# Mendelian Randomisation

James McMurray

PhD Student
Department of Empirical Inference

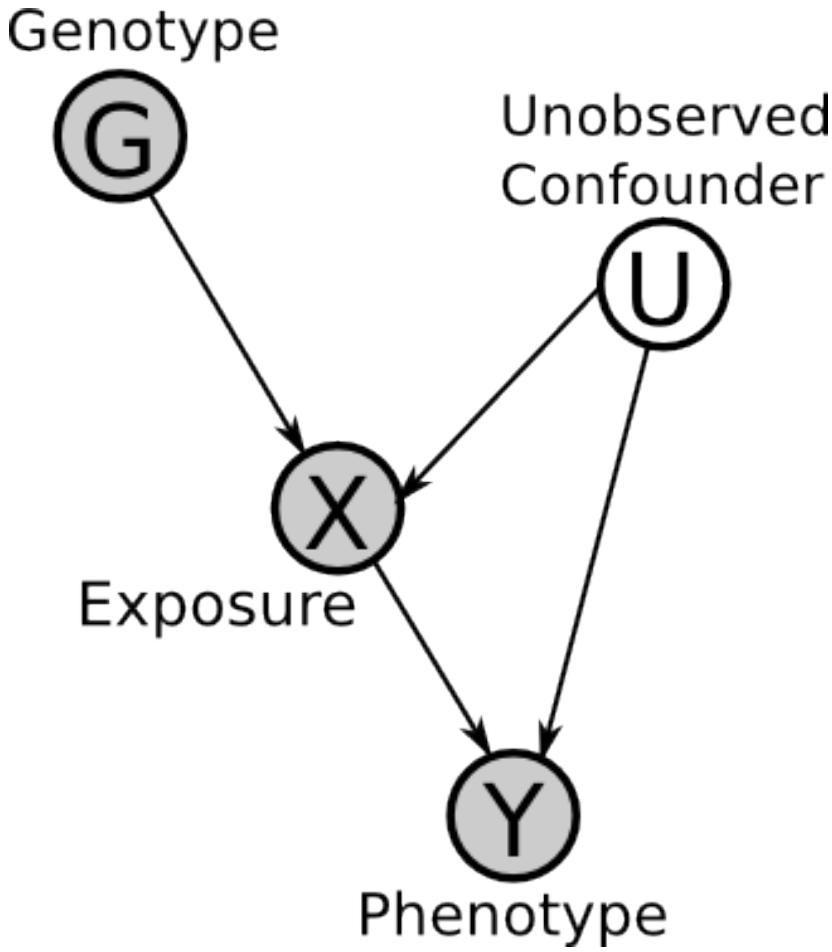
29/07/2014

# What is Mendelian Randomisation?

- Approach to test for a causal effect from observational data in the presence of certain confounding factors.
- Uses the measured variation of genes of known function, to bound the causal effect of a modifiable exposure (environment) on a phenotype (disease).
- Fundamental idea is that the genotypes are randomly assigned (due to meiosis).
- This allows them to be used as an instrumental variable.

MAX-PLANCK-GESELLSCHAFT

# What is Mendelian Randomisation?: DAG




MAX-PLANCK-GESELLSCHAFT

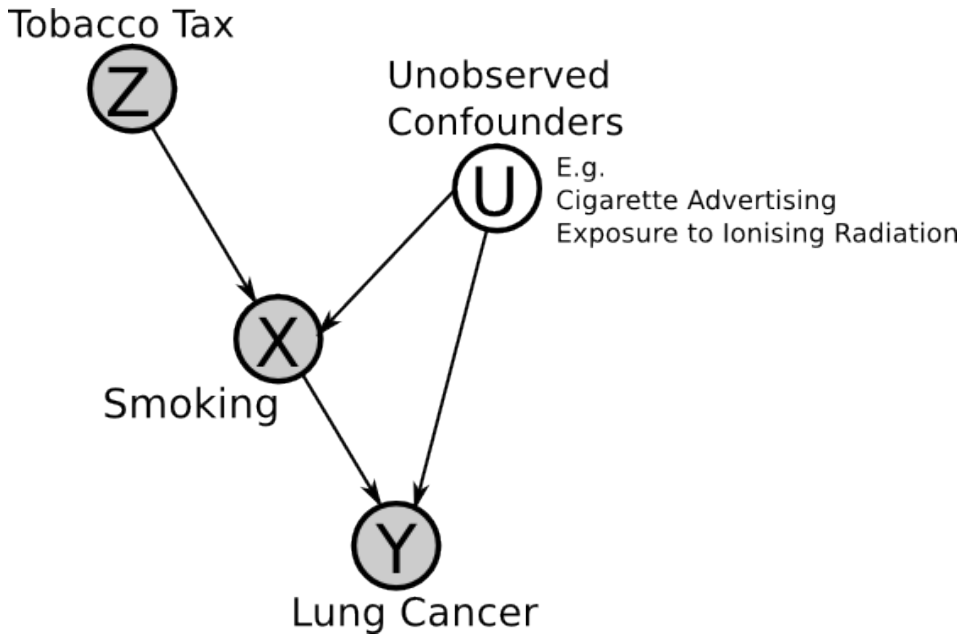
# Motivation: Why Observational Data?

- Randomised Control Trials (RCTs) are the gold standard for causal inference.
- However, it is often not ethical or possible to carry out RCTs.
- E.g. we cannot randomly assign a lifetime of heavy smoking and no smoking to groups of individuals.
- This leads to a need to use observational data.
- But this requires many assumptions.

# Instrumental Variables (IV)

- In the previous DAG, G is the instrumental variable (instrument).
- Because it affects Y only through X (exclusively)
- Therefore, under certain assumptions, if G is correlated with Y then we can infer the edge $X \rightarrow Y$
- First we will consider an example

# Instrumental Variables (IV): Example

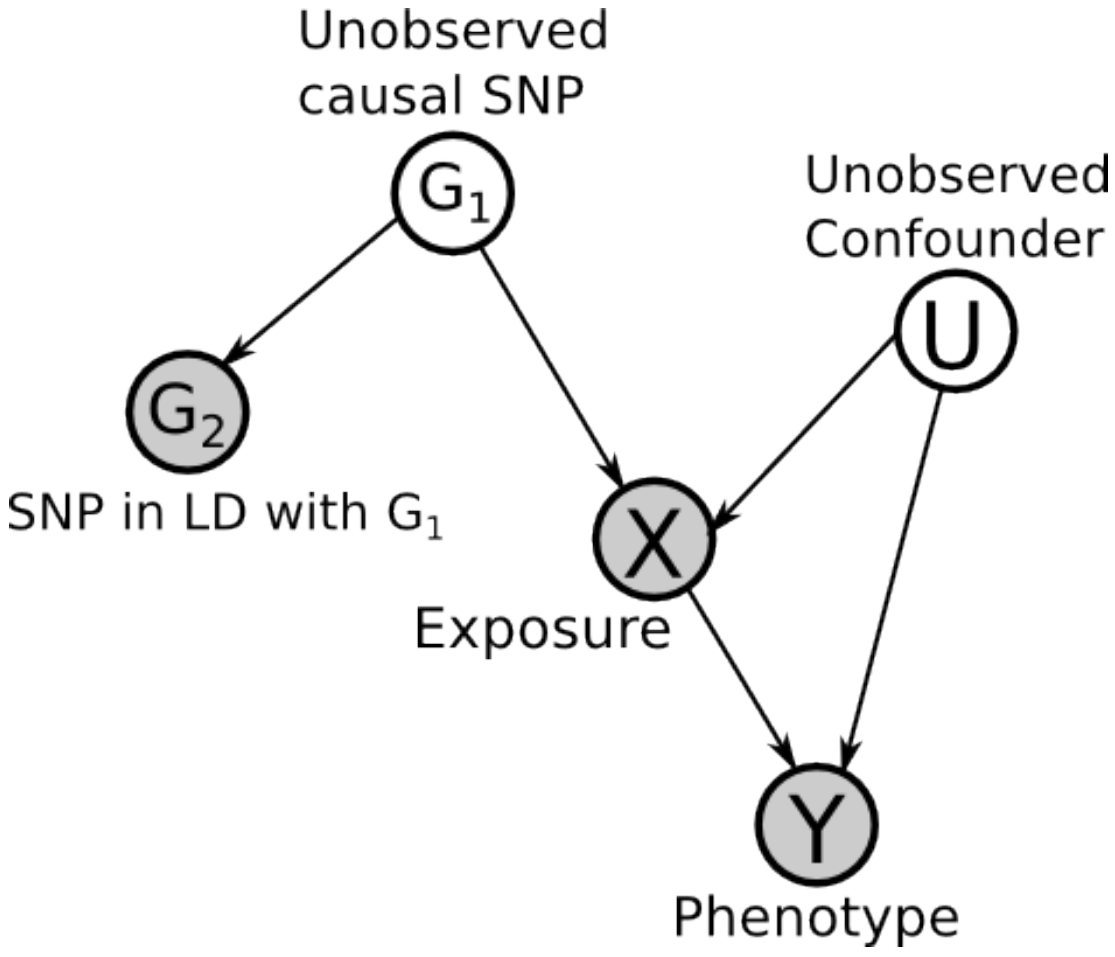



- If the assumptions are met, then if we observe that an increase in tax leads to a reduction in lung cancer, then one could infer that smoking is a "cause" of lung cancer (though possibly indirectly itself - i.e. it's not "smoking" per se, but the tar in the lungs, etc.).

# Instrumental Variables (IV): Assumptions: $Z \to X$

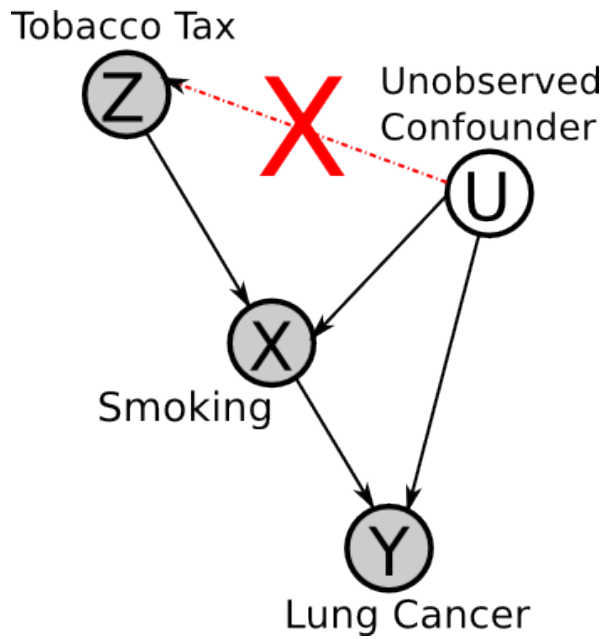
- We must know (a priori) that the causal direction is $Z \to X$ and not $X \to Z$.
- This is what makes the causal structure unique and identifiable.
- Note this does not mean that the Z we choose has to be the "true" cause of X. For example, if Z is a SNP, we can choose a SNP in linkage disequilibrium with Z, so long as it is independent of all the other variables but still correlated with X.
- The more correlated Z is with X, the better.

# Example: Can use associated SNP as instrument



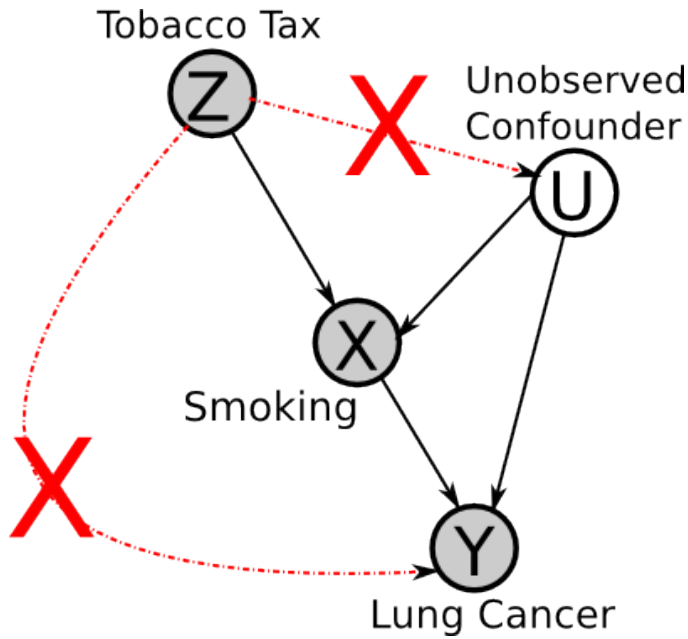
# Instrumental Variables (IV): Assumptions: $Z \perp\!\!\!\perp U$

- No factor can affect both the instrument and the effects. For example, there cannot be a factor that causes both higher taxes and less cancer (e.g. differences in health awareness in different countries).



# Instrumental Variables (IV): Assumptions: No $Z \rightarrow Y$

- Z cannot directly affect Y (or indirectly, except through X).
- I.e. there cannot exist other mechanisms by which Z affects Y (i.e. high tobacco tax increases substance abuse, leading to higher rates of cancer)

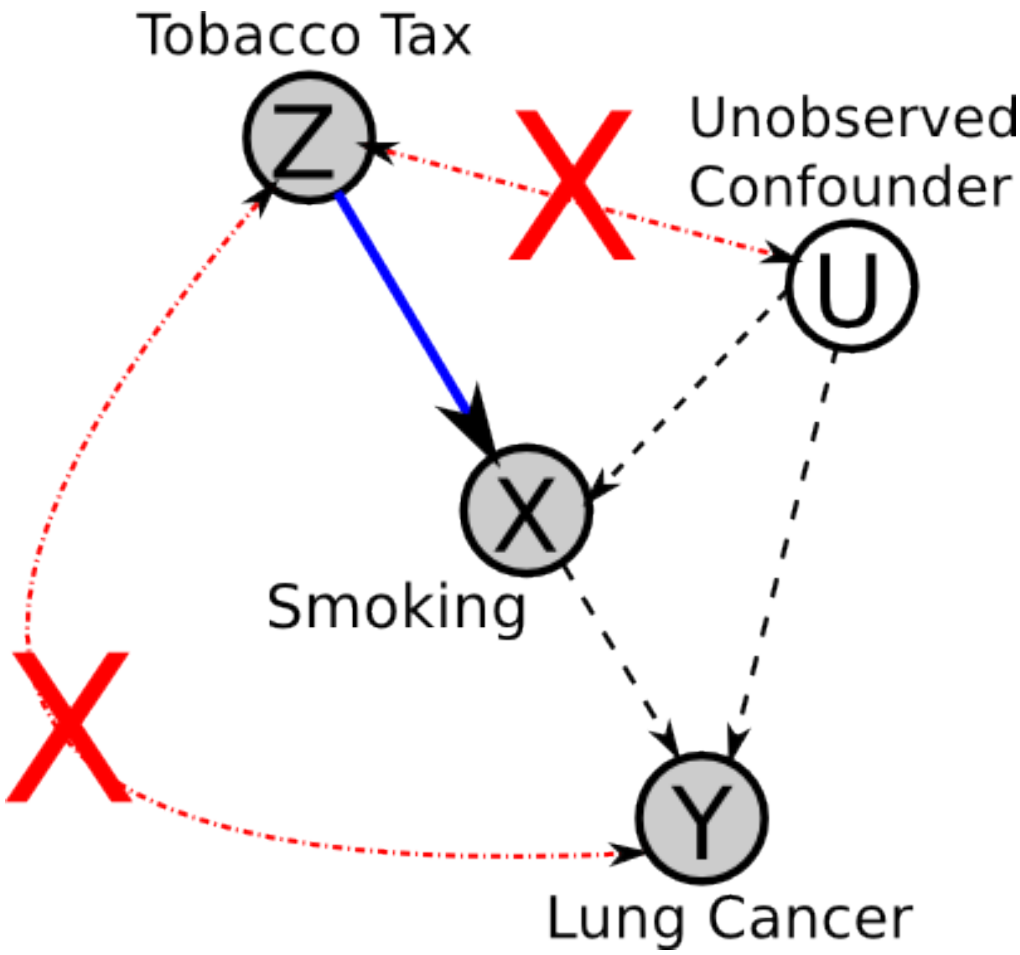


# Instrumental Variables (IV): Assumptions: Faithfulness

- Assume that the true underlying DAG manifests itself in the observed data
- I.e. causal effects do not cancel out
- Reasonable assumption, because the contrary would require very specific parameters
- But note that if relations are deterministic, implied Conditional Independencies do not hold and faithfulness is violated
- Note that, in practice, the sample size is important in testing the independencies in the data.

MAX-PLANCK-GESELLSCHAFT

# Instrumental Variables (IV): Assumptions: DAG



# What is Mendelian Randomisation?: Original example

- Katan MB (1986): *Apolipoprotein E isoforms, serum cholesterol, and cancer.*
- Do low serum cholesterol levels increase cancer risk?
- But maybe both cancer risk and cholesterol levels are affected by diet (confounders)
- Or latent tumours cause the lower cholesterol level (reverse causation)
- But patients with Abetalipoproteinemia (inability to absorb cholesterol) - did not appear predisposed to cancer
- Led Katan to idea of finding a large group genetically predisposed to lower cholesterol levels
- This is Mendelian Randomisation.
- Note this does not require that the genetic variants are direct determinants of health. But, uses the association to improve inferences of the effects of modifiable environmental risks on health.

# What is Mendelian Randomisation?: Original example

- Apolipoprotein E (ApoE) gene was known to affect serum cholesterol, with the ApoE2 variant being associated with lower levels.
- Many individuals carry ApoE2 variant and so have lower cholesterol levels from birth
- Since genes are randomly assigned during meiosis (due to recombination), ApoE2 carriers should not be different from ApoE carriers in any other way (diet, etc.), so there is no confounding via the genome - note these assumptions.
- Therefore if low serum cholesterol is really causal for cancer, the cancer patients should have more ApoE2 alleles than the controls - if not then the levels would be similar in both groups.

MAX-PLANCK-GESELLSCHAFT

# What is Mendelian Randomisation?: Original example

- Katan only provided the suggestion, but the method has since been used for many different analyses with some success, such as the link between blood pressure and stroke risk.

- However, some conclusions have later been disproved by Randomised Control Trials. To understand why, we must consider the biological assumptions.

MAX-PLANCK-GESELLSCHAFT

# Panmixia

- Recall the assumption that the genotype is randomly assigned - this implies panmixia
- That is, there is no selective breeding (so random mating)
- Implies that all recombination is possible
- In our DAG, this means that G is not influences by Y (or other variables)
- Not entirely accurate, as demonstrated by Population Stratification

# Population Stratification

- Systematic difference in allele frequencies between subpopulations, due to ancestry
- For example, physical separation leads to non-random mating
- Leads to different genetic drift in different subpopulations (i.e. changes in allele frequency over time due to random sampling)
- Means that the genotype is not randomly assigned when considered across sub-populations
- E.g. Lactose intolerance

MAX-PLANCK-GESELLSCHAFT

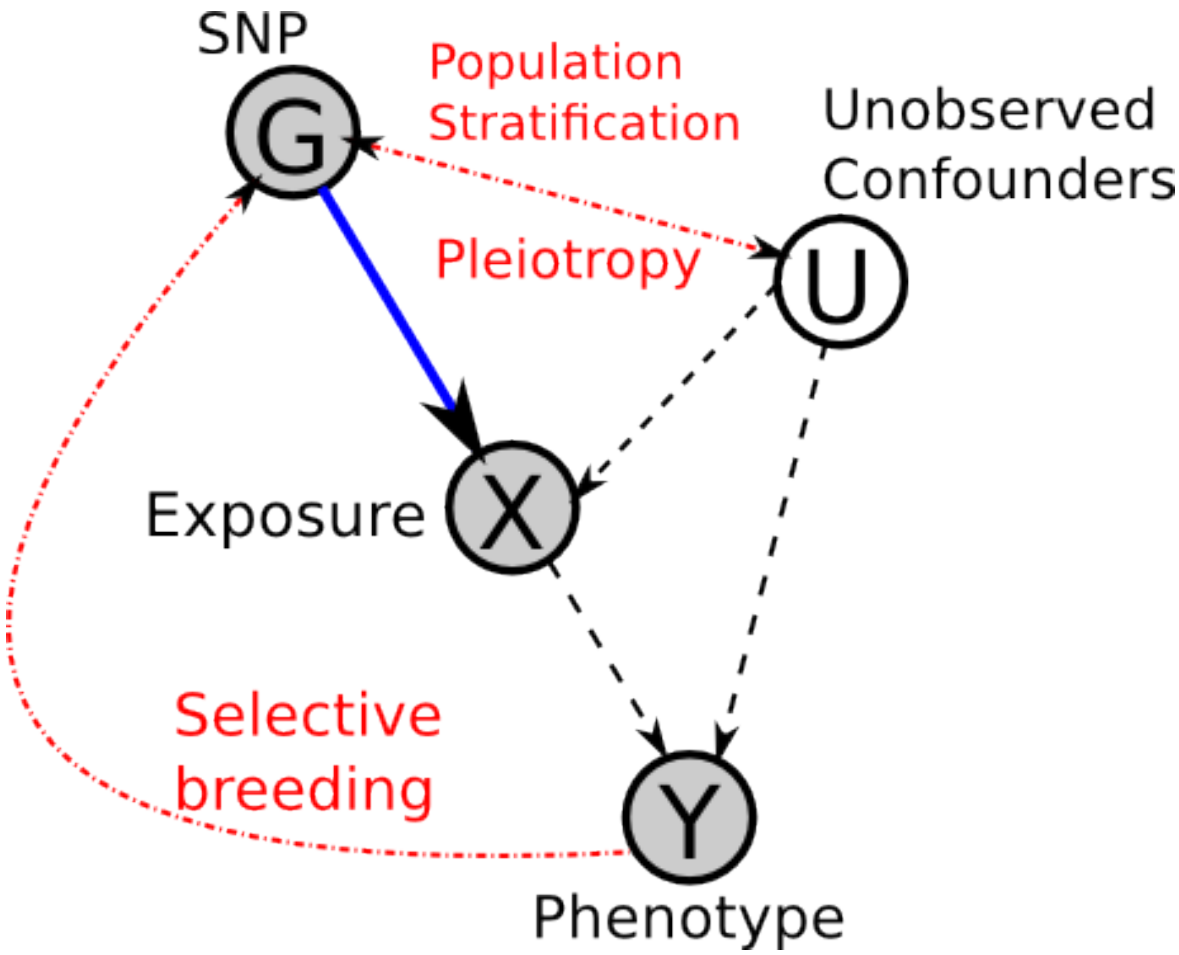
# Canalization

- Variation in robustness of phenotypes to genotype and environments
- Waddington Drosophilia experiment:
  - Exposed Drosophilia pupae to heat shock
  - Developed Cross-veinless phenotype (no cross-veins in wings)
  - By selecting for this phenotype, eventually appears without heat shock
  - Led to theory of organisms rolling downhill in to "canals" of the epigenetic landscape with development, becoming more robust to variation
  - Think of it like an optimisation problem
- Exact mechanisms unknown
- Acts as confounder between genotype, environment and phenotype

# Pleiotropy

- One gene can affect many (even seemingly unrelated) phenotypes
- Mendelian Randomisation makes the assumption of no pleiotropy
- In this case, this means that we know the genotype is only influencing the phenotype via the considered exposure
- I.e. ApoE2 only affects serum cholesterol levels, and cannot affect cancer risk by other, unobserved means.
- This is a big assumption, prior knowledge is necessary.
- If possible, using multiple, independent SNPs (instruments) helps to alleviate this issue (as if they are all consistent then it is unlikely that they all have other pathways causing the same change) - but note they must not be in Linkage Disequilibrium!

# The real underlying DAG?




MAX-PLANCK-GESELLSCHAFT

# Conclusion

- Instrumental variables are a method to infer causal relations from observational data, given certain assumptions.
- Applied in Genetic Epidemiology with Mendelian Randomisation.
- Has had some success, but underlying biology poses problems.
- Can we improve robustness with more measurements of intermediate phenotypes? (gene methylation, RNAseq, proteomics) - multi-step Mendelian Randomisation
- Can we improve identification of appropriate instruments? (e.g. whole genome sequencing makes it easier to identify population stratification)

Thanks for your time

Questions?

MAX-PLANCK-GESELLSCHAFT