Sketching alternate realities: An introduction to causal inference in genetic studies

Webinar for Quantitative Genetics Tools
NIDA Center of Excellence in Omics, Systems Genetics, and the Addictome

Saunak Sen
Professor and Chief of Biostatistics
Deparment of Preventive Medicine
University of Tennessee Health Science Center
sen@uthsc.edu ~ @saunaksen ~ http://www.senresearch.org

2020-11-20

Please mute your speakers to reduce any ambient noise that may interfere with others hearing.

Please ask questions using the chat option at the bottom of your screen.

Goals

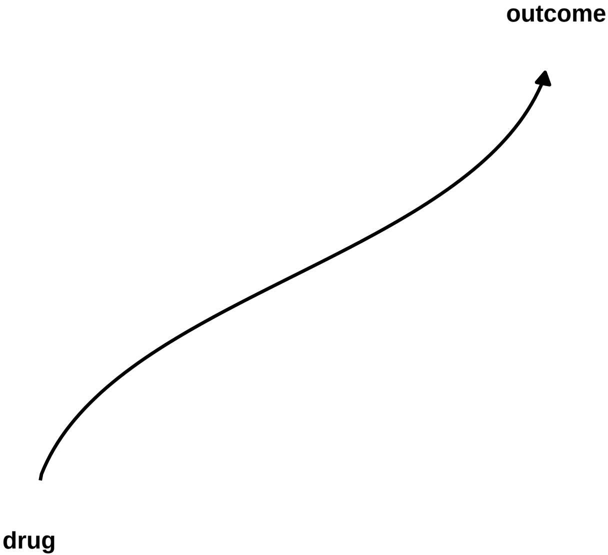
Express causal inference as a missing data problem

Outline assumptions needed for causal inference

Express causal information as (directed acyclic) graphs

Outline how to use graphs to guide analytic strategy

Drug trial



outcome

drug

Potential outcomes (counterfactual) framework

id	placebo	diuretic	effect
ind1	140	140	0
ind2	140	130	-10
ind3	145	135	-10
ind4	150	130	-20
ind5	135	105	-30
ind6	130	110	-20
inf7	125	135	10
ind8	155	155	0

The goal is to estimate the average causal effect using the potential outcomes for each individual.

Only one of the two potential outcomes is observed

id	placebo	diuretic
ind1		140
ind2	140	
ind3		135
ind4	150	
ind5		105
ind6		110
inf7	125	
ind8	155	

Only one of the two potential outcomes is observed

id	placebo	diuretic
ind1	140	
ind2		130
ind3		135
ind4		130
ind5		105
ind6	130	
inf7	125	
ind8	155	

Only one of the two potential outcomes is observed

id	placebo	diuretic
ind1	140	
ind2		130
ind3	145	
ind4	150	
ind5		105
ind6	130	
inf7		135
ind8		155

Estimation of causal effect

Problem: To estimate causal effect, we have to impute the missing (counterfactual) data.

Difficult problem, and needs assumptions.

The more variables involved, the more missing data we have to essentially "impute", increasing problem difficulty.

Easier to estimate average causal effects than individual causal effects.

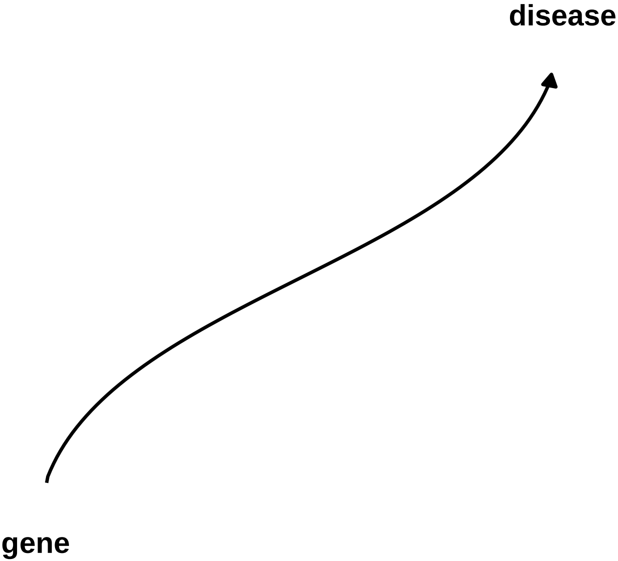
Estimation of causal effect

Average causal effect can be estimated if we assume that treatment assignment was independent of potential outcomes.

Assume no interference between units: Potential outcomes of a unit does not depend on other units' treatment values.

Other issues: selection bias, non-compliance, differential dropout, measurement error.

Gene-disease



disease

gene

Gene-disease associations

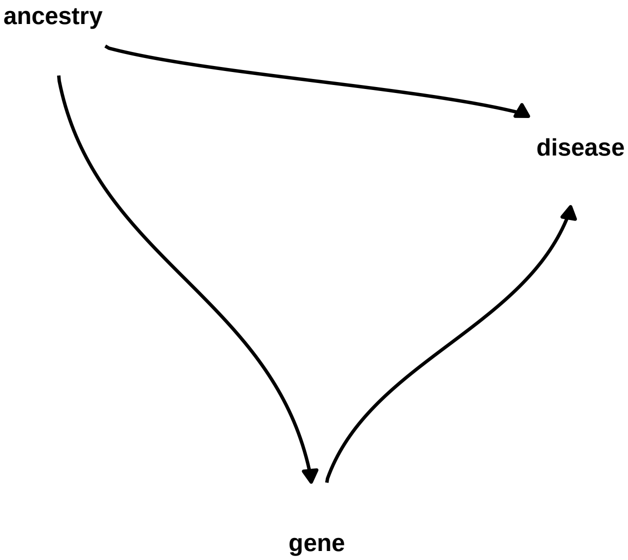
One may think of the gene (genotyope at a locus) as being randomized.

Potential confounding by ancestry, or family structure.

Selection effects if only fit individuals survive.

Potential interference there is competition between littermates.

Confounding by ancestry



Diabetes and Pima/Pipago ancestry

$Gm^{3;5,13,14}$ haplotype status found to be marginally associated with diabetes in Pima Indiams.

Haplotype	N	Diabetes (%)
Present	293	23 (8)
Absent	4,627	1,343 (29)

But... haplotype frequency depends on Pima ancestry, and diabetes prevalence depends on ancestry. Thus gene "assignment" depends on potential diabetes status.

Diabetes and Pima/Pipago ancestry

$Gm^{3;5,13,14}$ haplotype association dissapears when adjusted by ancestry (and age).

524

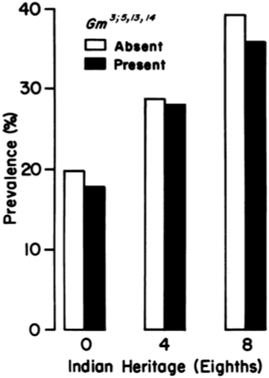
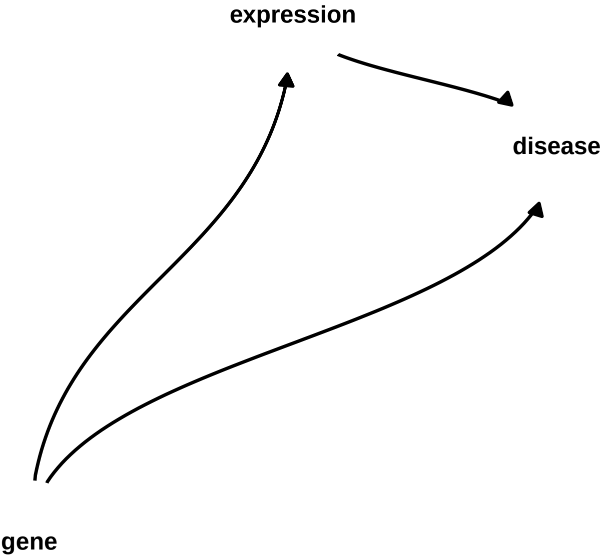


Figure 2 Age-adjusted prevalence of diabetes by the presence of the haplotype $Gm^{3;5,13,14}$, according to Indian heritage, among residents of the Gila River Indian Community.

Directed acyclic graphs (DAGs)

AKA: Bayesian networks, causal graphs, belief networks, decision networks



Directed acyclic graphs (DAGs)

Directed: Edges imply direction of causality

Acyclic: No cycles

Nodes/vertices: Variables (observed or unobserved)

Edges/arrows: Causal effect

Causal Markov assumption: Conditional on its direct causes, a variable is independent of any variables for which it is not a cause.

Implies that you can factor the joint distribution of the variables

$$p(v) = \prod_{j=1}^{M} p(v_j | causes(v_j))$$

Non-parametric framework, as conditional distributions can be arbitrary; pieces can be learned by statistical learning models.

Direction of causality

The graph

$$x \to y$$

implies

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

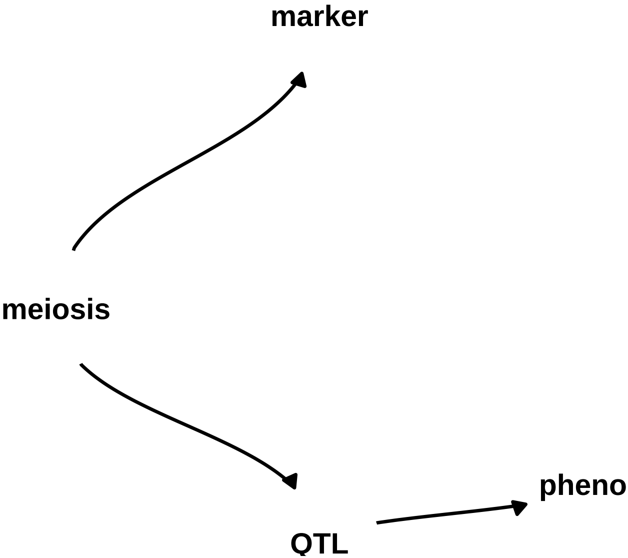
which is consistent with the graph

$$x \leftarrow y$$

Thus, it is not possible to infer causation from a sample of x and y alone without additional assumptions/information.

Note this is non-parametric, and $p(y|x)$ or $p(x|y)$ can be very general and we can use complex methods such as machine learning.

QTL study



QTL study

- Phenotype: y
- QTL: q
- Markers: m
- Meiosis: c

$$p(y, q, m, c) = p(y|q)\ p(q|c)\ p(m|c)$$

Integrating (averaging) over c:

$$\begin{aligned} p(y, q, m) &= p(y|q)\ p(q, m) \\ &= p(y|q)\ p(q|m)\ p(m) \end{aligned}$$

Conditional on the QTL, genetic model (phenotype given QTL) independent of linkage model (distribution of QTL given observed markers).

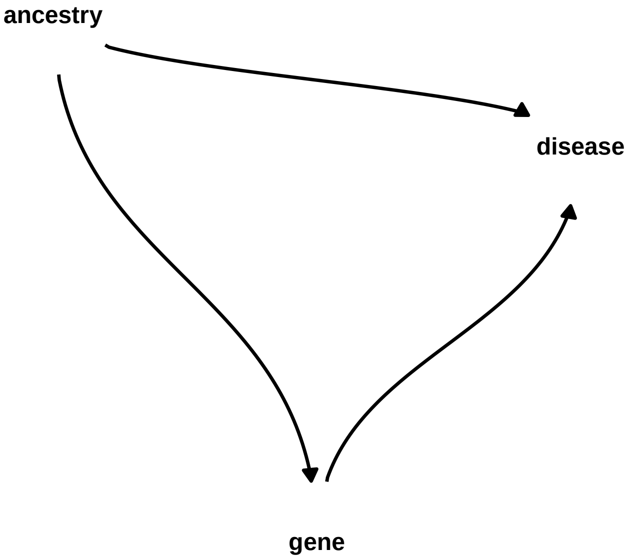
Causal diagrams and marginal independence

A path is causal if it consists entirely of edges with their arrows pointing in the same direction.

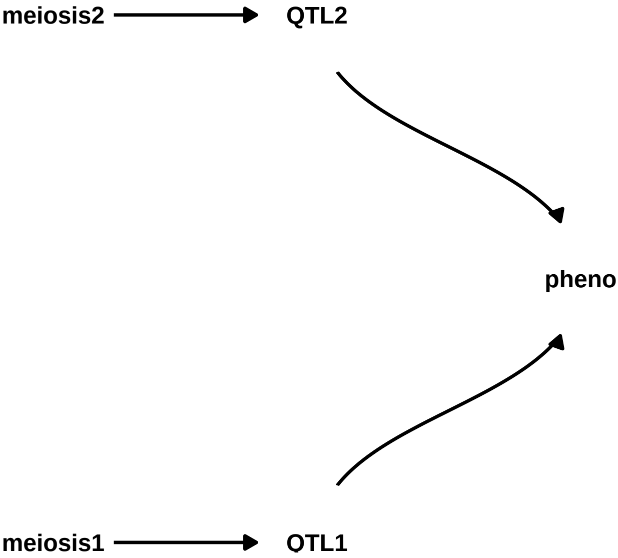
Otherwise, it is non-causal (association).

DAG framework equivalent to potential outcomes framework.

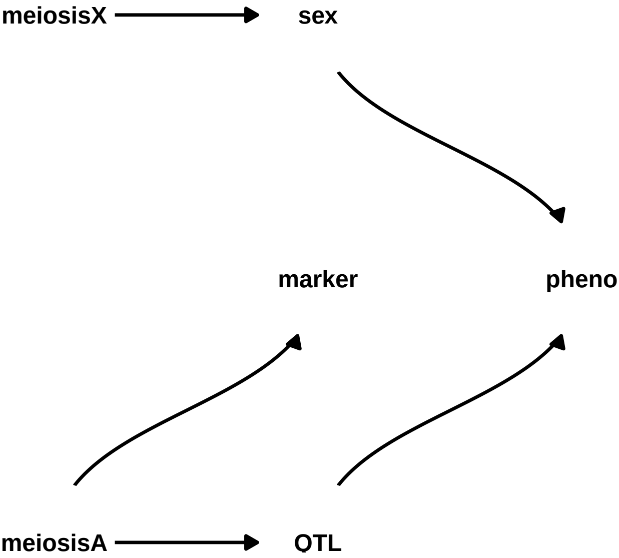
Confounding by ancestry



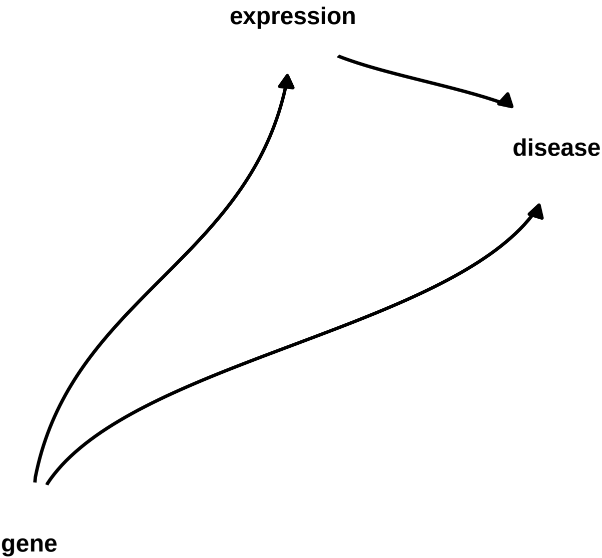
Unlinked QTL



Sex as covariate



Expression mediator



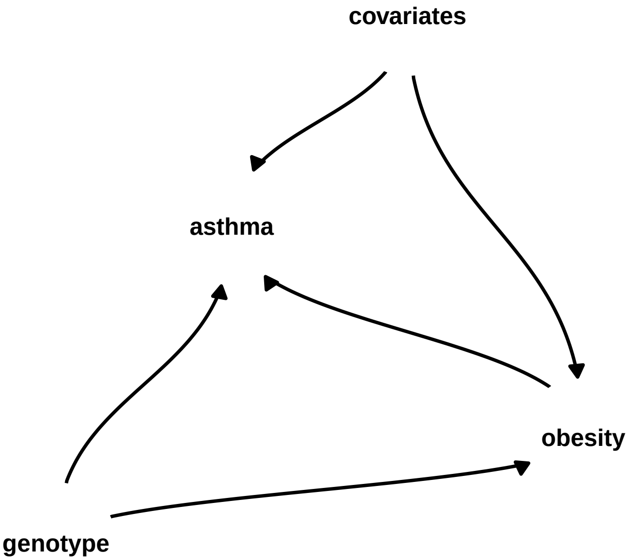
Causal diagrams and conditional independence

A collider is a node where two or more arrows “collide”.

A path is blocked if and only if it contains a non-collider that has been conditioned on, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on.

An unblocked path implies association; a blocked path implies conditional independence.

Asthma obesity


covariates

asthma

obesity

genotype

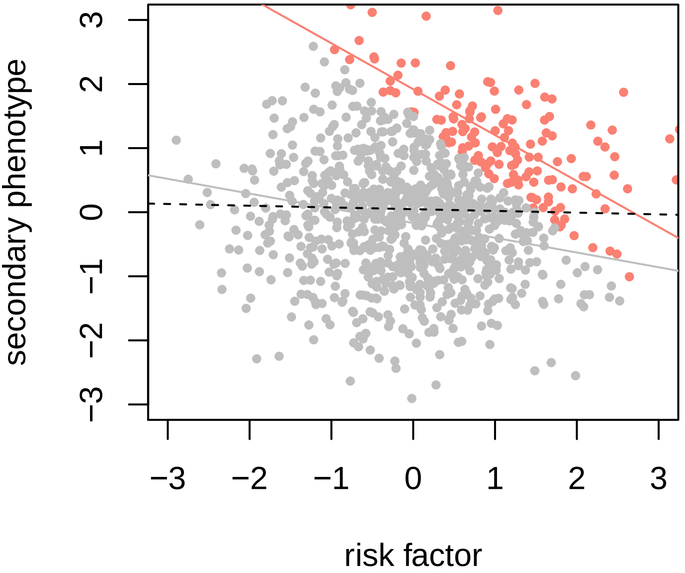
Asthma case-control study (analysis of secondary phenotype)

In an asthma case-control study, asthma status would be a collider.

It is well-known that associations between asthma and covariates or any risk factors are unbiased.

So, if we are interested in the association between a secondary phenotype (obesity), and a risk factor (genotype), that estimate will be biased.

Risk factor and secondary phenotype in case-control study



Instrumental variable estimation

A variable is called an instrument for a potential cause if it is only related to the outcome via the potential cause.

The randomization device in randomized clinical trials is an instrument.

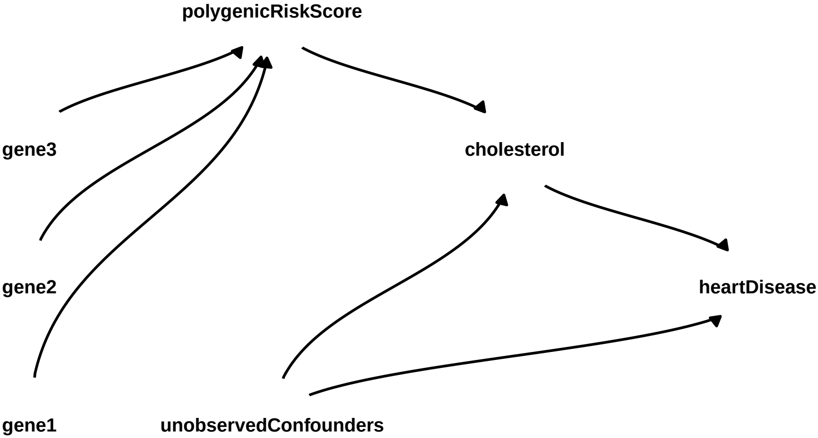
Natural experiments such as the Draft Lottery can be viewed as instruments.

In genetics and epidemiology, the polygenic risk score can be used as an instrument.

For continuous treatment (potential cause), the estimate is

$$\frac{Cov(\text{outcome}, \text{instrument})}{Cov(\text{treatment}, \text{instrument})}$$

Polygenic risk score (Medelian randomization)



Summary

The potential outcomes framework shows how causal inference can be framed as a missing data problem.

We need assumptions and/or devices (instruments, interventions) to infer causal effects.

DAGs can be used to express causal information.

DAGs can be used to evaluate what variables may be associated and what to condition on in analysis.

Framework is non-parametric; additional assumptions can be made to improve estimation.

Further reading and references

Hernan and Robins (2020) “Causal inference: What if”

Pearl (2018) “The book of why: The new science of cause and effect”

Greenland, Pearl, and Robins (1999) “Causal diagrams for epidemiological research”

Textor et. al. (2016) "Robust causal inference using directed acyclic graphs: the R package ‘dagitty’

Knowler et. al. (1988) “Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture”

Sen and Churchill (2001) “A statistical framework for quantitative trait mapping”