Mediation Analysis via Mendelian Randomization: Searching for Causality in Observational Studies

General background reading: Mendelian Randomization, Methods for Using Genetic Variants in Causal Estimation by S. Burgess and S. G. Thompson (2015).

Outline

- Observational and Experimental Studies
- Randomized Clinical Trials and Instrumental Variables (IVs).
- Mendelian Randomization
  - Estimating Causal Effects using Mendelian Randomization.
  - Potential Problems
- Mediation Analysis more generally
- Conclusions
- Extra: A published example for illustration.

Observational versus Experimental Studies

- **Experimental studies** are designed to understand causes and effects and **make causal inferences** by directly manipulating the amount of an exposure and which groups receive it.
- **Observational studies** also seek to understand cause and effects. However, unlike experiments, the researcher is not able to control how subjects are assigned to groups or what treatments they receive.
- Causality is harder to determine in observational studies due to residual confounding and the possibility of reverse causality.
- **Mediation analysis** is designed to help determine the pathway of cause to effect.

Some Correlations

- Individuals on Medicaid have worse health outcomes than individuals without health insurance in the US.
- Individuals with Parkinson disease are less likely to be smokers than controls.
- Stork populations and human birth rates are correlated in Europe.

Inferring Causality from these Studies?

The New York Times | https://nyti.ms/2tDUzgB

The Upshot

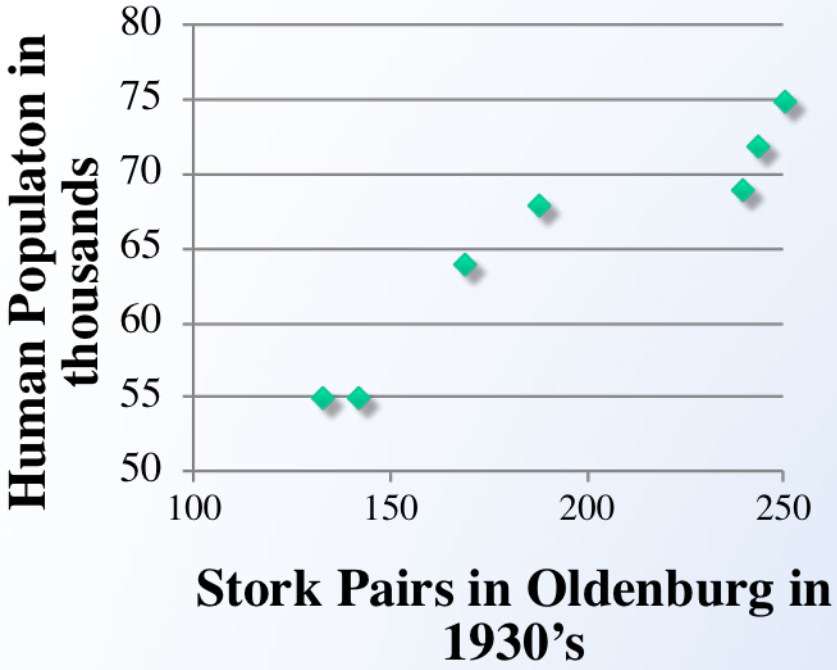
Medicaid Worsens Your Health? That's a Classic Misinterpretation of Research

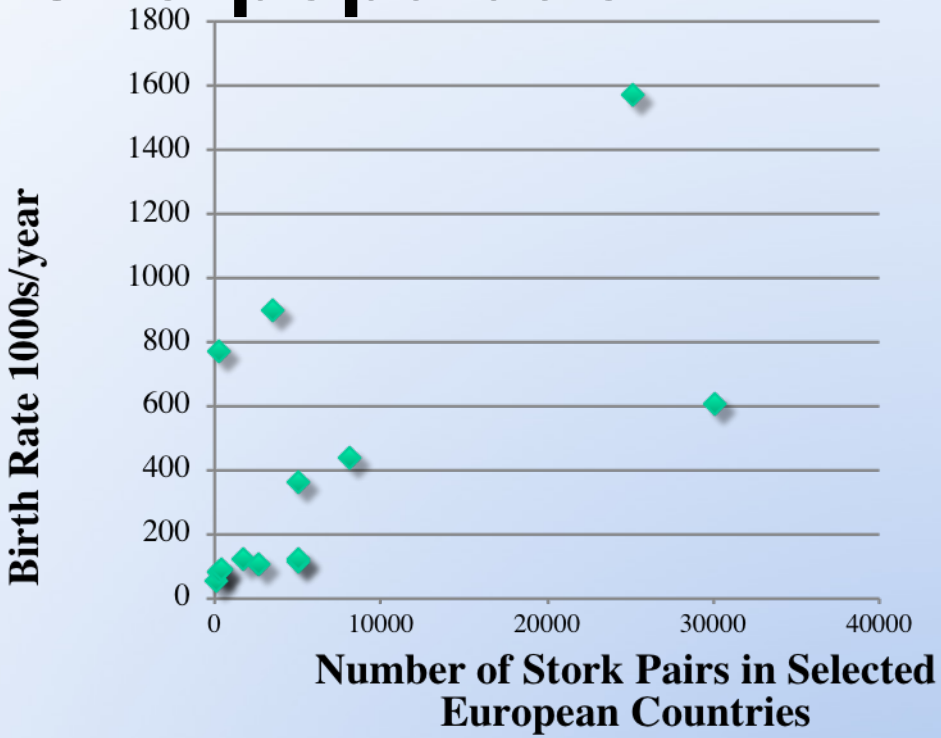
The New Health Care

By AARON E. CARROLL and AUSTIN FRAKT JULY 3, 2017

- Smoking is protective of Parkinson's disease (so smoke if you are at risk)?
- AND …

Reducing the Number of Storks will reduce Human Overpopulation





Correlation does not imply Causation

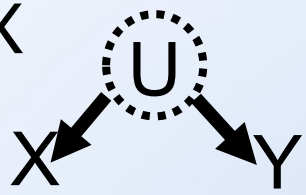
- Alternative?
- **Reverse Causality:** The outcome is actually the predictor. That is:

  $Y \longrightarrow X$ not $X \longrightarrow Y$

- For our examples, possible reverse causality:
  - People can sign up for Medicaid retroactively, so becoming ill can lead to Medicaid enrollment.
  - Individuals with (subclinical) Parkinson's may stop smoking because it makes them feel worse.
  - Babies bring Storks?

Correlation does not imply Causation Continued

- **Confounding:** A third variable (or set of variables) U influences both Y and X



  - Medicaid enrollees are of lower socioeconomic status than even the uninsured and so less opportunity for good health practices.
  - Unknown/unmeasured confounder in the case of Parkinson Disease and Smoking
  - Amount of rural/agricultural area, storks and babies.

Which Genomic Correlations are Causal?

R.C. Richmond, G. Hemani, K. Tilling, G. Davey Smith, C.L. Relton
R150 | *Human Molecular Genetics, 2016, Vol. 25, No. R2*

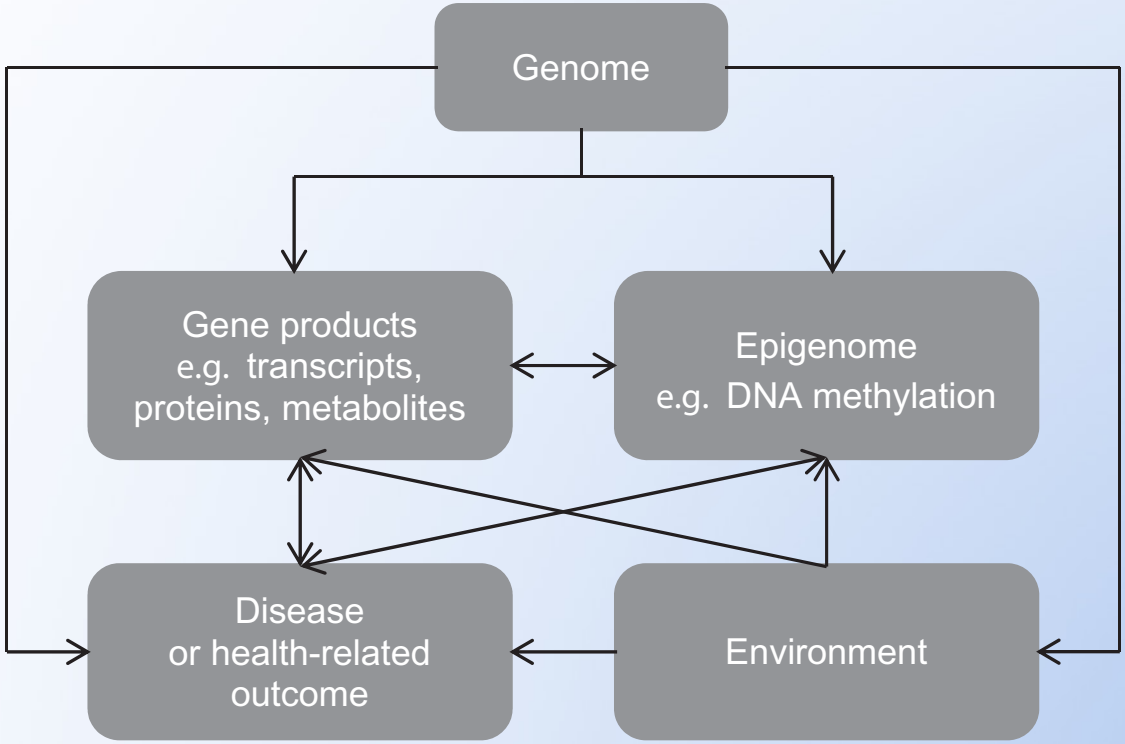
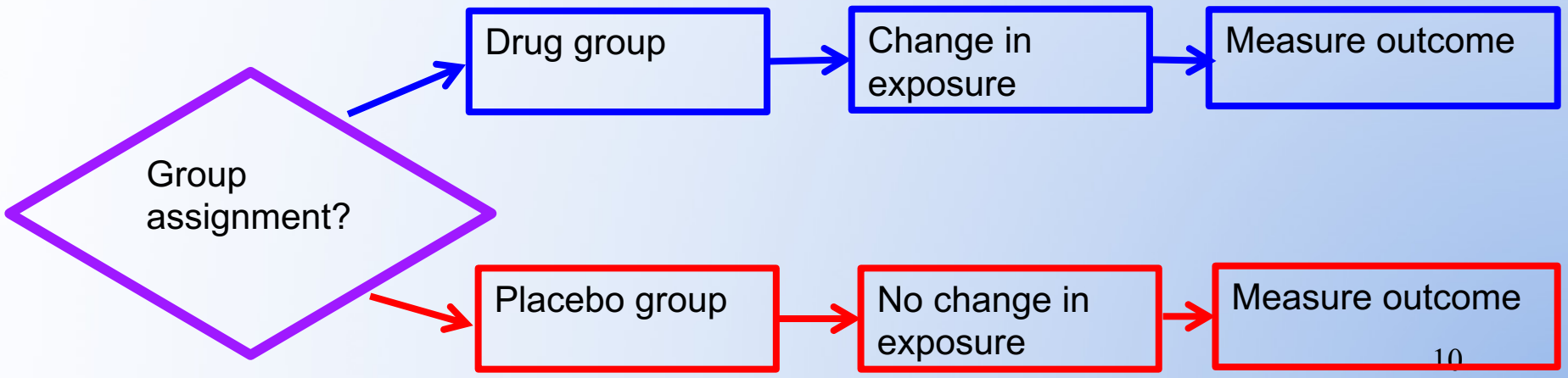


Figure 1. The interplay between genomics, other "omics" and environmental factors in relation to disease or health-related outcomes.

Medical Research has Relied on Randomized Clinical Trials to Demonstrate Causality

- How do RCT work?
  - Randomly assign individuals to treatment groups, then prospectively determine if groups differ in their outcomes.
  - Determine the average effect of being assigned to the treatment group versus control group as an estimate of the causal effect in of treatment in the population.



Why do Randomized Clinical Trials allow Causal Inference?

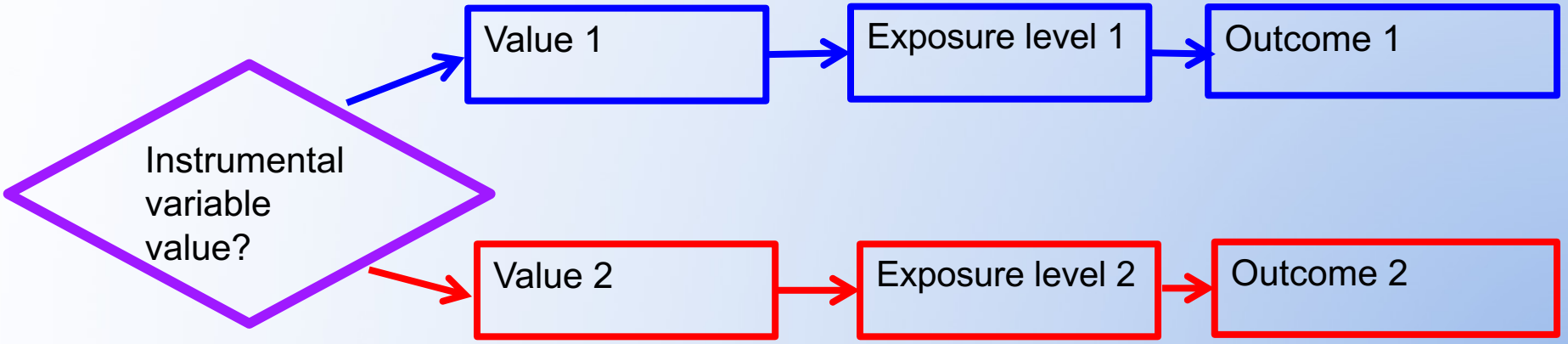
- In well designed randomized clinical trials:
  - Assignment affects treatment but is not directly influencing outcome.
  - Random assignment insures that the two groups the same in terms of levels of other risk factors (confounders) and thus exchangeable.

Problems with Randomized Clinical Trials

- Not always practical. As examples:
  - May need to follow up subjects for a long period of time.
  - If the outcome is onset of a rare disease then very few at risk will ever have the outcome
- Not always generalizable to the population
  - Subjects tend to be healthier, more motivated, more compliant than the general population.
  - Intent to treat is not the same as the “as treated” effect.

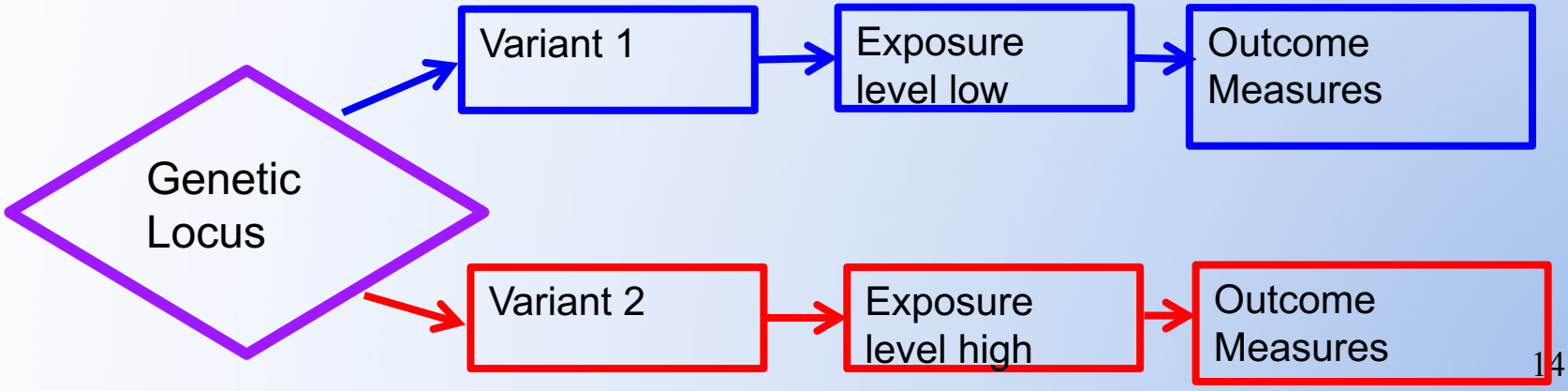
Instrumental Variable Analysis is an Example of Mediation Analysis

- Find a variable that is correlated with the exposure but does not influence the outcome (except indirectly through the exposure level).
- This variable should not be associated with any confounders of the exposure-outcome association.
- In effect, find a “natural experiment” where some variable has randomized individuals to exposure groups that are exchangeable except for exposure level.
- Differs from classical RCT in that we want to determine the causal effect of exposure on outcome using the association of the IV with exposure and the association of IV with outcome.



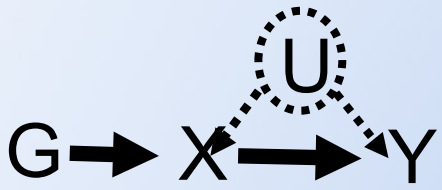
Mendelian Randomization

- Choose a genetic locus as the IV such that:
  - It is correlated with the exposure but does not influence the outcome (except indirectly through the exposure level).
  - It is not associated with any confounders of the exposure outcome association.
- The division of the population into subgroups by variant is independent of competing risk factors and so these groups are exchangeable.
- Use the association of variant with exposure and the association of variant with outcome to estimate the causal effect of exposure on outcome.



The Mendelian Randomization Model represented as a Directed Acyclic Graph (DAG)

- G denotes the locus (or loci)
- X the exposure
- U (possibly unmeasured) confounder(s)
- Y the outcome



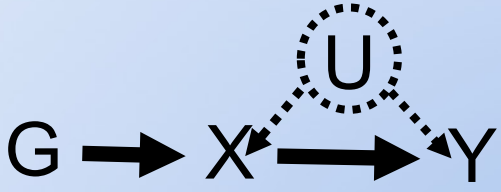
- Acyclic because there are no feed back loops.

Estimating Causal Effects with MR

- Additional assumptions
  - SUTVA (stable unit treatment value assumption). Outcome is not affected by how treatment was assigned and an individual's outcome depends only on his/her risk factors and treatment (not anyone else's).
  - Monotonicity: the value of the IV should effect at least one person's exposure. All those affected are affected in the same direction.
- A number of analysis methods exist. (see e.g. V. Didelez et al. (2010) Stat. Sci. 25:22-40 or S. Burgess and S. Thompson (2015) chapter 4 for reviews)

Examples of MR Analysis Methods:

- The form of the estimates depends on whether the outcome is continuous or dichotomous, the exposure is continuous or dichotomous, the IV is polychotomous or dichotomous, and the statistical approach.
- Generalized linear model based:
  - Ratio of coefficients (Wald type statistics)
  - Two stage methods
  - Likelihood and Bayesian Methods
- Semi-parametric Methods.



Estimates from Ratios

- When the outcome and exposure are continuous, then the average causal effect can be estimated as

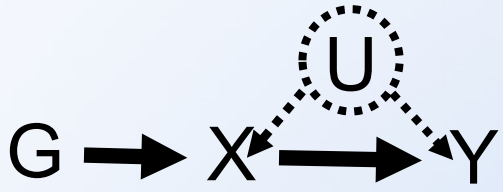
$$\beta = Cov(Y,G)/Cov(X,G).$$

- When Y is dichotomous and exposure is continuous, then the causal odds ratio (COR) is approximated as

$$\log(COR) = \log(OR_{Y|G})/\beta_{X|G} \Rightarrow COR = OR_{Y|G}^{1/\beta_{X|G}}$$

- These Wald like estimates let us use summary statistics, however they require large sample sizes to be accurate estimates and are subject to bias.
- Standard errors and confidence intervals are approximate.

Two Stage Regression



- Regress X on G then regress Y on E(X).
- If outcome, Y, is continuous then regression is linear if not use generalized linear models.
- With a single IV and continuous Y get same estimate as the ratio method but model extends so that multiple IVs (multivariate regression) can be used simultaneously.

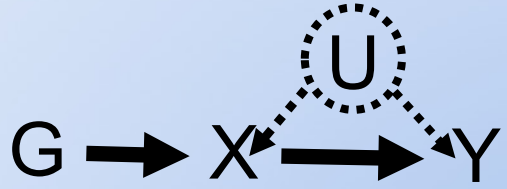
Likelihood and Bayesian Approaches

- Model X and Y jointly and find maximum likelihood estimates for the parameters.
- Simple likelihood example:
- Let i =1, …, N individuals, let X and Y be continuous, and let there be K unlinked loci.
- The model:

$$x_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} + e_{x_i}$$

- $$y_i = \beta_0 + \beta_1 x_i + e_{y_i}$$

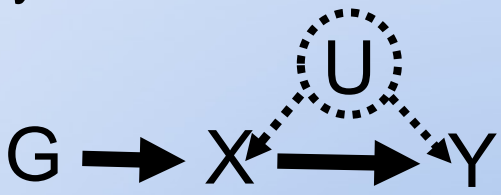
$$\begin{pmatrix} e_X \\ e_Y \end{pmatrix} \sim N(0, \Omega)$$



- Where β_1 is the causal parameter.

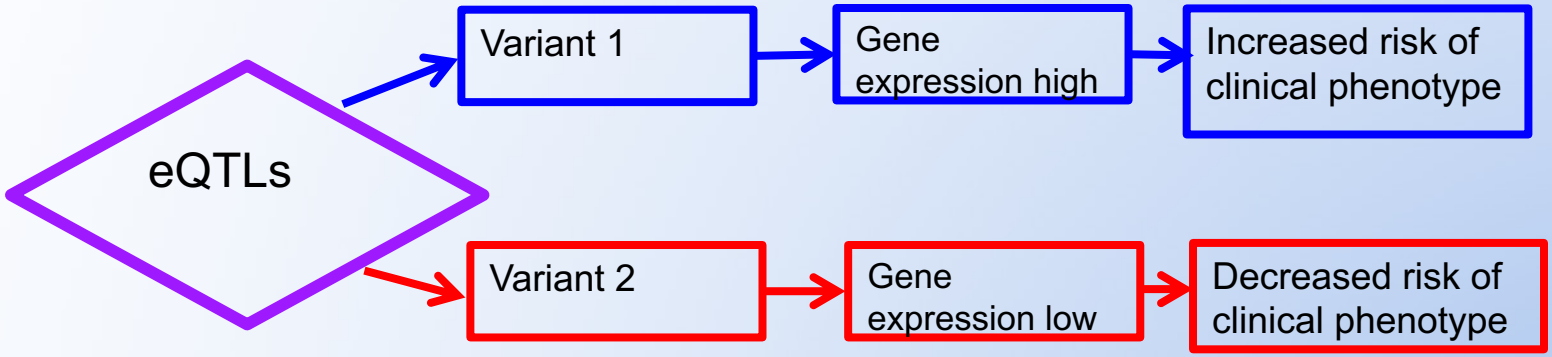
Potential Problems with Mendelian Randomization

- "Assignment" of alleles not random: e.g. assortative mating or selection with regards to the locus used as an IV or ascertainment induced (variant effects likelihood of being in the study, e.g. survivor bias).
- Pleiotropy effects of the IV induce another connection between the locus and the outcome.
- Association between locus and exposure is weak. There are issues both with loss of power and upwardly biased estimates. Bigger sample sizes, covariate adjustment and combining studies through meta analysis can help.
- Developmental compensation (Canalization) can reduce the exposure difference for individuals with different variants.
- Bias introduced if assumptions not met and it is difficult to assess validity of modeling assumptions.



Mendelian Randomization with Genomic Data

- Question of interest: Can Mendelian randomization help us understand the causality of pathways implicated with expression data?



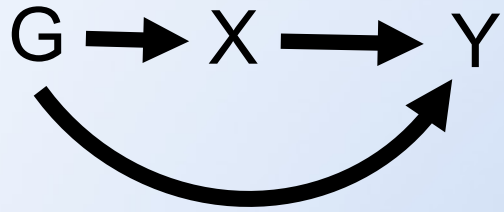
- Do assumptions of MR apply? What are the pitfalls? This is an active area of research.

MR with Family Data?

- Yes, with linear mixed models. see computer exercises for details.

Note: Mediation Analysis is a much Bigger Field than just Mendelian Randomization

- Forms of Mediation Analysis cover questions MR can't address. E.g. suppose we are interested in determining the extent that genetic variation at a locus (the exposure, G) acts on a phenotype (outcome, Y) through a molecular intermediate (mediator, X)



- MR is no longer the appropriate approach. G is no longer an IV because we allow for a direct effect.
- A number of methods exist to address this question. It can be tackled in a generalized linear regression framework. Example: Natural Effect Models.

Summary

- Mendelian randomization provides a way to assess causality in observational studies.
- Randomized clinical trials may provide better evidence but are not always possible to perform.
- Care must be taken in the selection of the loci and in the analysis methods.
- Mediation analysis and causal inference methods abound in epidemiology – MR is just one example.

A Few Select References – (there are a lot more; it's a big field)

- Mendelian Randomization
  - S.Burgess and S. G.Thompson (2015). Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation.Chapman & Hall/CRC Press.
  - I. Pichler et al. (2013) Serum iron levels and the risk of Parkinson's disease: a Mendelian randomization study. Plos Medicine 19:10.1371
  - K. Baicker and A. Chandra (2017) Evidence-Based Health Policy. N Engl J Med 377:2413-2415
  - V. Didelez etal (2010). Assumptions of IV methods for observational epidemiology. Stat.Sci. 25:22-40.

- Mediation analysis applied to genomic data
  - R.C. Richmond et al. (2016) Challenges and novel approaches for investigating molecular mediation. Hum. Mol. Genet. 25(R2):R149-R156
  - R. Barfield et al. (2017) Testing for the indirect effect under the null for genome-wide mediation analysis. Genet. Epid. 41:824-833.

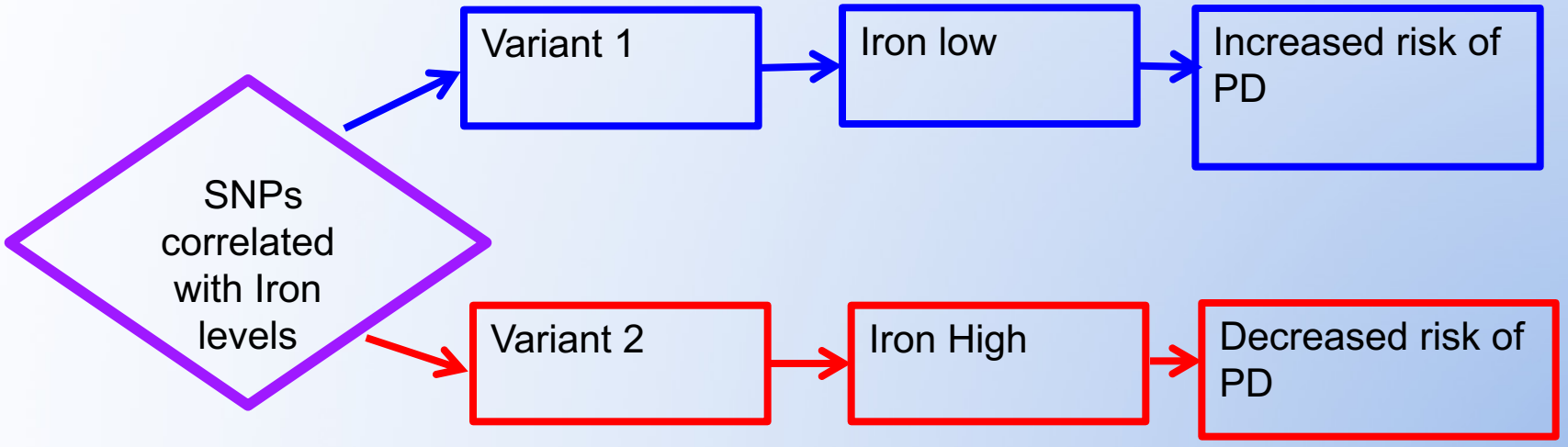
- Mediation analysis and Causal Inference Theory
  - Judea Pearl, (2014) Interpretation and Identification of Causal Mediation. Psych. Methods 19:459-481
  - Judea Pearl (2009) Causality: models,, reasoning and inference. Camb. U. Press.
  - T.J. VanderWeele (2015) Explanation in causal inference: methods for mediation and interaction. Oxford. U. Press.
  - T.J. VanderWeele (2016) Mediation Analysis: A Practitioner's Guide. Anm Rev. Pub. Health. 37:17-32

Illustrative Mendelian Randomization Study

- Serum iron levels and risk of Parkinson's disease. I. Pichler et al. (2013) PLOS Medicine 10:e1001462
- Prior evidence from Observational Studies is confusing:
  - Autopsy study in which increased iron found in PD brains versus unaffected brains.
  - Most but not all, case (Parkinson's patients) - control studies of serum Iron levels show reduced levels in cases.

Iron - Parkinson's Disease Continued

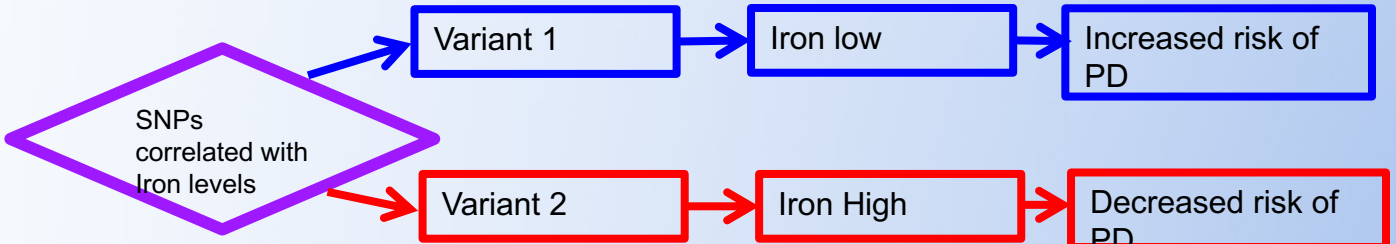
- Pichler et al. used three variants: two in *HFE* (not in LD) and one in *TMPRSS6*.
- Effect of the variant on serum iron comes from GWAS meta analysis. Effect of the variant on PD risk from meta analysis of GWAS and candidates studies. Serum iron and PD risk GWAS were different.



Iron-Parkinson's Example Con't.

- For each locus:
  - Iron levels = quantitative exposure, standardized so mean = 0, SD = 1.
  - PD = dichotomous
  - locus = counts (0,1,2) of variant associated with increased iron. Assumed independence of allelic effects.
  - Meta-analysis estimates of (1) per allele OR for PD and (2) per allele in increase in iron values (number of SDs).
  - Use Wald type estimate to get the MR estimate,

    $\log(OR_{PD|iron}) = \log(OR_{PD|allele})/\beta_{iron|allele}$
- Pool MR estimates across the three loci.
  - MR estimate = 0.88 (approximate 95% CI, 0.82 – 0.95, pvalue =0.001)
  - Corresponds to 0.3% relative reduction in PD risk per 1 μg/dl increase in serum iron over lifetime (Small effect size).



Iron-Parkinson's Example Con't.

- Validity checks conducted:
  - All three loci assessed to be strong IVs.
  - Three IVs show similar results - No evidence of heterogeneity (suggesting the no pleiotropy assumption holds).
  - Sensitivity analysis (by excluding particular studies used in meta-analyses) in order to determine if population stratification a factor provided similar results to the primary analysis.