Causal Inference using Instrumental Variables in an Epidemiological Application

Nuala Sheehan & Vanessa Didelez

University of Leicester & University College London

Causal Inference in Epidemiology

One important reason is to **find** and **assess** the size of the effect of **modifiable** risk factors (e.g. diet) on diseases so that public health **interventions** can be informed.

Example:

Observational studies consistently show positive association between homocysteine levels and coronary heart disease (CHD).

Homocysteine levels are reduced by folate intake.

If the relationshiop is causal, we can reduce CHD risk by adding folate to the diet.

Problem: Association $\neq$ Causation

We might find an association but the intervention turns out to be useless.

Example: Beta-carotene and lung cancer

- Peto et al. (1981): increased intake of vitamin beta-carotene "reduces" risk of smoking related cancers
- Could not be reproduced in randomised controlled trials (1994)

Need to distinguish between association and causation so that we know whether an intervention will be useful.

Problem: Association $\neq$ Causation

Fisher (1926): **Randomised** experiments render **reverse causation** and **confounding** highly unlikely.

Randomised/controlled experiments not always possible—ethical, practical or financial problems.

Require **causal inferences** from **observational** data.

Confounding problems—exposures and diseases of interest often related to socioeconomic or behavioural factors. We can try to adjust for confounding but need to know and measure the confounding factors.

Mendelian Randomisation

Katan(1986)—letter to the Lancet:

Hypothesis under debate in mid-1980s: low serum cholesterol increases risk of cancer.

Have to satisfactorily eliminate

1. **Reverse causation:** Does presence of hidden tumours induce a lowering of cholesterol in future cancer patients?
2. **Confounding:** Are other factors such as diet and smoking affecting both cholesterol levels and cancer risk?

Mendelian Randomisation

Katan(1986)—letter to the Lancet:

Rare disease abetalipoproteinaemia $\longrightarrow$ practically zero cholesterol levels. No evidence of premature cancer.

Larger sample of individuals genetically predisposed to having low cholesterol levels?

Known that alleles E2 (8%), E3 (77%) and E4 (15%) of Apolipoprotein E (APOE) polymorphism associated with different levels of cholesterol.

APOE2 associated with lower levels.

Mendelian Randomisation

Katan(1986)—continued:

- Many E2 allele carrriers: majority have relatively low levels of serum cholesterol from birth.
- Crucially, similar on average to those carrying E3 and E4 alleles in all other respects.
- Mendel's Second Law: APOE genes assigned randomly during meiosis and independently of confounding factors.
- Hence, no need for a prospective study. Just compare APOE genotypes in cancer patients and controls.

Mendelian Randomisation

Katan(1986)—continued:

- If low serum cholesterol level is really a risk factor for cancer, then patients should have more E2 alleles and controls should have more E3 and E4 alleles.
- On the other hand, if the reported associations are indeed spurious, APOE alleles should be equally distributed across both groups.

Conjecture: we should find an association between genotype and disease **if and only if** the phenotype is causal for the disease.

Formal Approach

- Let X be the phenotype and Y the disease of interest;
- Let G be the genotype related to X;
- Let U stand for unmeasued confounders.

e.g.

$X =$ homocysteine level

$Y =$ 1 if CHD, 0 if no CHD (binary)

$G =$ MTHFR genotype—typically dichotomised

$U =$ lifestyle.

Interventions

Want notation to distinguish between association and causation.

Intervention: setting X to a value x: use $do(X = x)$.

$P(Y|do(X = x))$ not necessarily same as $P(Y|X = x)$.

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

Causal Effect

The Average Causal Effect ACE is

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

i.e. average difference in Y between setting $X = x_1$ and $X = x_2$.

Alternatively:

Consider Causal Odds Ratio or Causal Relative Risk—but Maths more difficult.

Identifiability

The ACE is identifiable if we can estimate it consistently from observational data.

Mathematically: ACE is identifiable if it can be re-expressed **without** $do(X)$ notation and only using distribution of **observable** variables.

Note: If sufficient confounders U are measured, it can be shown that the ACE can be identified as

$$\sum_u (E(Y|X = x_1, U = u) - E(Y|X = x_2, U = u))P(U = u)$$

—usual adjustment.

Identifiability via Mendelian Randomisation

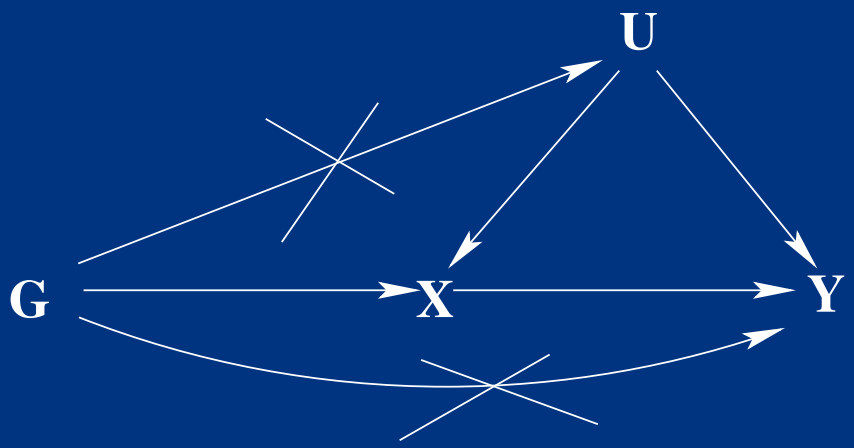
MR permits identification of ACE when genotype satisfies the properties of an **Instrumental Variable** (IV).

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

Note:

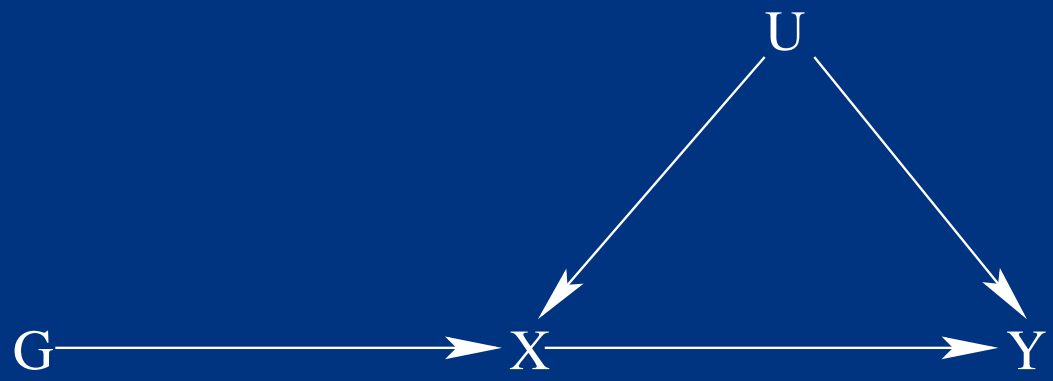
- No causal assumptions here
- Assumptions 1 and 3 cannot be tested without measuring $U \longrightarrow$ justification has to be based on background/subject matter knowledge.

Core Conditions—Graphically



- G does not affect Y other than through X;
- G is not associated with the unobserved confounders.

Core Conditions—Graphically



Equivalent to factorisation

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

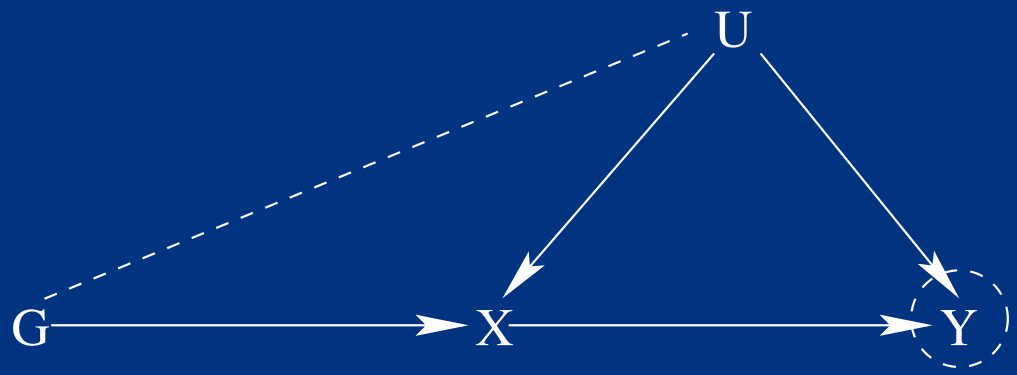
And under intervention in X

$$p(y, u, g|do(X = x_0)) = p(y|u, x_0))p(u)p(g)$$

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

Case-Control Scenario

Beware: Everything is conditional on Y.



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

$$\Rightarrow p(g, u|y) \neq p(g|y)p(u|y) \text{ despite } p(g, u) = p(g)p(u).$$

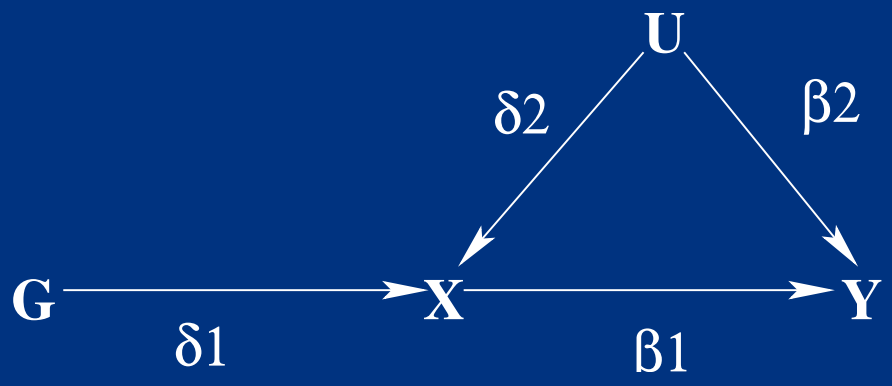
Selection effect: "moral" edge between G and U.

Results

If core conditions are satisfied

- Can test for causal effect by testing for an association between G and Y—Katan's original idea.
- For linear models without interactions, can find consistent point estimator for causal effect ACE.
- For binary/categorical variables, can find bounds on causal effect.
- With binary response, causal effect for subgroups of population can be estimated under certain model assumptions (**local causal effect**).

Graphical Illustration for Linear Case



- Wanted: β_1
- Regression of Y on G gives $\beta_1 \delta_1$
- Regression of X on G gives δ_1
- Obtain β_1 as ratio
- works only for linear / no–interactions case!

Mendelian Randomisation in Practice

Typically

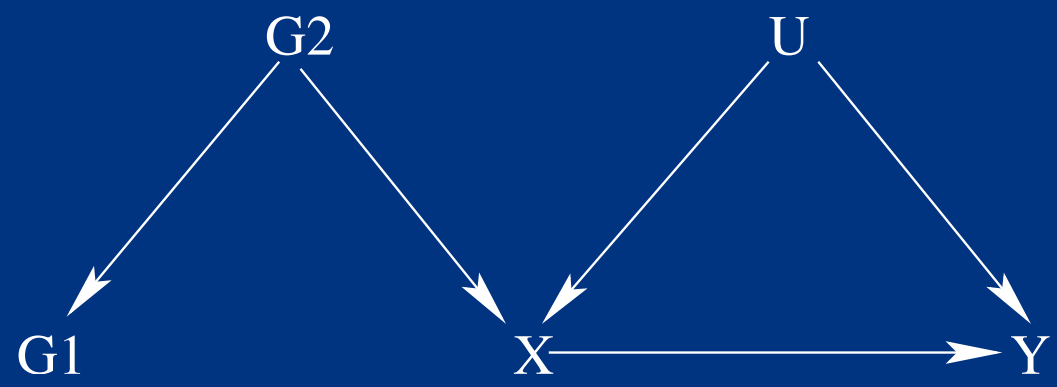
- G—Binary,
- X—Continuous,
- Y—Binary.

Therefore $p(y|x,u)$ is usually non-linear e.g. logistic.

Can't use ratio of regression coefficients to estimate causal effect of X on Y,

This has been misunderstood.

Chosen Gene is not "Causal"



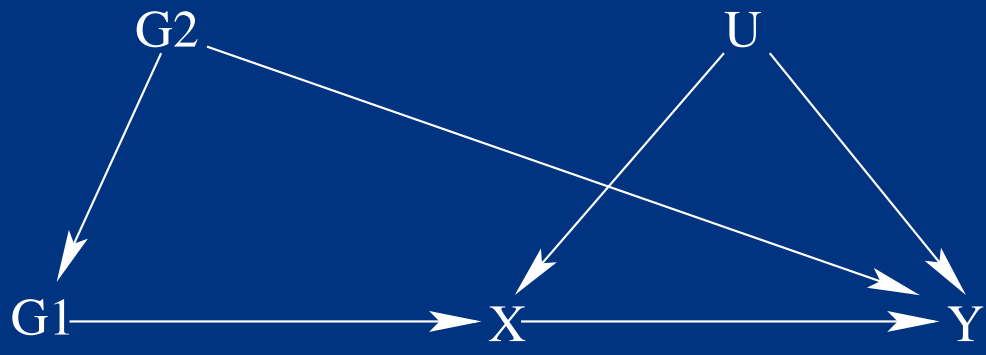
- G_2 is causal for X.
- G_1 and G_2 are associated.
- We are using G_1 as the Instrumental Variable.

$\Rightarrow$ **All Core Conditions are still satisfied!**

For our purposes, we do not have to find the "right" gene.

Linkage Disequilibrium

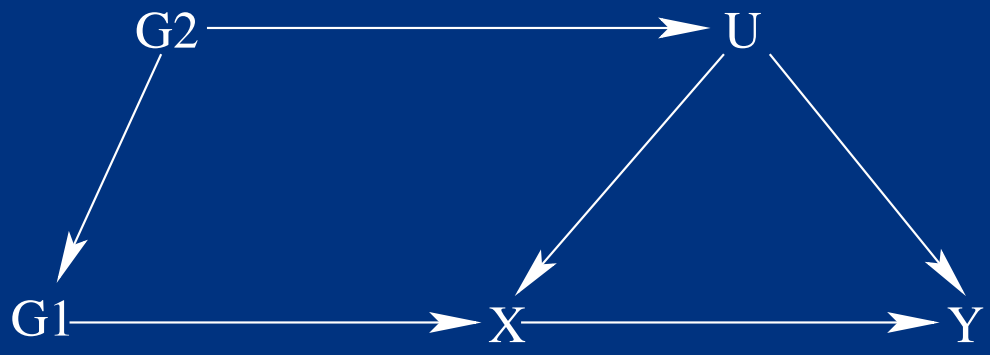
Population association between alleles at different loci.



(a)

- Chosen instrument in linkage disequilibrium with gene having a direct effect on Y.
- $Y \perp\!\!\!\perp G_1 | (X, U)$ is violated.

Linkage Disequilibrium

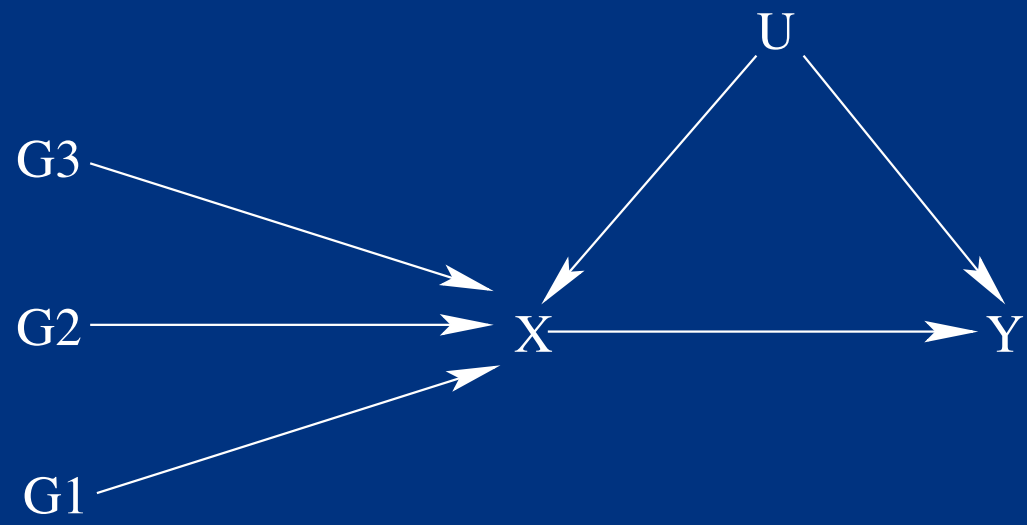


(b)

- Chosen instrument in linkage disequilibrium with gene affecting Y indirectly via confounders.
- $G_1 \perp\!\!\!\perp U$ violated.

Genetic Heterogeneity

More than one gene affects the phenotype.



- Okay if other genes are not both associated with G_1 *and* influence Y other than via X.
- Suspect weaker $G_1 - X$ association (bad instrument).

Conclusion

A formal causal framework is imperative for these epidemiological applications

- for a precise statement of what the relevant causal parameter is;
- to formalise the relationship between associational findings and causal implications in order to estimate this parameter.

Conclusion

- Causal inference always requires background knowledge for verification of necessary assumptions.
- Mendelian randomisation: background knowledge $\longleftrightarrow$ genetics.
- Hence, we can decide when IV assumptions are met by Mendelian randomisation.
- Can use this to test for and estimate the causal effect in situations where confounding is believed to be likely and not fully understood.

Problems and Open Questions

- Estimation requires additional (strong?) parametric assumptions which are unlikely to be satisfied in the case of a binary/categorical response.
- If all variables are binary, can only calculate bounds for the causal effect without making any assumptions besides core conditions.
- Causal parameters other than ACE are more difficult to identify.
- Estimation in retrospective studies more complicated—only odds ratio can be used.