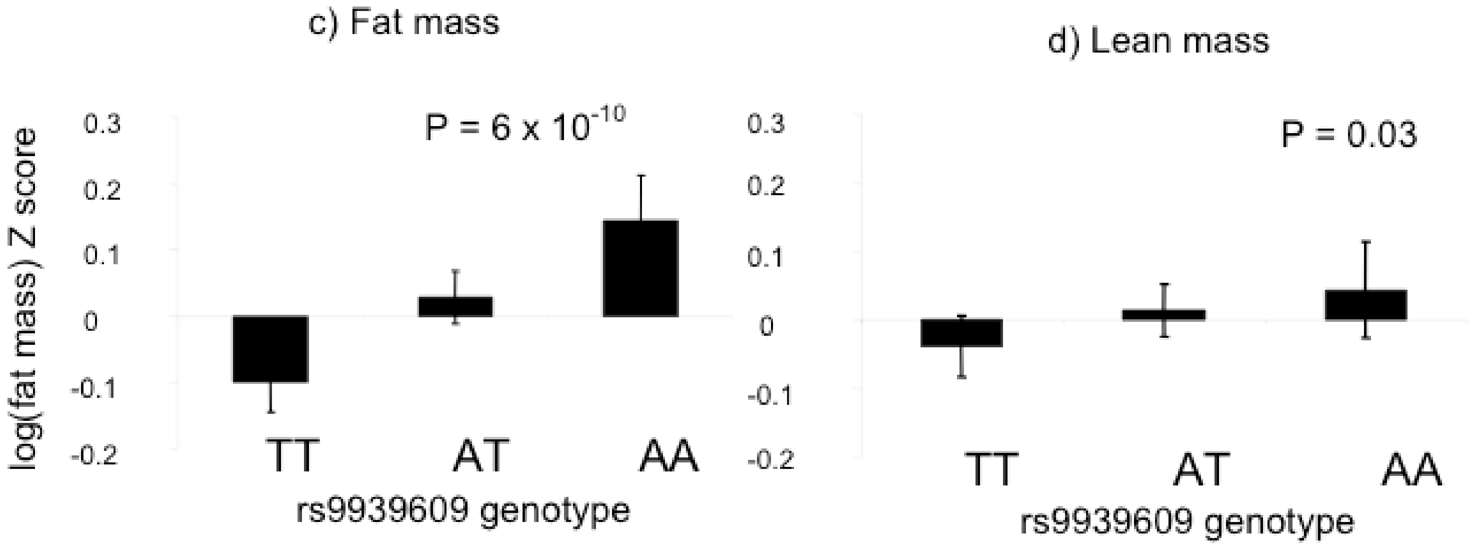
Genome wide association studies:
gene–phenotype associations often weak, low power, not reproducible.

But even when strong/reproducible association found, functionality of genes not well understood if only based on association studies.
⇒ cannot be as confident that core conditions satisfied.

Example: recently found *FTO* gene associated with BMI / fat mass but functionality not (yet?) understood. (Frayling et al., 2007)

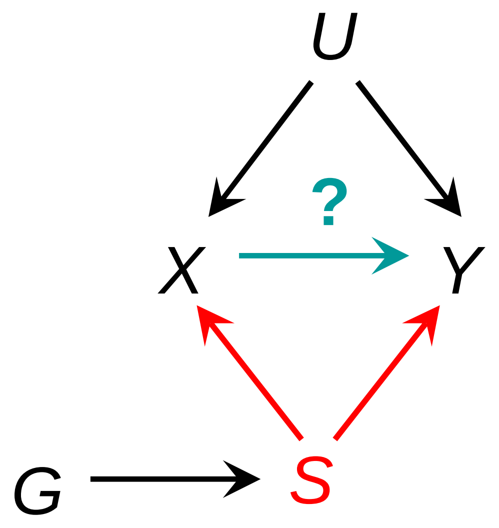
Example: FTO — fat/lean mass

From genome wide association studies:



Finding Genetic IV — Association Studies

Alternative explanation: G causes a condition S, which in turn causes the phenotype and disease of interest.

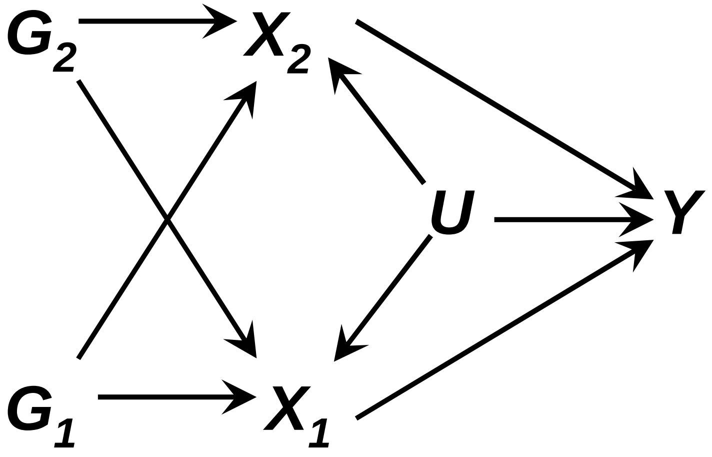


Here $G \not\!\perp\!\!\!\perp X$ (genotype and phenotype are associated), $G \perp\!\!\!\perp U$.

But $G \not\!\perp\!\!\!\perp Y \mid (X, U)$ — and cannot check this from data.

More Complex Situation

Often: several genotypes and phenotypes interacting
⇒ need to generalise available methods



Conclusions

IVs enable (limited) causal inference in presence of unobserved confounding.

Graphical models useful to

- represent assumptions underlying IVs / Mendelian randomisation;
- identify potential for violations;
- support arguments that assumptions are satisfied.