Post GWAS analysis and Causal Inference

Yongjin Park, UBC Path + Stat, BC Cancer

07 March 2024

Learning objectives

- Understand the basic idea of summary-based post-GWAS analysis.
- Understand when and how we can bring causal interpretation.

TL; DR

- Mendelian Randomization

```
library(MendelianRandomization)
```

- Establish causal relationship: $M \rightarrow Y$ using genotype X
- Given $X \rightarrow M \rightarrow Y$, the MR effect size is

$$\beta_{M \rightarrow Y} \approx \frac{\beta_{X \rightarrow Y}}{\beta_{X \rightarrow M}}$$

GWAS (previous lecture), *then*, what's next?

- GWAS until 2010s: heavy focuses on **mapping**
  - GWAS map: genetic variants $\longrightarrow$ a phenotype
  - Stringent genome-wide p-value cutoff
  - Study design, meta analysis
- NHGRI-EBI GWAS Catalog:
  `https://www.ebi.ac.uk/gwas/`

GWAS (previous lecture), then, *what's next?*

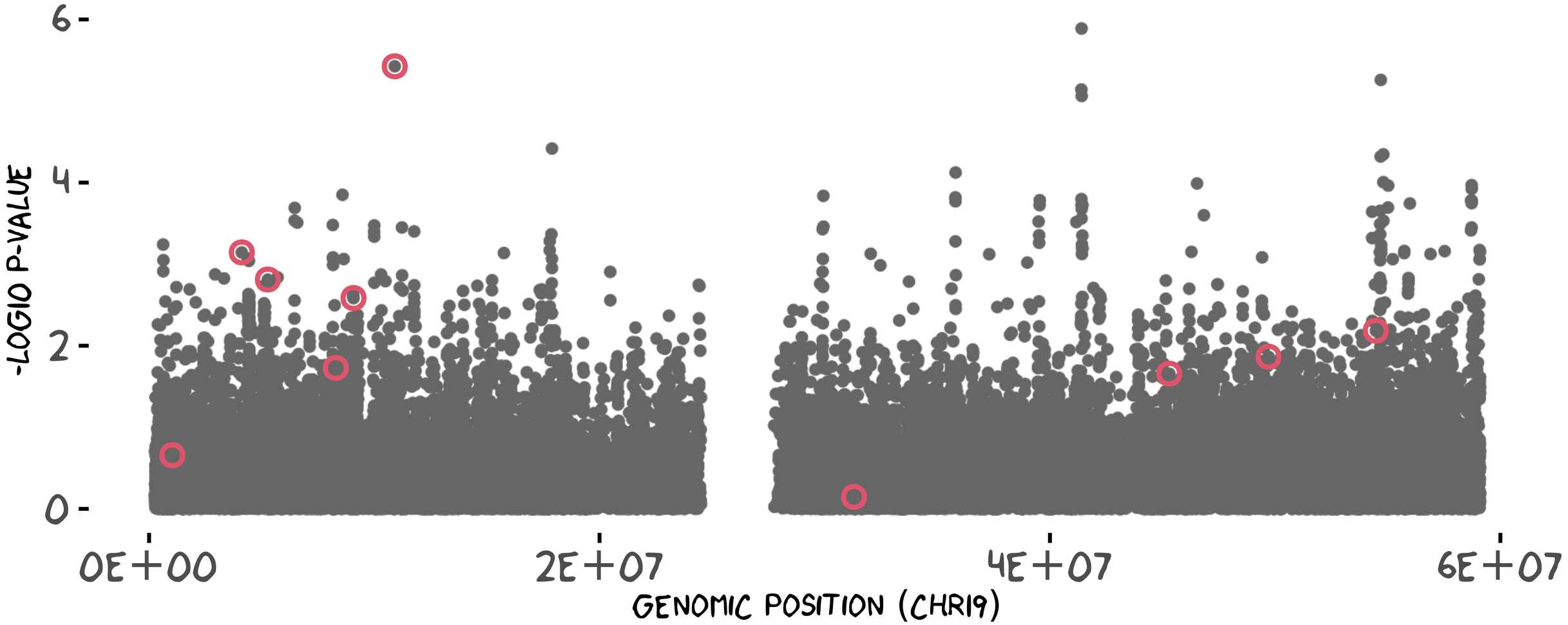
- GWAS since 2010s: more emphases on what to do with GWAS
  - Can we turn GWAS results to a prediction model?
  - Can we elucidate **the mechanisms** implicated by GWAS?
  - Machine learning, data integration, **causal inference**

Today's lecture

1. Summary Statistics-based post-GWAS
2. Intro to causal inference
3. Mendelian Randomization

We may not have a full data (especially phenotypes)

Instead, we could have just summary statistics (e.g., Manhattan plot).



We could have full access to univariate stats

```
head(sumstat.dt[, .(chromosome, marker.ID, physical.pos, beta,
    se, pv)])
```

```
##    chromosome marker.ID physical.pos         beta         se         pv
##         <int>    <char>        <int>        <num>      <num>      <num>
## 1:         19 rs8100066       260912 -0.013392212 0.03209562 0.6765244
## 2:         19 rs8102615       260970 -0.038096443 0.02945621 0.1960189
## 3:         19 rs8105536       261033 -0.014187124 0.03210357 0.6585875
## 4:         19 rs2312724       266034 -0.004839159 0.01335465 0.7171156
## 5:         19 rs4897933       266888  0.003032540 0.01061138 0.7750691
## 6:         19 rs1020382       267039 -0.007183893 0.03202064 0.8225019
```

Figure out how these univariate stats were calculated

A variant-by-variant model:

$$Y_i \sim X_{ij}\beta_j + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Figure out how these univariate stats were calculated

A variant-by-variant model:

$$Y_i \sim X_{ij}\beta_j + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Summary statistics for each variant j:

$$\hat{\beta}_j = \frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}^2}, \quad \hat{\mathbb{V}}[\beta_j] = \frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1} X_{ij}^2}, \quad Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathbb{V}}[\beta_j]}}$$

We know that a phenotype is polygenic

Thus, we assume there are multivariate effects

$$Y_i = \sum_{j=1}^{p} X_{ij}\theta_j + \epsilon, \quad \forall i \in [n], \text{ or}$$

$$\mathbf{y} = X\theta + \epsilon,$$

where $\theta_j \neq \beta_j$

Goal: How can we recover the multivariate one?

Data:

$$\hat{Z}_j = \sum_{i=1}^{n} X_{ij} Y_i / \sigma_\epsilon \sqrt{n}$$

for all $j \in [p]$.

Goal: Recover $p \times 1$ multivariate effect size vector θ.

Inference $\approx$ reversing the generative process

Data:

$$\hat{Z}_j = \sum_{i=1}^{n} X_{ij} Y_i / \sigma_\epsilon \sqrt{n}$$

for all $j \in [p]$.

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Inference $\approx$ reversing the generative process

With a matrix-vector notation (to save space),

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \underset{n \times 1 \text{ phenotype}}{\mathbf{y}}$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Inference $\approx$ reversing the generative process

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \underset{n \times 1 \text{ phenotype}}{\mathbf{y}} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Inference $\approx$ reversing the generative process

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \underset{n \times 1 \text{ phenotype}}{\mathbf{y}} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \frac{\sqrt{n}}{\sigma} \underbrace{\left(\frac{1}{n} X^\top X\right)}_{\text{LD}} \theta + \frac{1}{\sigma\sqrt{n}} X^\top \epsilon$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Inference $\approx$ reversing the generative process

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \underset{n \times 1 \text{ phenotype}}{\mathbf{y}} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \mathbf{R}\frac{\sqrt{n}}{\sigma}\theta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

where $\mathbf{R} = n^{-1} X^\top X$ is an empirical LD matrix.

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Running SuSiE in a summary statistics mode

The underlying model:

$$\mathbf{z} \sim \mathcal{N}\left(\frac{\sqrt{n}}{\sigma} R\theta, R\right)$$

```
library(susieR)
z <- sumstat.dt$beta/sumstat.dt$se
R <- cov(X)
susie.out <- susie_rss(z, R, n = nrow(X))
```

Running SuSiE in a summary statistics mode

The underlying model:

$$\underset{\text{univariate z}}{\mathbf{z}} \sim \mathcal{N}\left(\underbrace{\frac{\sqrt{n}}{\underbrace{\sigma}_{\text{LD}}}R}_{\text{LD}} \underbrace{\theta}_{\text{multivariate effect}}, \underset{\text{LD}}{R}\right)$$

```
library(susieR)
z <- sumstat.dt$beta/sumstat.dt$se
R <- cov(X)
susie.out <- susie_rss(z, R, n = nrow(X))
```

LD-score: another type of summary-based analysis

What is a generative model for a χ^2_j, which is $(= Z^2_j)$, statistics vector?

We have seen this relationship in the fine-mapping model:

$$\underset{\text{univariate, summary stat}}{Z_j} = \frac{\sqrt{n}}{\sigma} \sum_k \underset{\text{LD between j and k}}{R_{jk}} \underset{\text{multivariate, true effect}}{\theta_k} + \epsilon_j$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

Bulik-Sullivan *et al*., *Nature Genetics* (2014); Finucane *et al*., *Nature Genetics* (2015)

LD-score: another type of summary-based analysis

What is a generative model for a χ^2_j, which is $(= Z^2_j)$, statistics vector?

$$\mathbb{E}\left[\chi^2_j\right] = \mathbb{E}\left[Z^2_j\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

LD-score: another type of summary-based analysis

What is a generative model for a χ^2_j, which is $(= Z^2_j)$, statistics vector?

$$\mathbb{E}\left[\chi^2_j\right] = \mathbb{E}\left[Z^2_j\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

If the effects are independent of each other, i.e., $\mathbb{E}\left[\theta_k\theta_j\right] = 0$ for all $k \neq j$,

$$\mathbb{E}\left[\chi^2_j\right] = n\underbrace{\sum_k R^2_{jk}}_{\text{LD-score}}\mathbb{E}\left[\theta^2_k\right] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

where p is the total number of SNPs,

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \stackrel{\text{def}}{=} \sum_k R_{jk}^2,$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \stackrel{\text{def}}{=} \sum_k R_{jk}^2,$$

We get

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}[\theta_k^2] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \overset{\text{def}}{=} \sum_k R_{jk}^2,$$

We get

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}[\theta_k^2] + 1 = \underset{\text{sample size}}{n} \underset{\text{LD score}}{l_j} \underset{\text{per SNP heritability}}{\frac{\tau}{p}} + 1$$

where p is the total number of SNPs.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

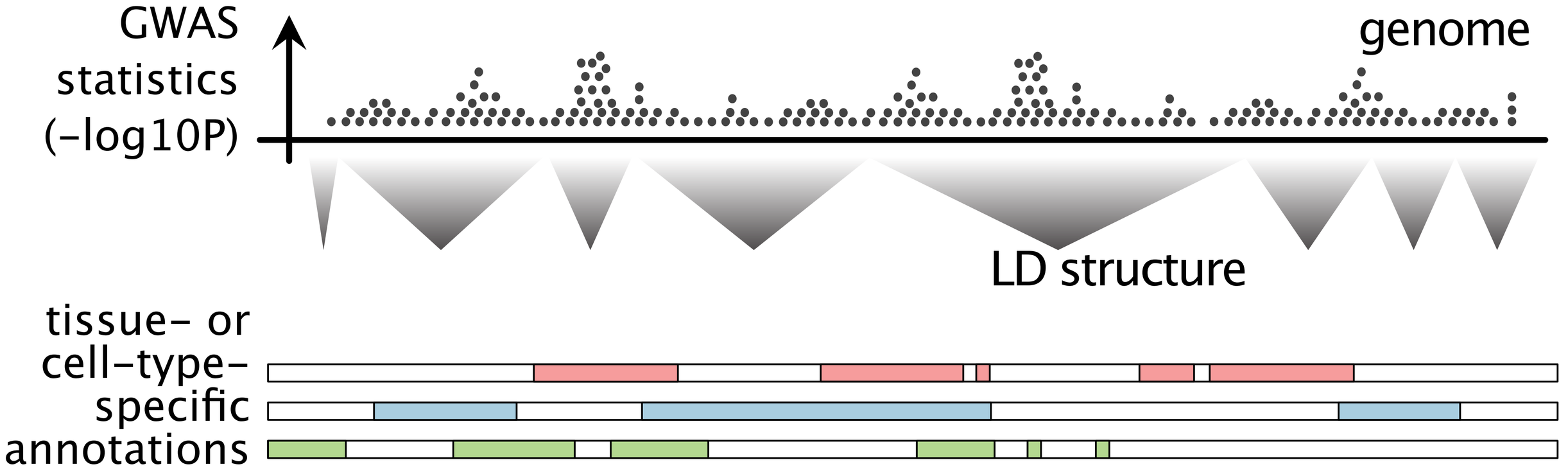
We can treat the relationships as a regression model and find the heritability parameters by regressing the observed χ^2 statistics on the reference LD scores l_j:

$$\begin{pmatrix} \chi_1^2 \\ \vdots \\ \chi_j^2 \\ \vdots \end{pmatrix} \sim \frac{n}{p} \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underbrace{\tau}_{\text{per SNP heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

If the intercept of $\{\chi_j^2\}$ deviate from 1, we can interpret that the GWAS statistics are inflated by some unadjusted population structures or other confounding factors.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Stratified LD-score regression partitions total heritability into multiple genomic annotations



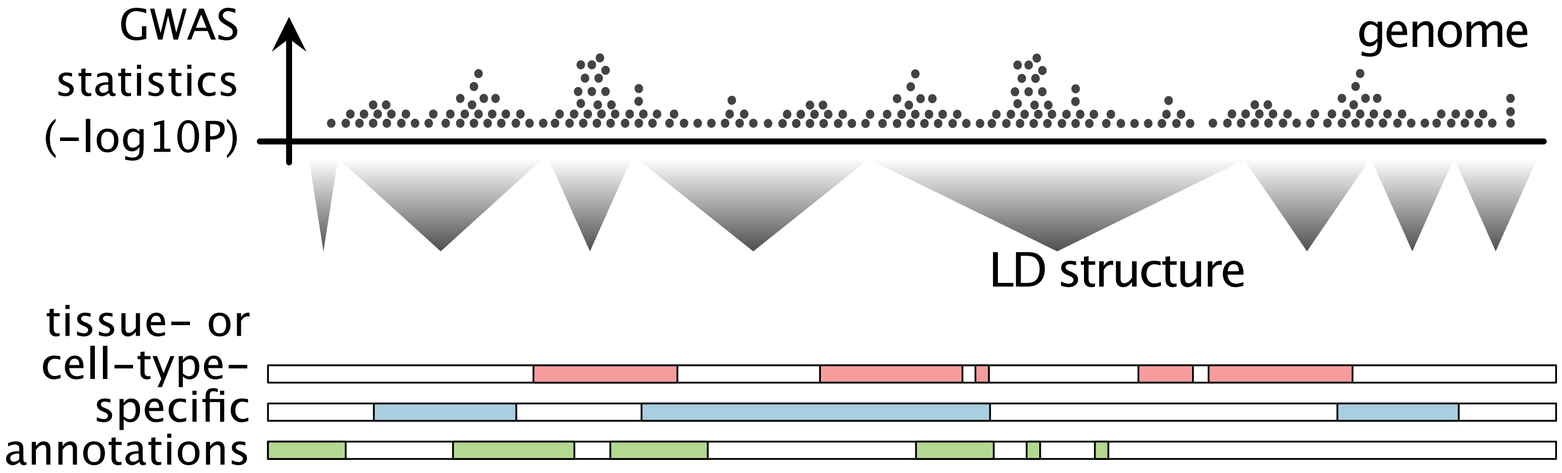
Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$



Finucane *et al.*, *Nature Genetics* (2015)

Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$

Instead of assuming a single parameter for the overall per-SNP heritability τ, we can "partition" this total heritability into annotation-type-specific ones, $\{\tau_t\}$.

Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\ \underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

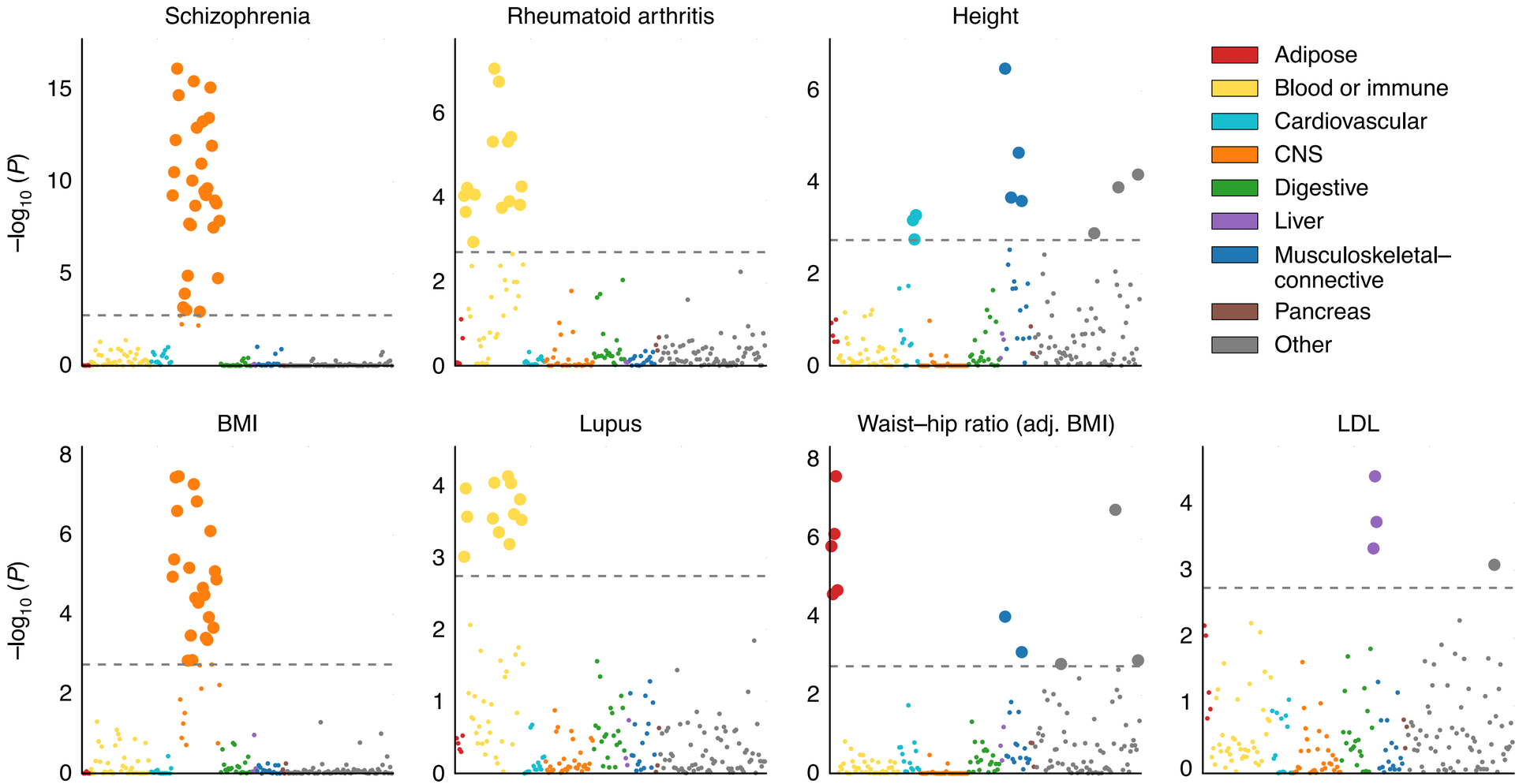
$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$

More explicitly,

$$\begin{pmatrix} \chi_1^2 \\ \vdots \\ \chi_j^2 \\ \vdots \end{pmatrix} \sim \frac{n}{p} \underbrace{\begin{pmatrix} l_{11} & l_{12} & l_{1t} & \cdots \\ & \vdots & & \\ l_{j1} & l_{j2} & l_{jt} & \cdots \\ & \vdots & & \end{pmatrix}}_{\text{stratified LD scores}} \underbrace{\begin{pmatrix} \tau_1 \\ \vdots \\ \tau_t \\ \vdots \end{pmatrix}}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

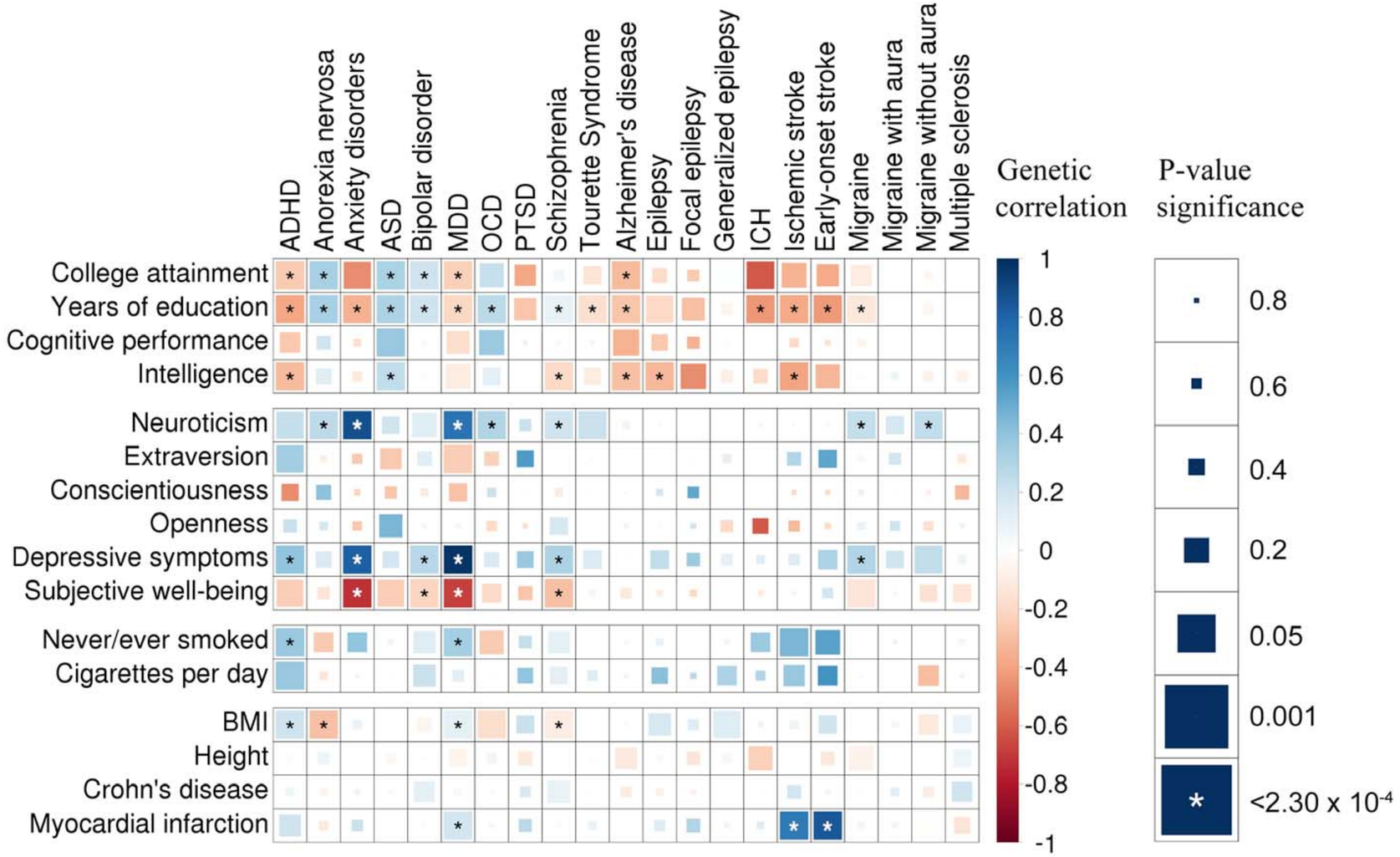
Finucane *et al.*, *Nature Genetics* (2015)

Stratified LDSC can identify tissue-specific enrichment of GWAS signals



Finucane … Prices, *Nature Genetics* (2018)

When multiple GWAS were done, post-GWAS analysis begins



Bivariate LD-score regression

Instead of one χ^2 vector, we need to deal with the element-wise product of two vectors of z-scores (between a trait 1 and 2):

$$\begin{pmatrix} z_1^{(1)} z_1^{(2)} l_1 \\ \vdots \\ z_j^{(1)} z_j^{(2)} l_j \\ \vdots \end{pmatrix} \sim \frac{\sqrt{N_1 N_2}}{p} \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underbrace{\rho}_{\text{genetic correlation}} + \underbrace{\frac{\rho_0 N_s}{\sqrt{N_1 N_2}}}_{\text{sample sharing}}$$

where N_1 and N_2 count sample size of the GWAS 1 and 2; N_s is the number of control individuals shared between the two traits.

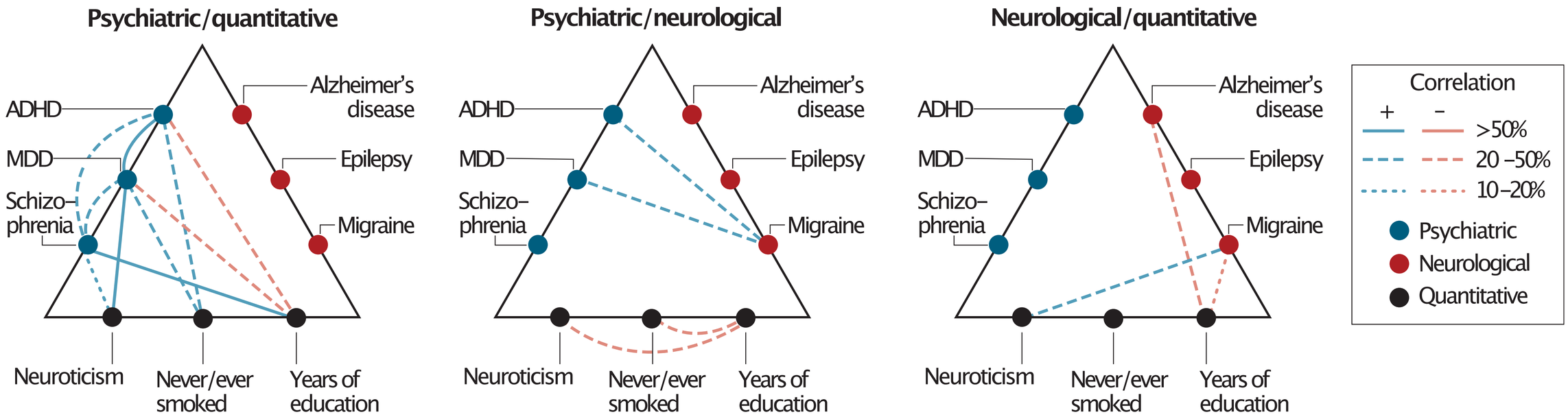
Bulik-Sullivan *et al.*, *Nature Genetics* (2015)

Post-GWAS analysis example: genetic correlations across many traits

Psychiatric disorders				Neurological disorders			
Disorder	**Source**	**Cases**	**Controls**	**Disorder**	**Source**	**Cases**	**Controls**
Attention deficit hyperactivity disorder	PGC-ADD2	12,645	84,435	Alzheimer's disease	IGAP	17,008	37,154
Anorexia nervosa	PGC-ED	3495	10,982	Epilepsy	ILAE	7779	20,439
Anxiety disorders	ANGST	5761	11,765	Focal epilepsy	"	4601*	17,985*
Autism spectrum disorder	PGC-AUT	6197	7377	Generalized epilepsy	"	2525*	16,244*
Bipolar disorder	PGC-BIP2	20,352	31,358	Intracerebral hemorrhage	ISGC	1545	1481
Major depressive disorder	PGC-MDD2	66,358	153,234	Ischemic stroke	METASTROKE	10,307	19,326
Obsessive-compulsive disorder	PGC-OCDTS	2936	7279	Cardioembolic stroke	"	1859*	17,708*
Posttraumatic stress disorder	PGC-PTSD	2424	7113	Early onset stroke	"	3274*	11,012*
Schizophrenia	PGC-SCZ2	33,640	43,456	Large-vessel disease	"	1817*	17,708*
Tourette syndrome	PGC-OCDTS	4220	8994	Small-vessel disease	"	1349*	17,708*
				Migraine	IHGC	59,673	316,078
				Migraine with aura	"	6332*	142,817*
				Migraine without aura	"	8348*	136,758*
				Multiple sclerosis	IMSGC	5545	12,153
				Parkinson's disease	IPDGC	5333	12,019
Total psychiatric		158,028	365,993	Total neurologic		107,190	418,650

The Brainstorm Consortium, *Science* (2018)

Bivariate LDSC reveals disease comorbidity at the common genetic variants' level



The Brainstorm Consortium, *Science* (2018)

What we need to know in this section

1. GWAS summary statistics contain good amount of information
2. In the summary data, we need to deal with LD structure R

Today's lecture

1. Summary Statistics-based post-GWAS
2. Intro to causal inference
3. Mendelian Randomization

Why causal inference?

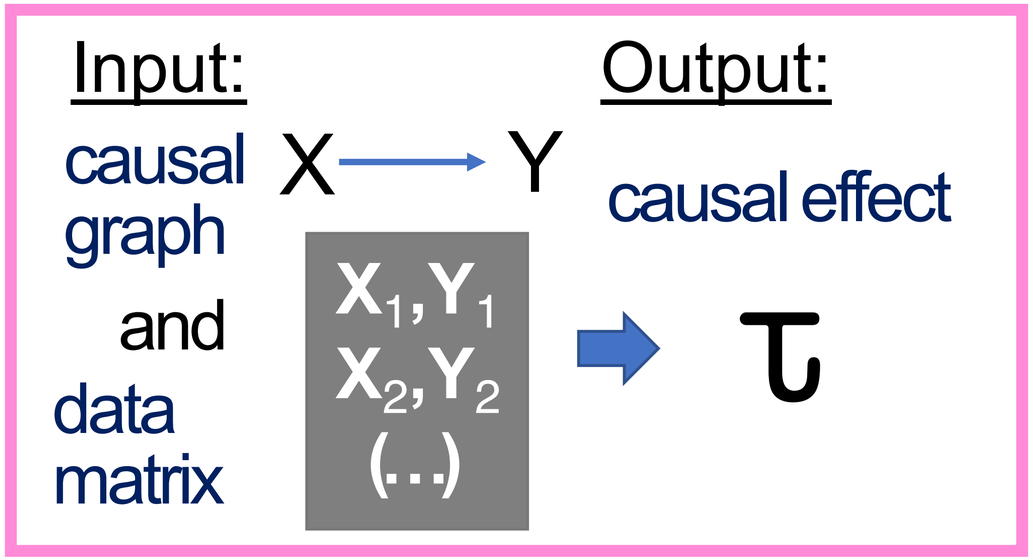
Mendelian + Randomization

Why causal inference?

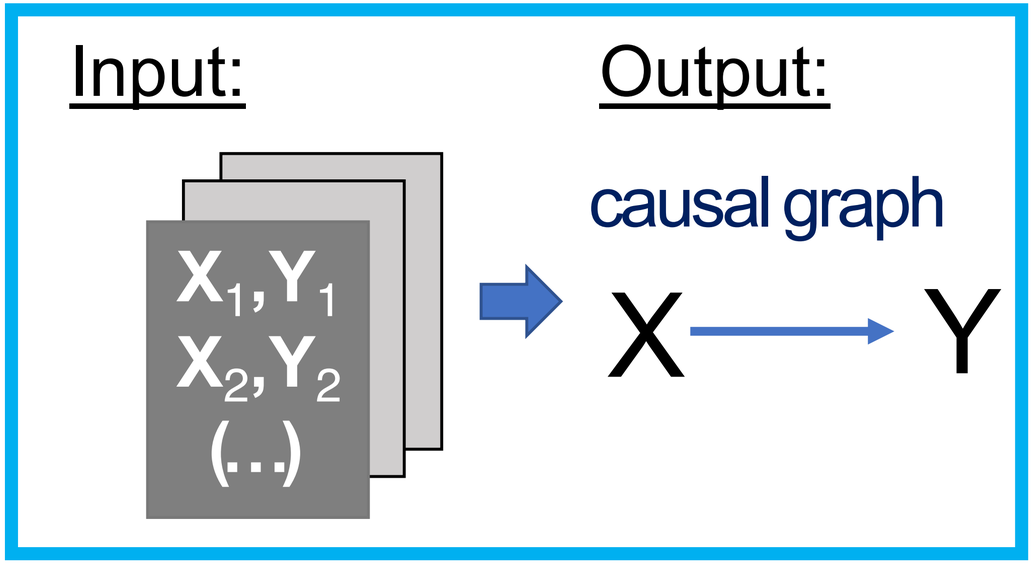
$$\underbrace{\text{Mendelian}}_{\text{genetics}} + \underbrace{\text{Randomization}}_{\text{causal inference}}$$

Background: What is causal inference?

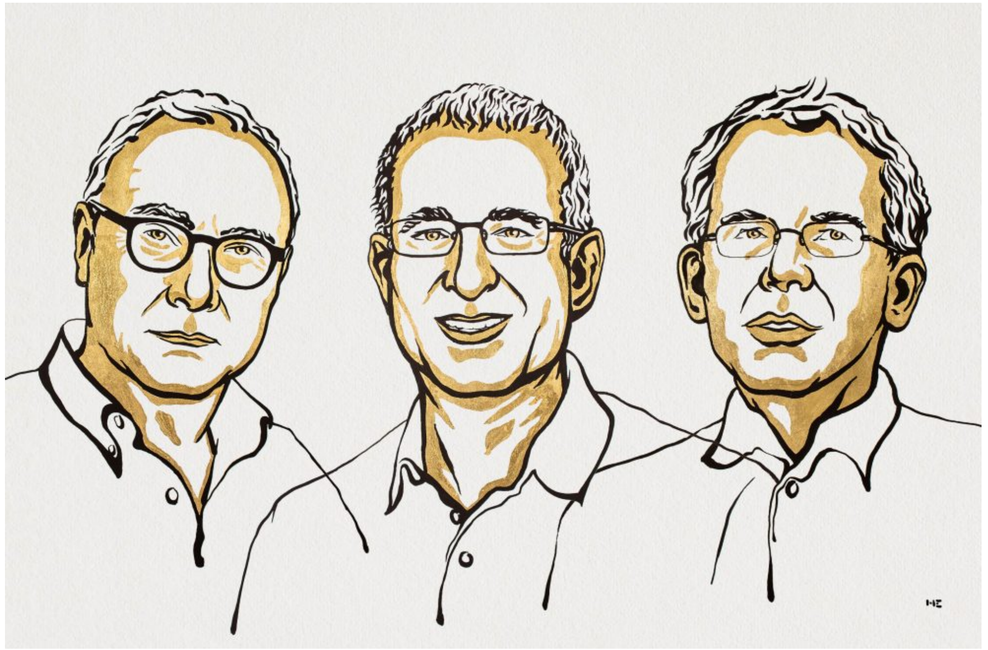
Causal Effect Inference



Causal structure discovery



The Nobel Prize 2021 in Economical Sciences



David Card, Joshua Angrist (instrumental variable), Guido Imbens (matching, propensity)

"For the methodological contributions to the analysis of causal relationships"

The goal of causal effect inference

1. Instead of **”seeing”** what happened (data analysis), we want to measure **the effect of ”doing”** an experiment (intervention).

The goal of causal effect inference

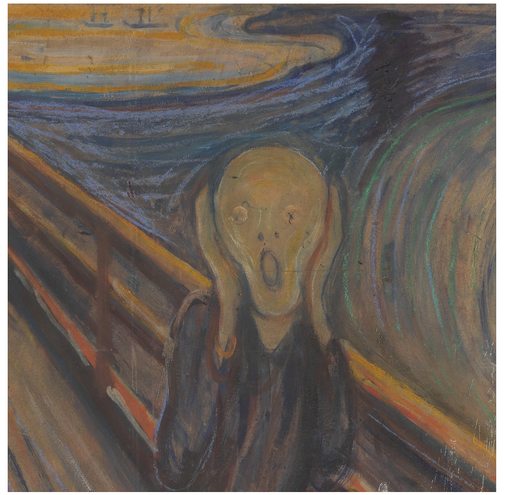
1. Instead of **”seeing”** what happened (data analysis), we want to measure **the effect of ”doing”** an experiment (intervention).
2. Because correlation is not causation.

The goal of causal effect inference

1. Instead of **”seeing”** what happened (data analysis), we want to measure **the effect of ”doing”** an experiment (intervention).
2. Because correlation is not causation.
3. If it is not causation, then what is it?

Common misconception - correlation implies causation

Henry Niles also wrote, *“To contrast ‘causation’ and ‘correlation’ is unwarranted* ***because causation is simply perfect correlation****”*



Pearl, *The Book of Why*, p.78

- How do we measure causal effects? Will any experiment do?

A working example: Why correlation is not causation

- U_i: The age at which a person i enrolled to this study.
- X_i: An indicator variable for a physician decided to administer this new preventative drug $X_i = 1$ or not $X_i = 0$.
- Y_i: Biomarker protein concentration (e.g., Amyloid-beta protein disrupts normal brain functions).

A working example: How the data had been generated.

- A person i's age U_i.

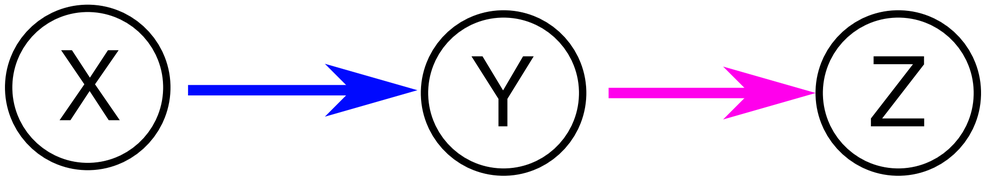
1. Prescribe the preventative drug based on the subject's age U_i (as they might need one):

$$P(X_i = 1|U_i) = \sigma\left(U_i\delta\right)$$

2. A biomarker protein concentration:

$$\mathbb{E}[Y_i|X, U] = X_i\tau + U_i\beta$$

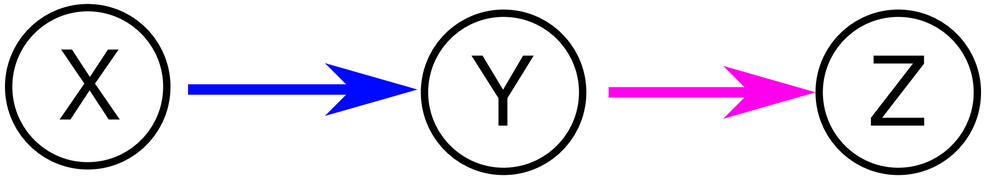
Directed Acyclic Graph as a language of causal inference



$$p(X, Y, Z) = p(Y|X) \quad p(Z|Y) \quad p(X)$$

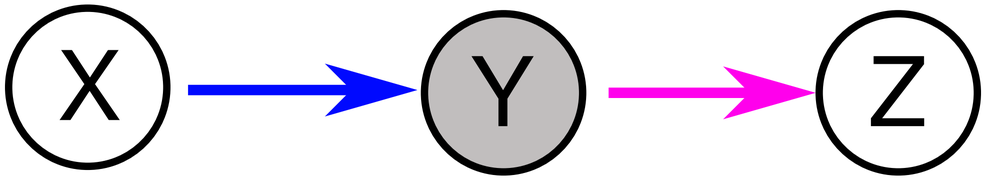
first edge second edge first node

DAG: Causal path/trail and reachability



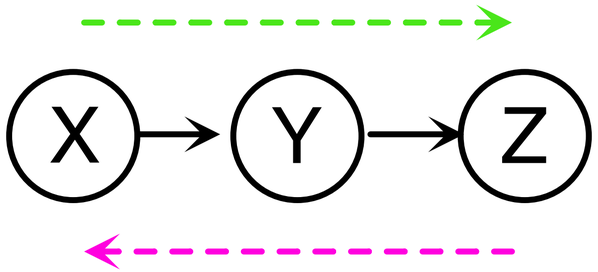
- There is a path between X and Y
- Also between Y and Z
- Also from X to Z

DAG: Conditioning (and/or adjustment)



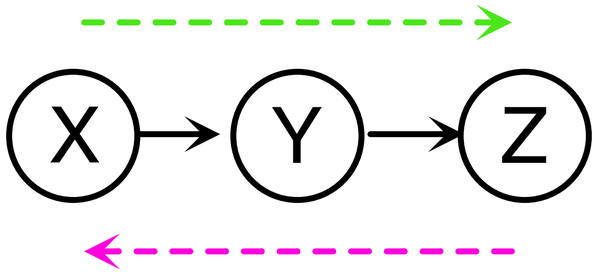
- A shaded node = conditioning like $P(Z|Y = y^\star)$ and $P(Y = y^\star|X)$
- There is a path between X and Y
- Also between Y and Z
- **But** no path from X to Z

d-separation: testing conditional independence (flow vs. no flow)

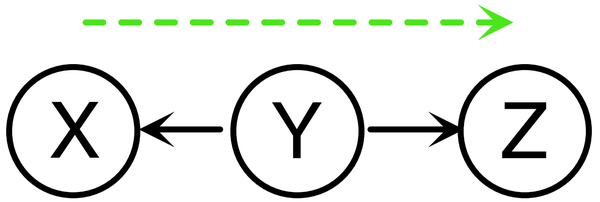


Flow: The outcome of Z depends on the effect of X

d-separation: testing conditional independence (flow vs. no flow)

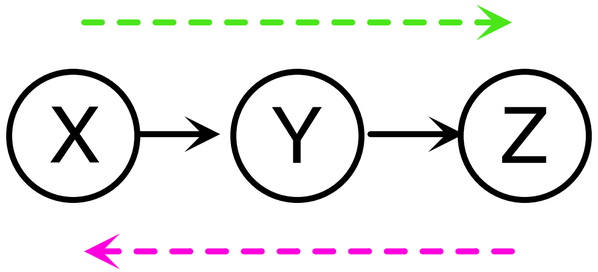


Flow: The outcome of Z depends on the effect of X

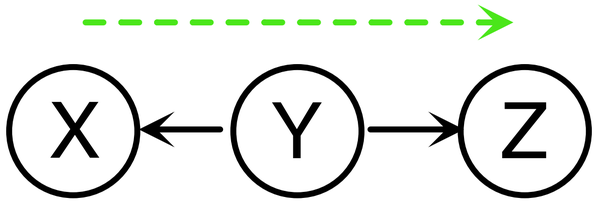


Flow: The outcome of Z depends on the effect of Y; so does the outcome of X on that of Y

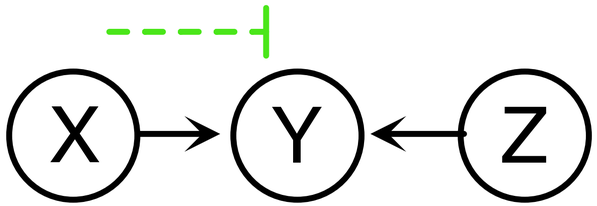
d-separation: testing conditional independence (flow vs. no flow)



Flow: The outcome of Z depends on the effect of X

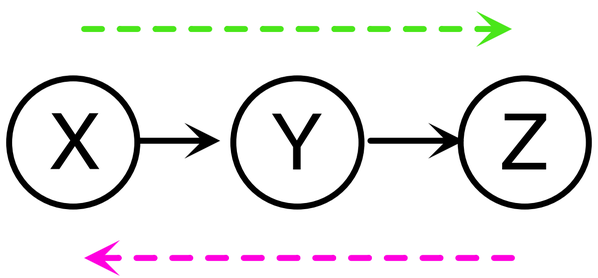


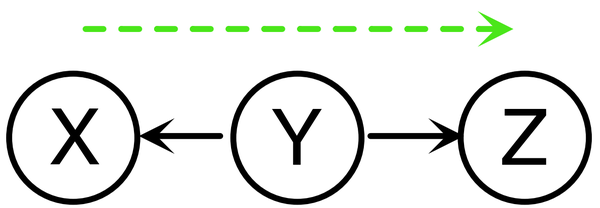
Flow: The outcome of Z depends on the effect of Y; so does the outcome of X on that of Y

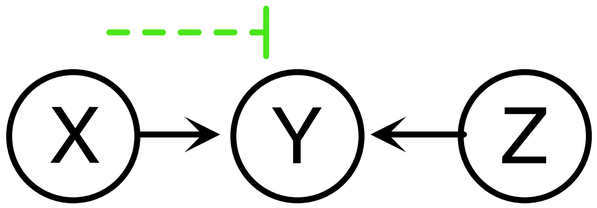


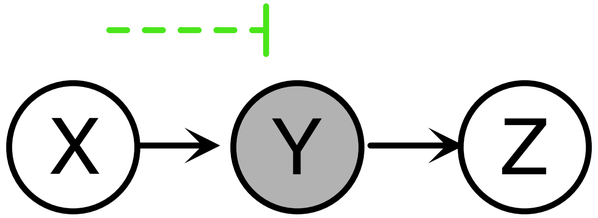
No Flow: The outcome of Z only affects Y; do does X

d-separation: testing conditional independence (flow vs. no flow)



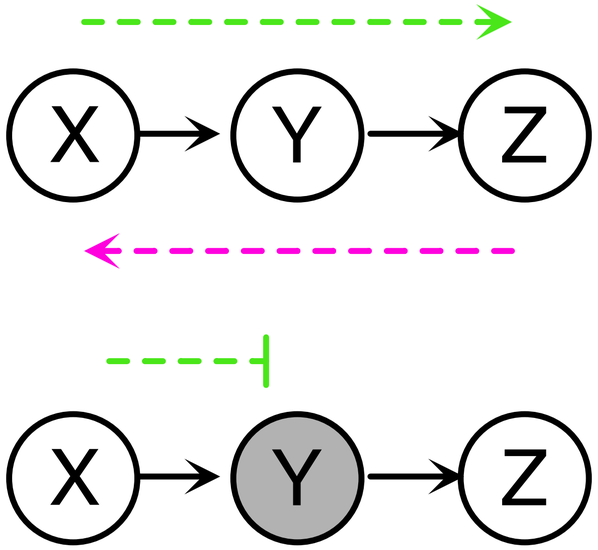




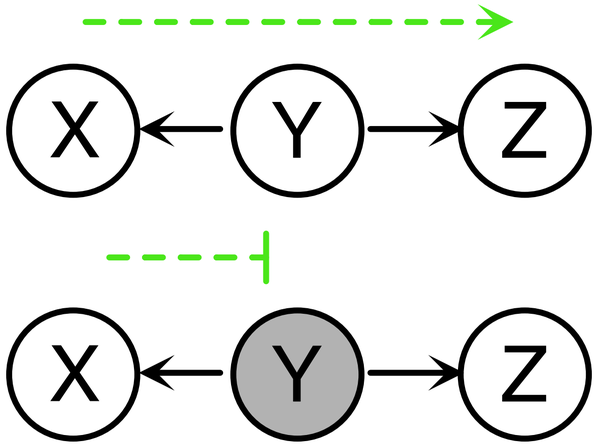


No Flow: The outcome of Z solely depends on $Y = y^*$, not X

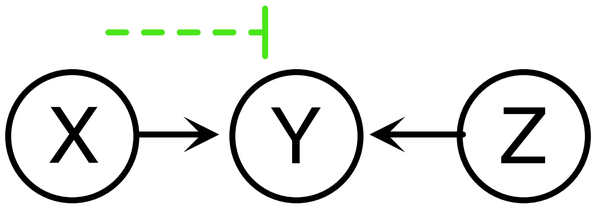
d-separation: testing conditional independence (flow vs. no flow)



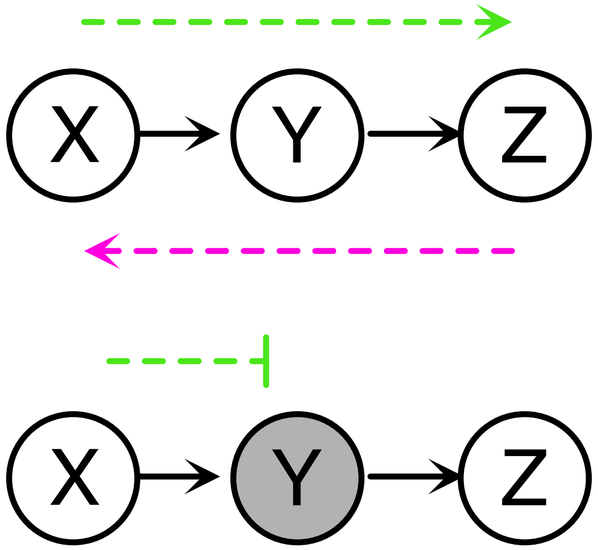
No Flow: The outcome of Z solely depends on $Y = y^*$, not X



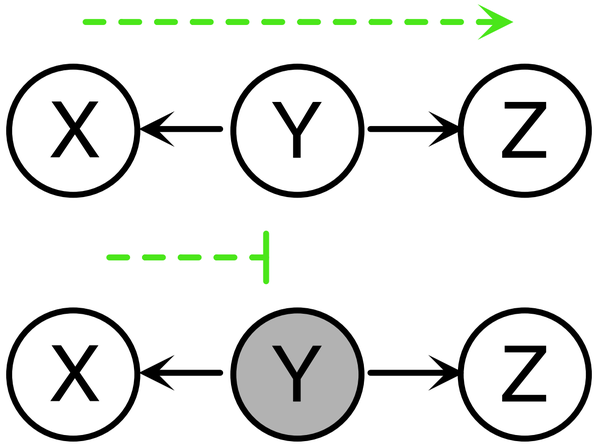
No Flow: The outcome of Z solely depends on $Y = y^*$, not X



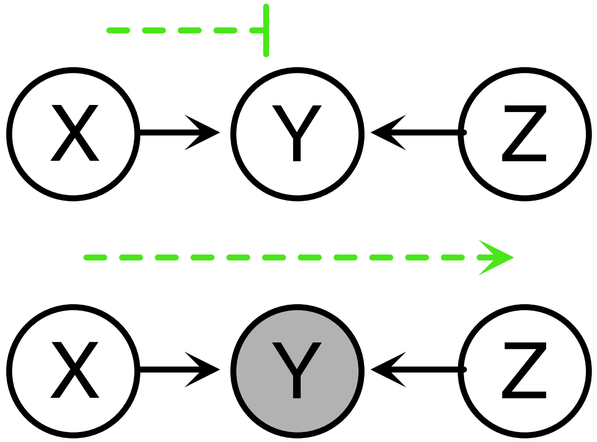
d-separation: testing conditional independence (flow vs. no flow)



No Flow: The outcome of Z solely depends on $Y = y^*$, not X

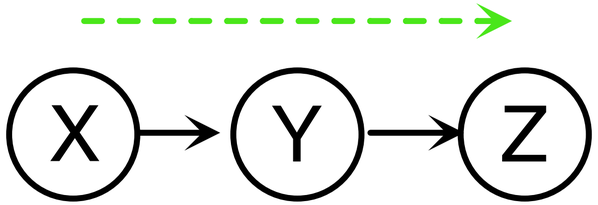


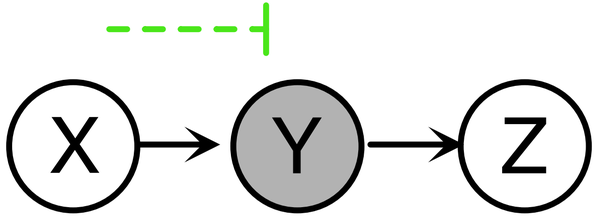
No Flow: The outcome of Z solely depends on $Y = y^*$, not X



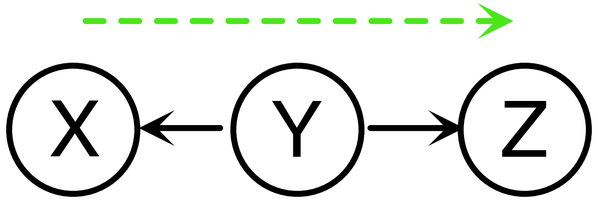
Flow: The observed outcome $Y = y^*$ could stem from X or/and Z

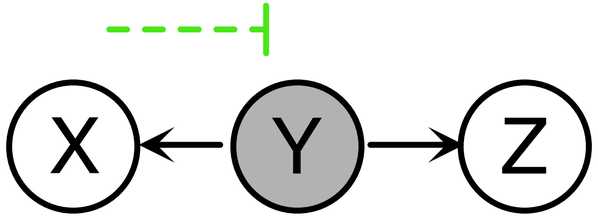
d-separation: testing conditional independence (flow vs. no flow)



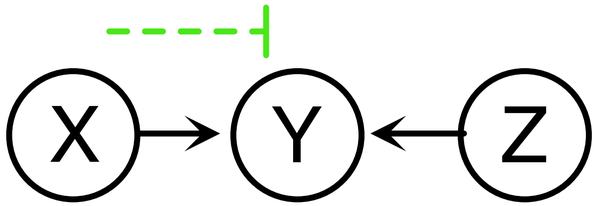


No Flow: The outcome of Z solely depends on $Y = y^*$, not X

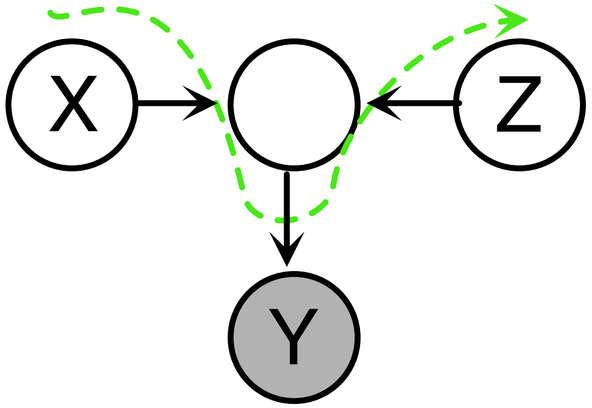




No Flow: The outcome of Z solely depends on $Y = y^*$, not X



Flow: deeper v-structure



Going back to the same working example

```r
## Simulation parameters
n = 200
.delta = 2
.tau = -1
.beta = 2
.eps = 0.3
sgm <- function(x) 1/(1 + exp(-x))

## A generative scheme
uu <- rnorm(n)  # standardized age
xx <- rbinom(n = n, size = 1, prob = sgm(uu * .delta))
yy <- xx * .tau + uu * .beta + rnorm(n) * .eps
```

- The variable U confounds X and Y.
- How can we represent it graphically?

If there were no confounding:

```r
yy.unc <- xx * .tau + rnorm(n) * .eps
```

Going back to the same working example

```r
## Simulation parameters
n = 200
.delta = 2
.tau = -1
.beta = 2
.eps = 0.3
sgm <- function(x) 1/(1 + exp(-x))

## A generative scheme
uu <- rnorm(n)  # standardized age
xx <- rbinom(n = n, size = 1, prob = sgm(uu * .delta))
yy <- xx * .tau + uu * .beta + rnorm(n) * .eps
```

- The variable U confounds X and Y.
- How can we represent it graphically?

If there were no confounding:

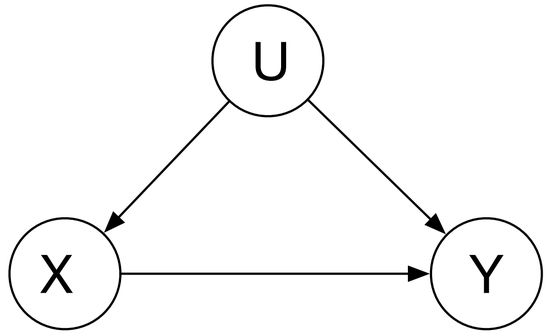
```r
yy.unc <- xx * .tau + rnorm(n) * .eps
```

Going back to the same working example

```r
## Simulation parameters
n = 200
.delta = 2
.tau = -1
.beta = 2
.eps = 0.3
sgm <- function(x) 1/(1 + exp(-x))

## A generative scheme
uu <- rnorm(n)  # standardized age
xx <- rbinom(n = n, size = 1, prob = sgm(uu * .delta))
yy <- xx * .tau + uu * .beta + rnorm(n) * .eps
```

- The variable U confounds X and Y.
- How can we represent it graphically?



1. How can we identify the causal effect, $X \to Y$?

If there were no confounding:

```r
yy.unc <- xx * .tau + rnorm(n) * .eps
```

Going back to the same working example

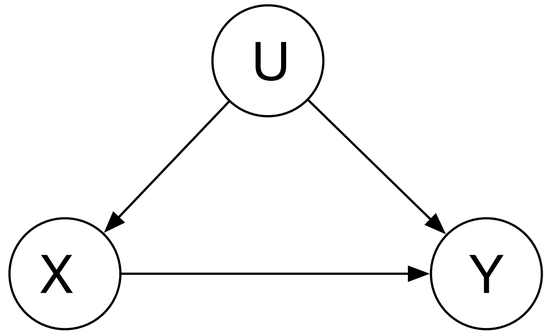
```r
## Simulation parameters
n = 200
.delta = 2
.tau = -1
.beta = 2
.eps = 0.3
sgm <- function(x) 1/(1 + exp(-x))

## A generative scheme
uu <- rnorm(n)  # standardized age
xx <- rbinom(n = n, size = 1, prob = sgm(uu * .delta))
yy <- xx * .tau + uu * .beta + rnorm(n) * .eps
```

If there were no confounding:

```r
yy.unc <- xx * .tau + rnorm(n) * .eps
```

- The variable U confounds X and Y.
- How can we represent it graphically?



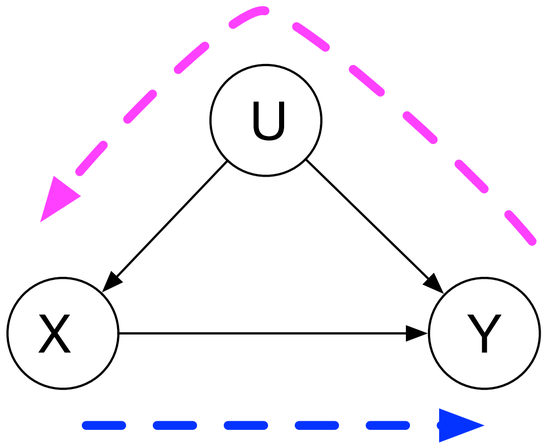
1. How can we identify the causal effect, $X \rightarrow Y$?
2. Why can't we just report the correlation/association between X and Y?

Going back to the same working example

```r
## Simulation parameters
n = 200
.delta = 2
.tau = -1
.beta = 2
.eps = 0.3
sgm <- function(x) 1/(1 + exp(-x))

## A generative scheme
uu <- rnorm(n)  # standardized age
xx <- rbinom(n = n, size = 1, prob = sgm(uu * .delta))
yy <- xx * .tau + uu * .beta + rnorm(n) * .eps
```

- The variable U confounds X and Y.
- How can we represent it graphically?



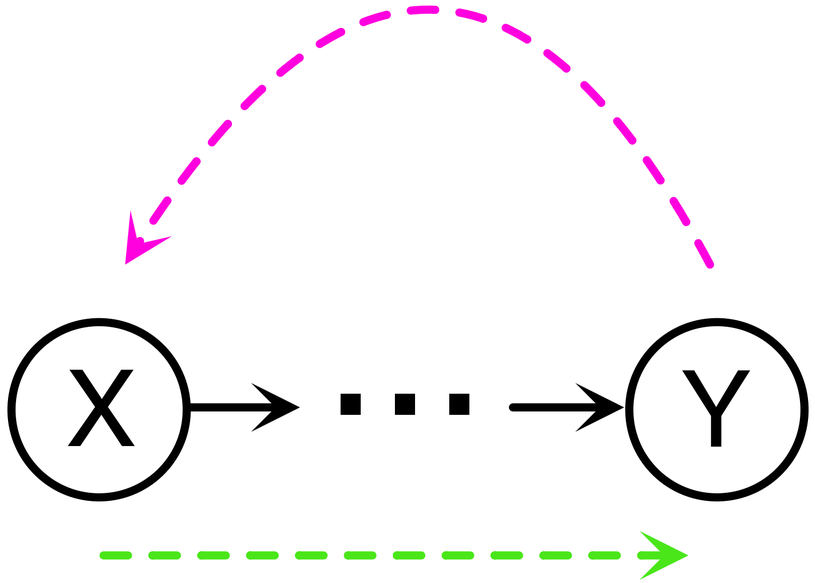
If there were no confounding:

```r
yy.unc <- xx * .tau + rnorm(n) * .eps
```

1. How can we identify the causal effect, $X \to Y$?
2. Why can't we just report the correlation/association between X and Y?

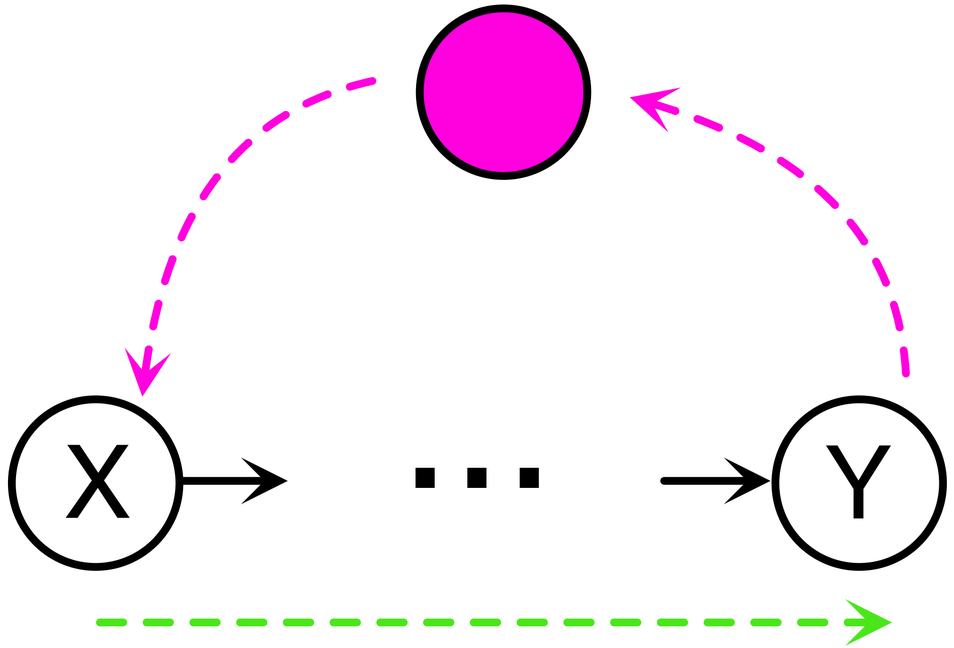
It happened because

Question: Which nodes should be “conditioned” and/or “adjusted” to block a reverse path from Y to X?

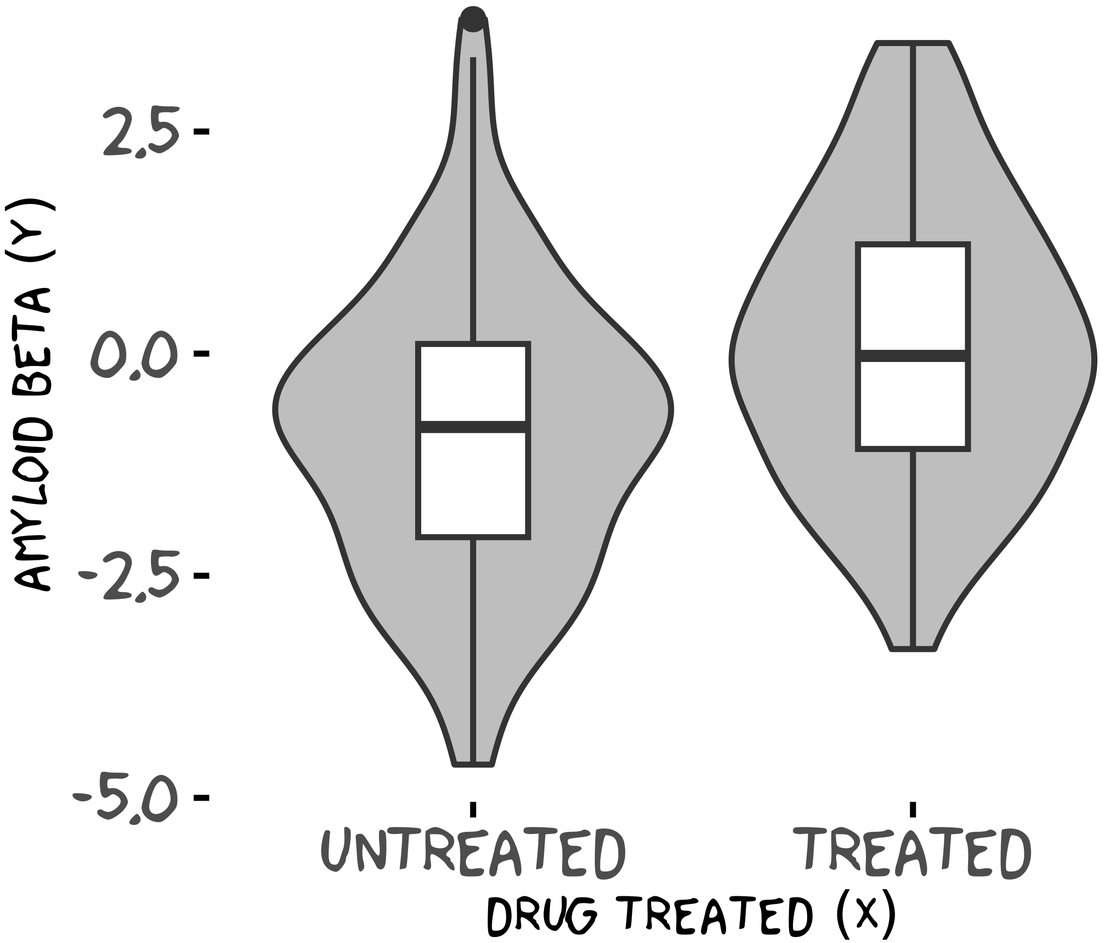


It happened because

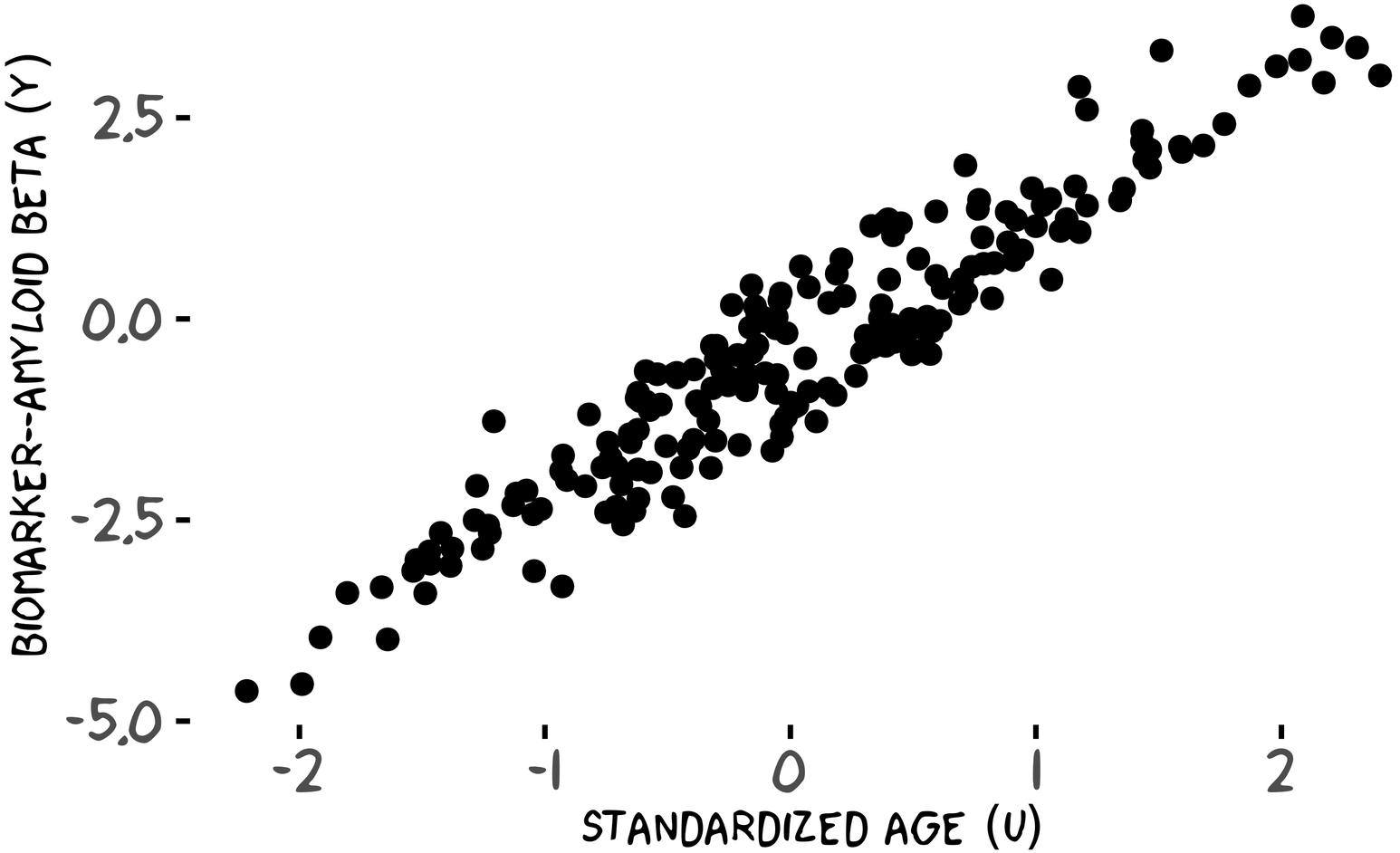
Question: Which nodes should be “conditioned” and/or “adjusted” to block a reverse path from Y to X?



Differential biomarker analysis suggests that our drug can exacerbate AD

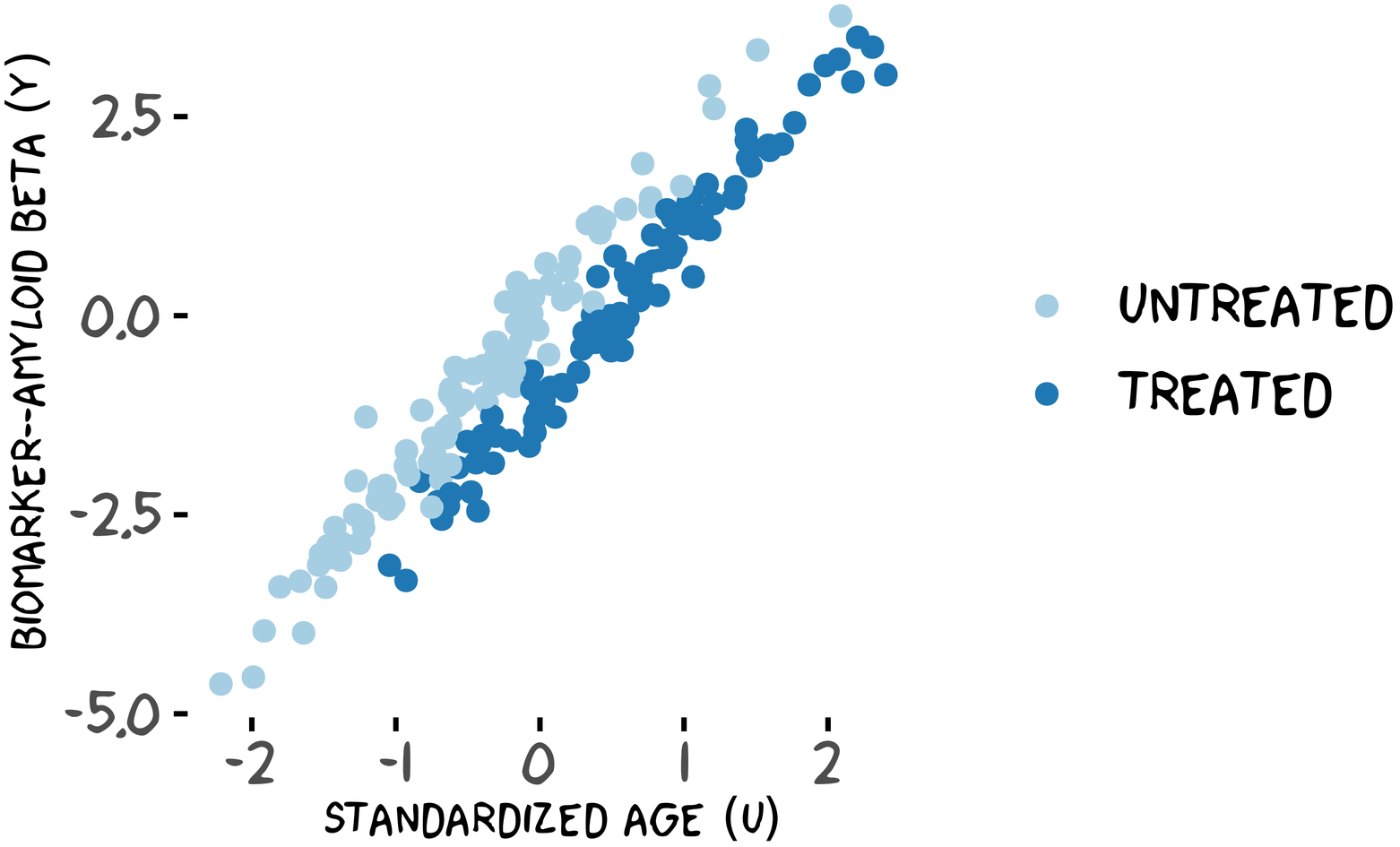


Prevalence of AD increase with age



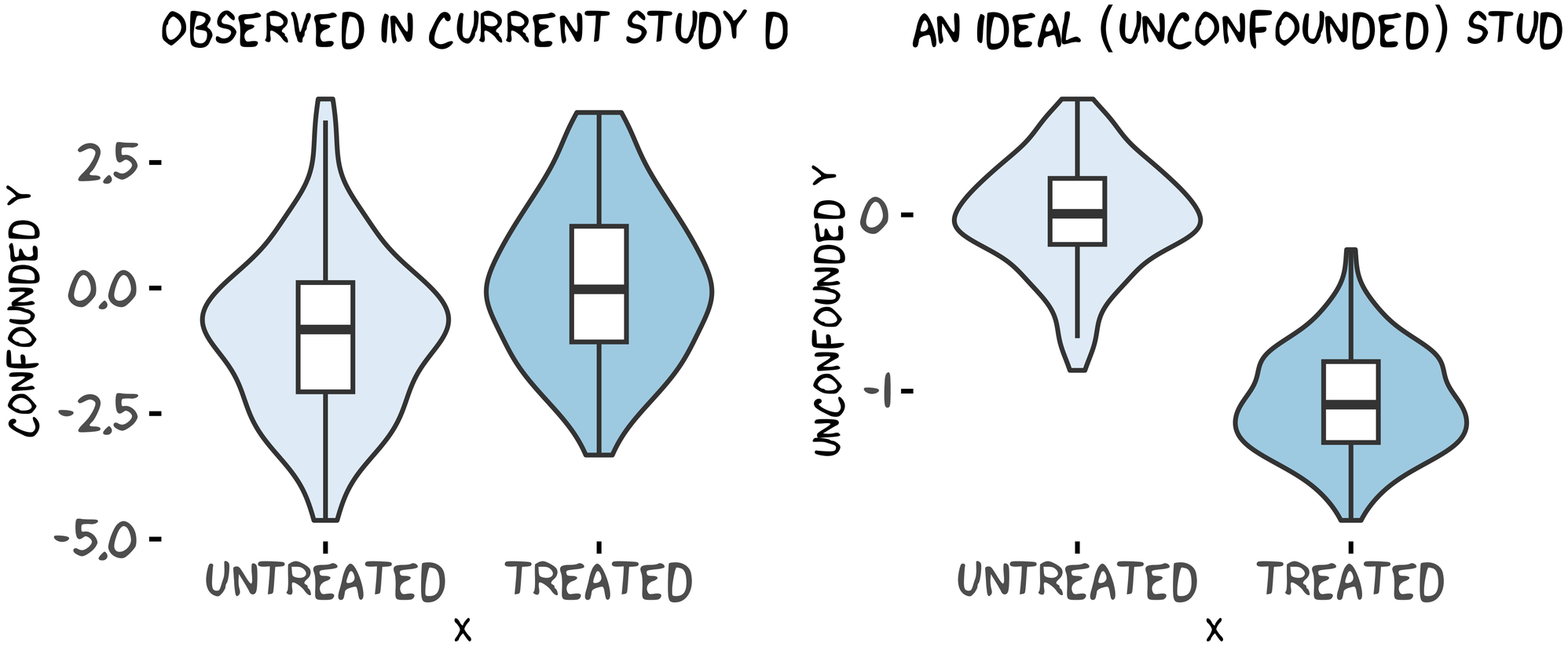
Doctors could have prescribed our drug more frequently to older people because the prevalence of AD tends to increase with age.

Prevalence of AD increase with age



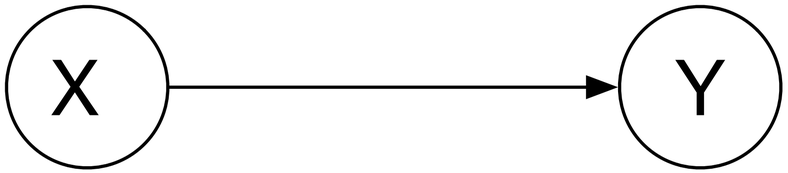
Doctors could have prescribed our drug more frequently to older people because the prevalence of AD tends to increase with age.

Only if we could observe the unconfounded data



What could have been an ideal experimental design?

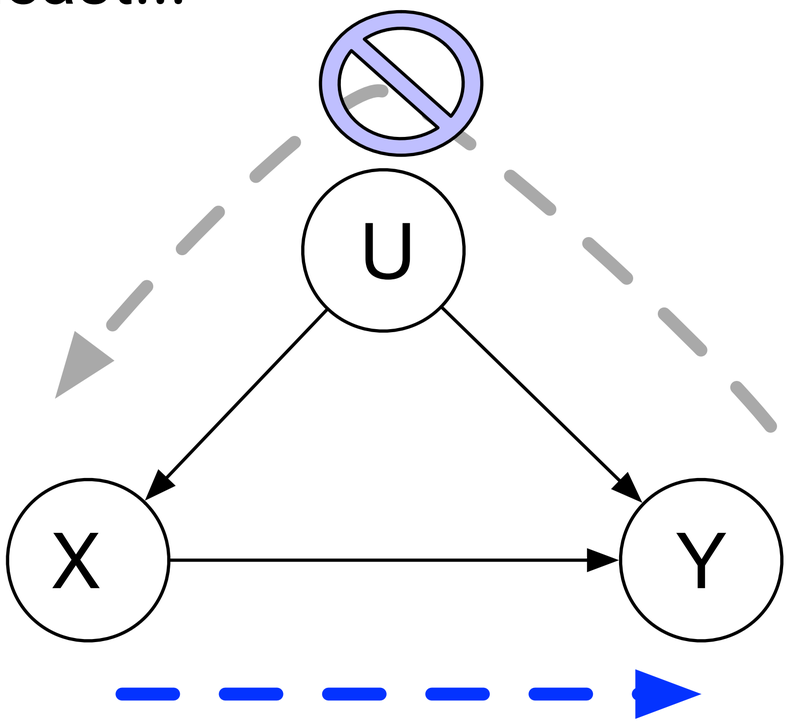
There is no "backdoor" between a putative cause variable (X) *and outcome variable* (Y)*.*



What could have been an ideal experimental design?

There is no "backdoor" between a putative cause variable (X) *and outcome variable* (Y)*.*

At least...



RCT: an ideal experiment (feat. RA Fisher)

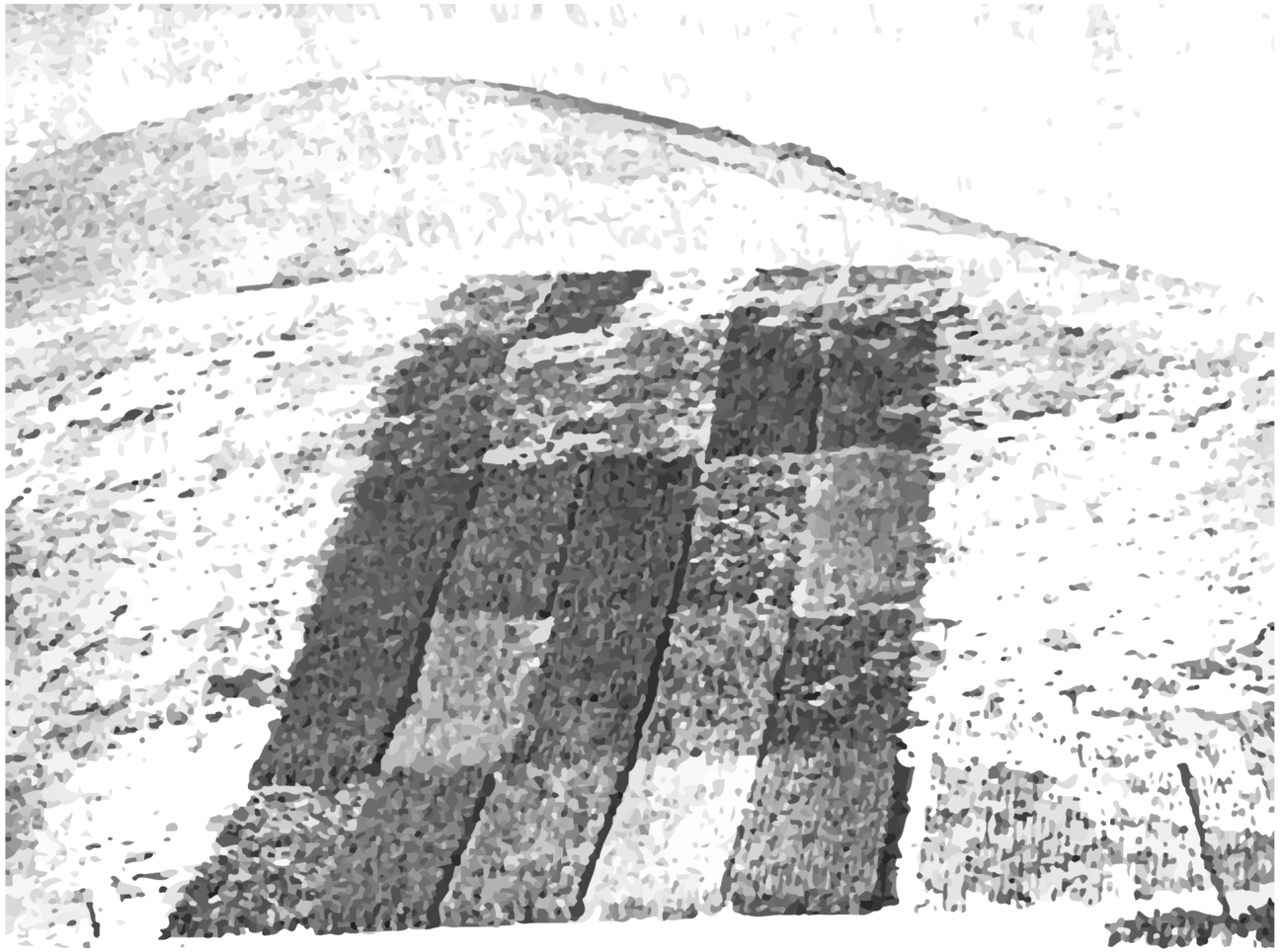
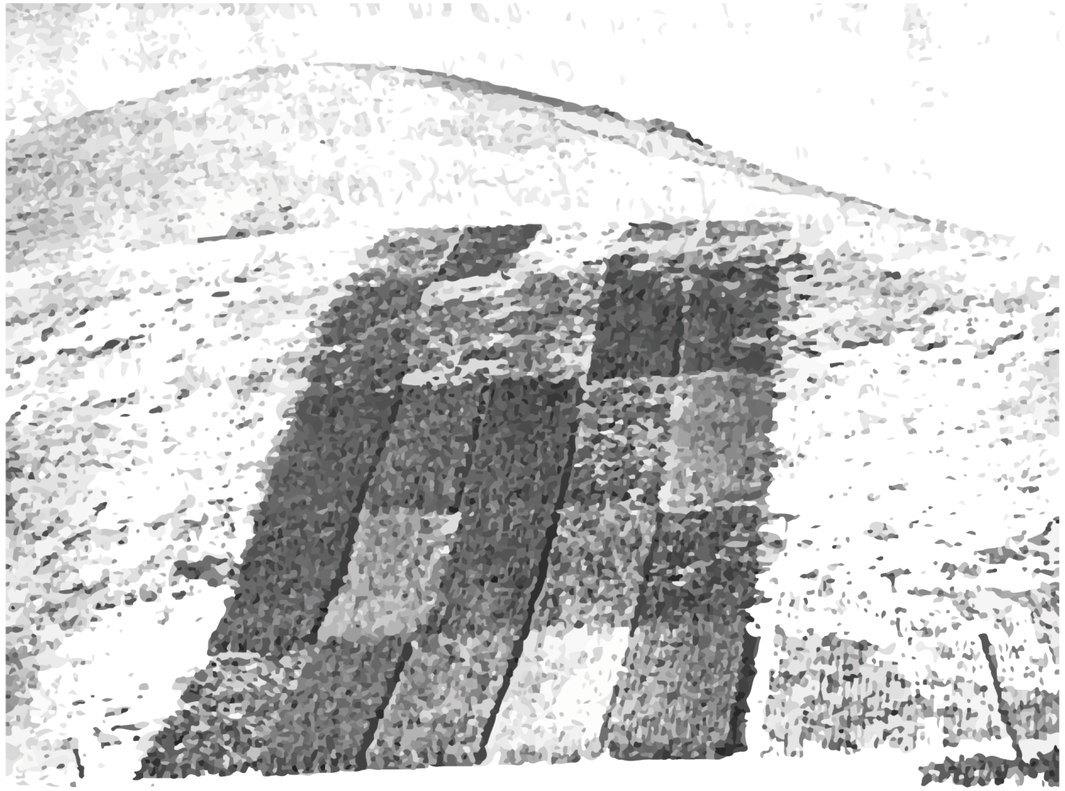


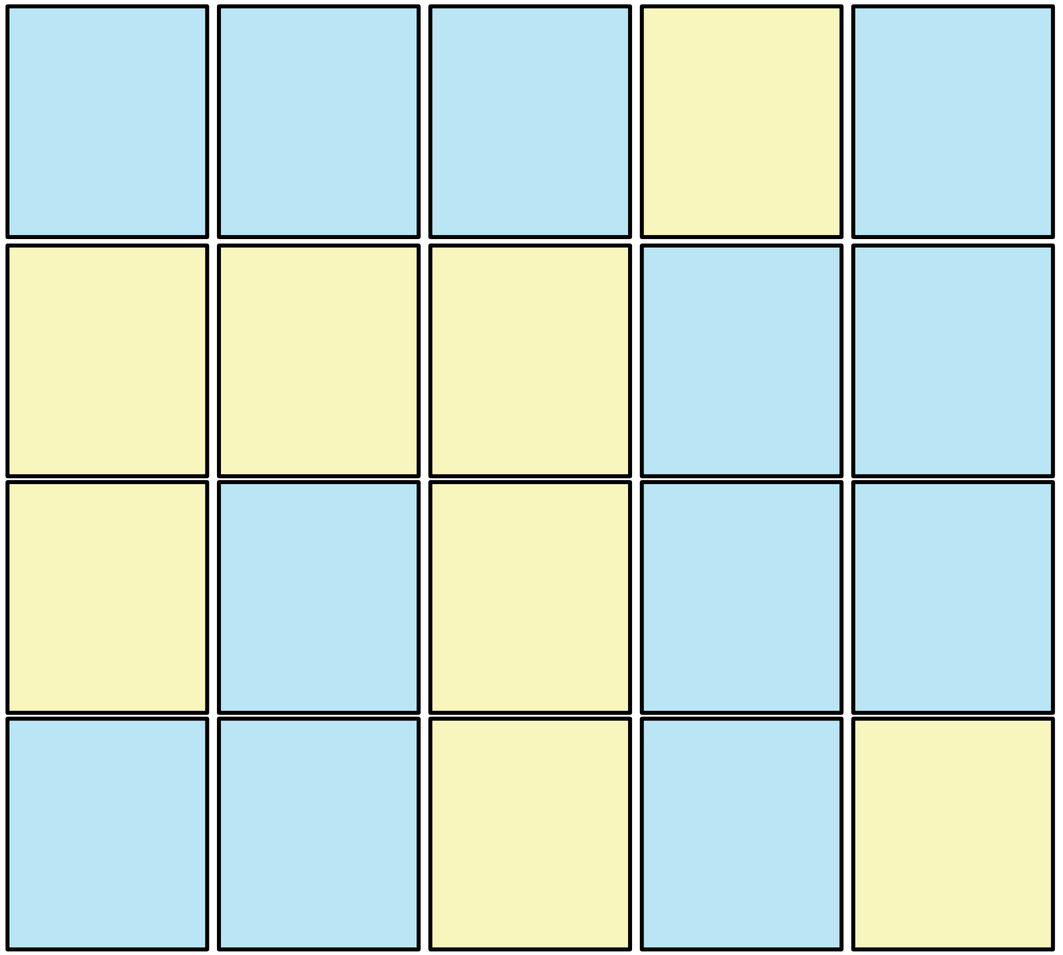
Image source: ‘www.adelaide.edu.au‘

RCT: an ideal experiment (feat. RA Fisher)

Crops grown in the plots

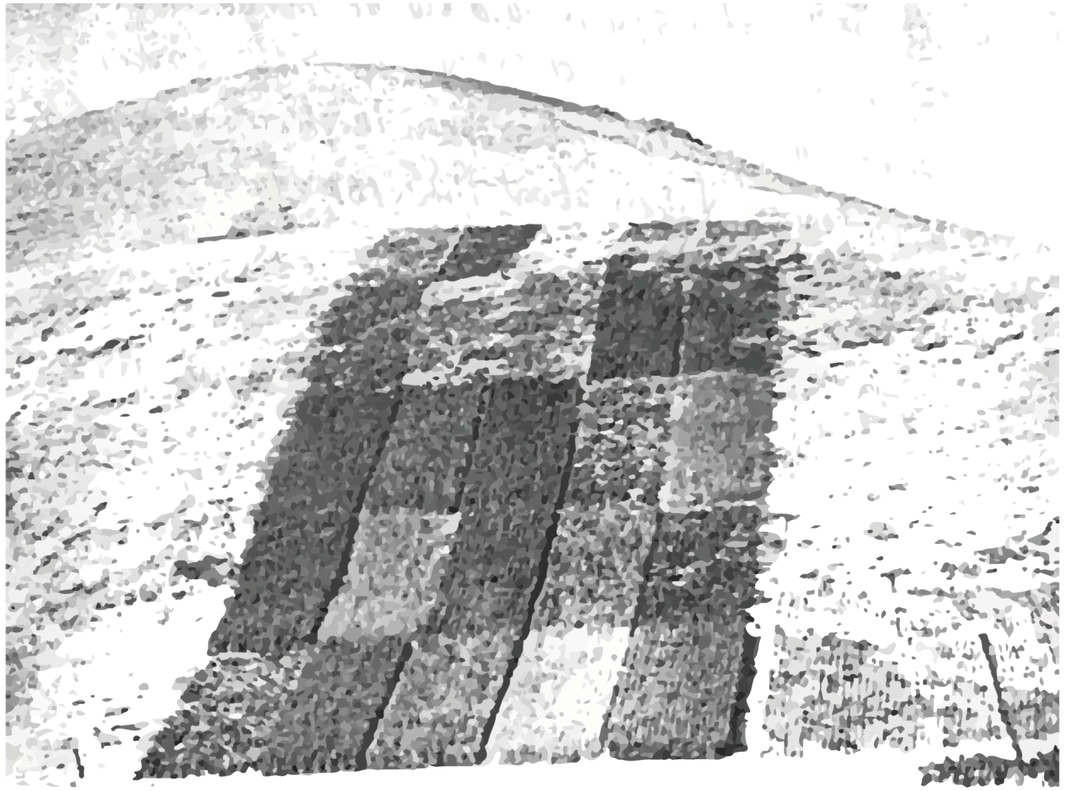


Experimental design

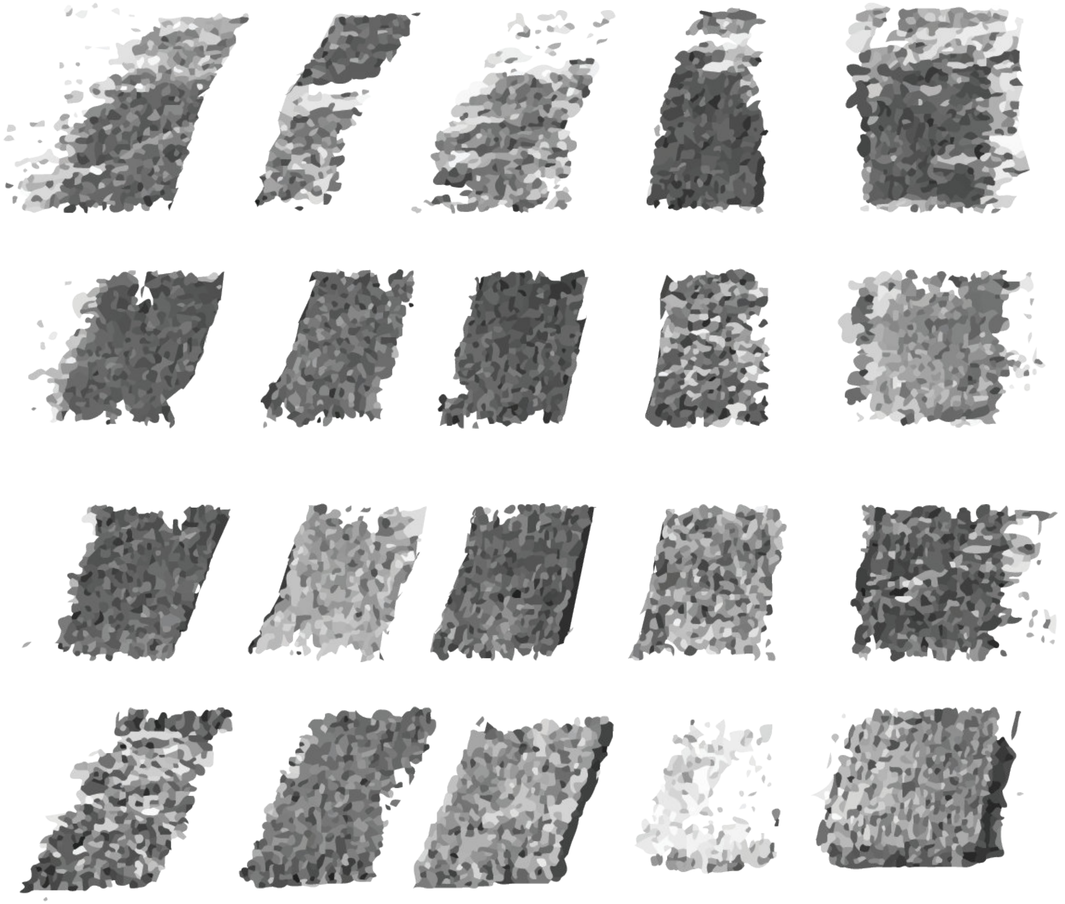


RCT: an ideal experiment (feat. RA Fisher)

Crops grown in the plots

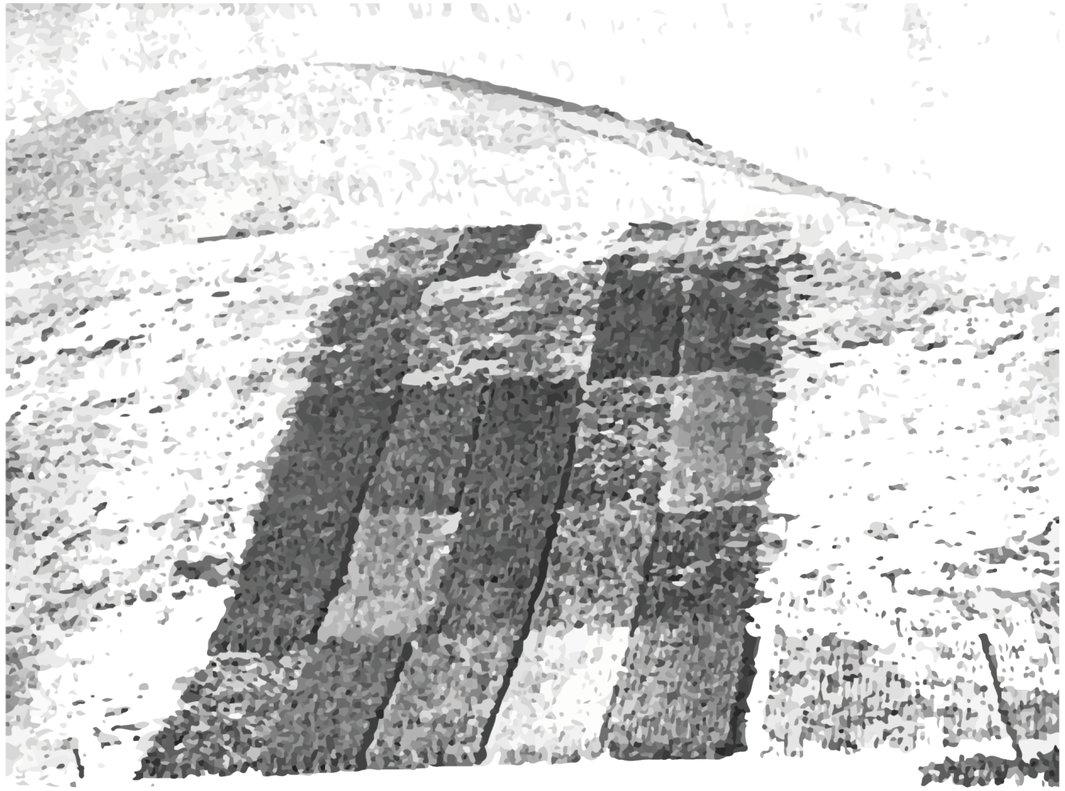


Random assignment of fertilizer

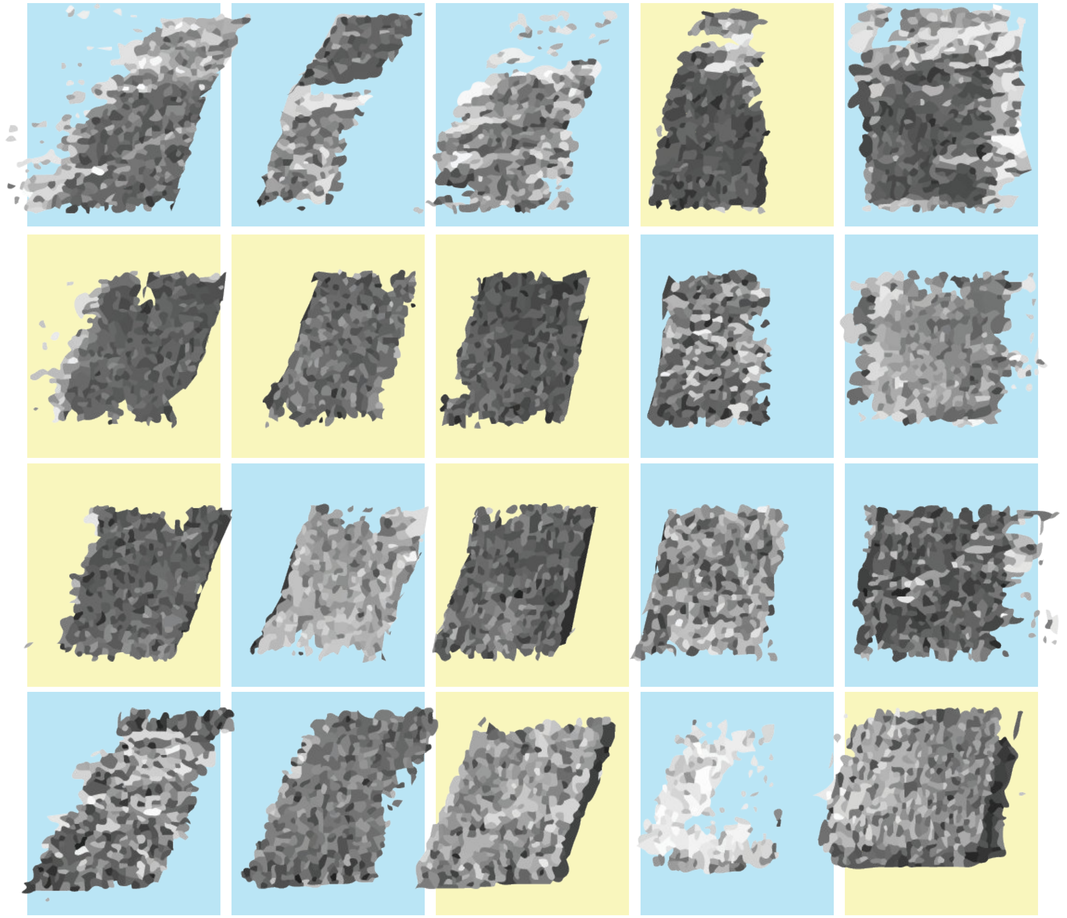


RCT: an ideal experiment (feat. RA Fisher)

Crops grown in the plots

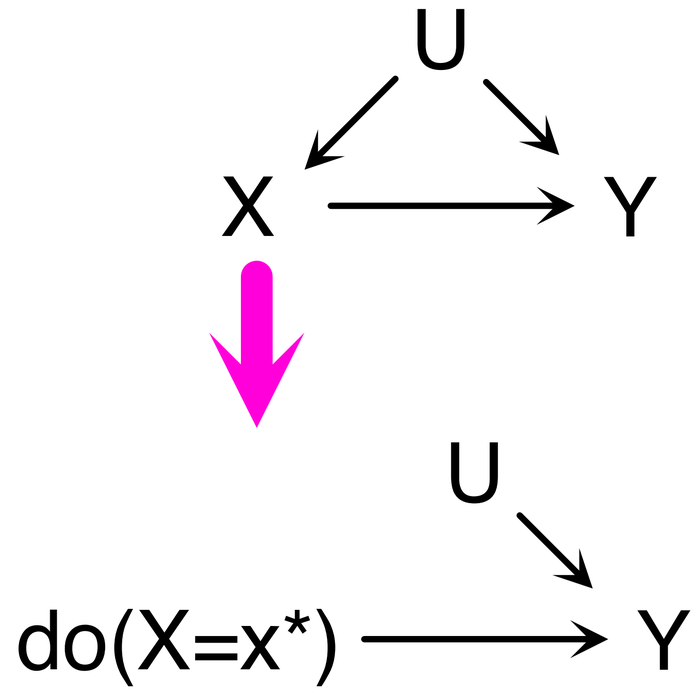


The treated vs. untreated

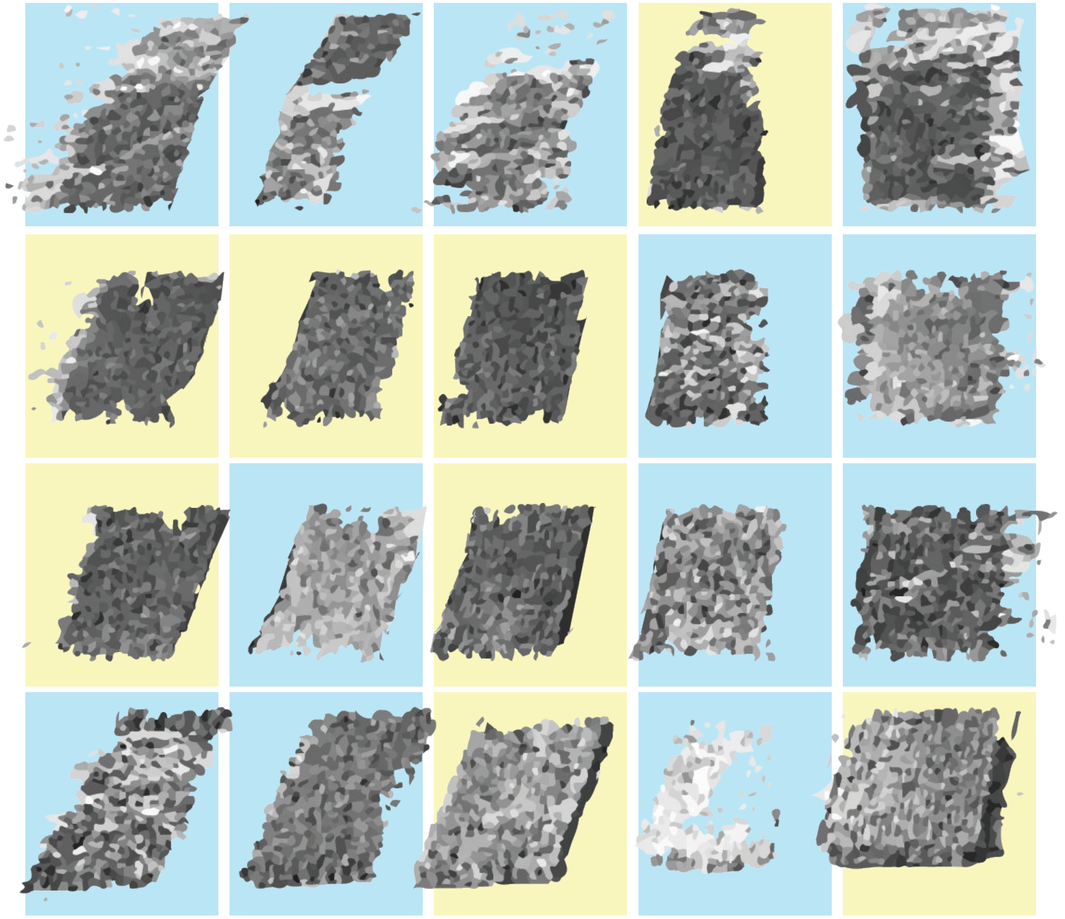


RCT: an ideal experiment (feat. RA Fisher)

Randomized Control Trial $\Longrightarrow$ **"doing"**

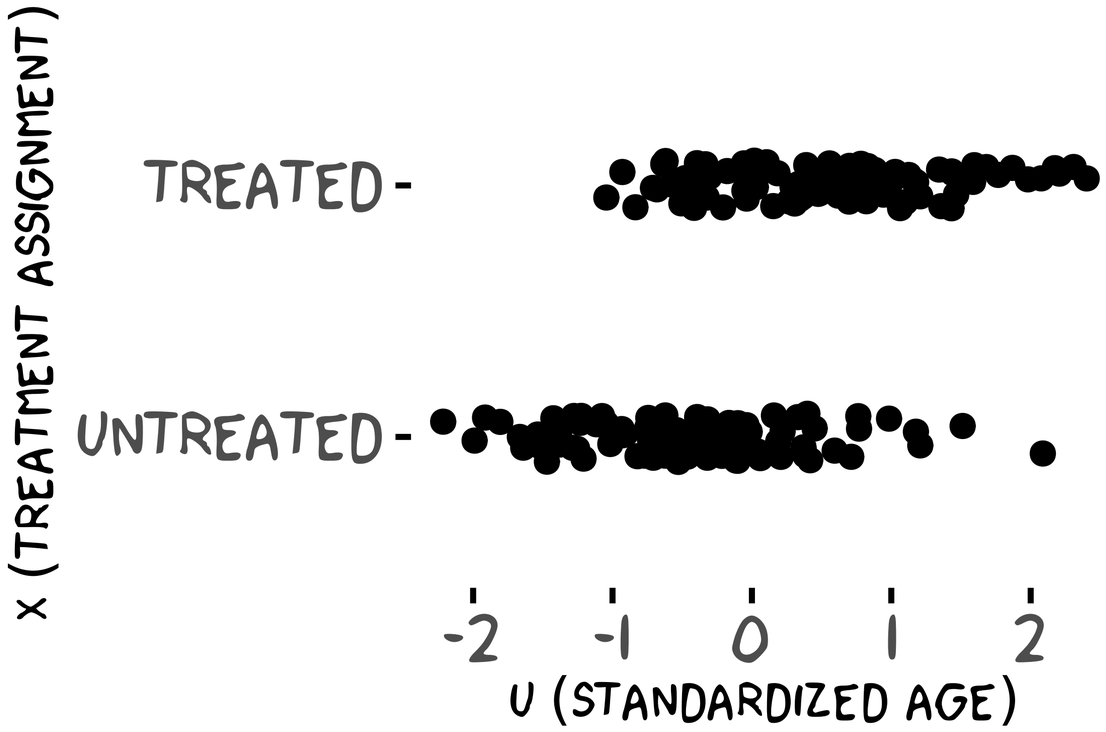


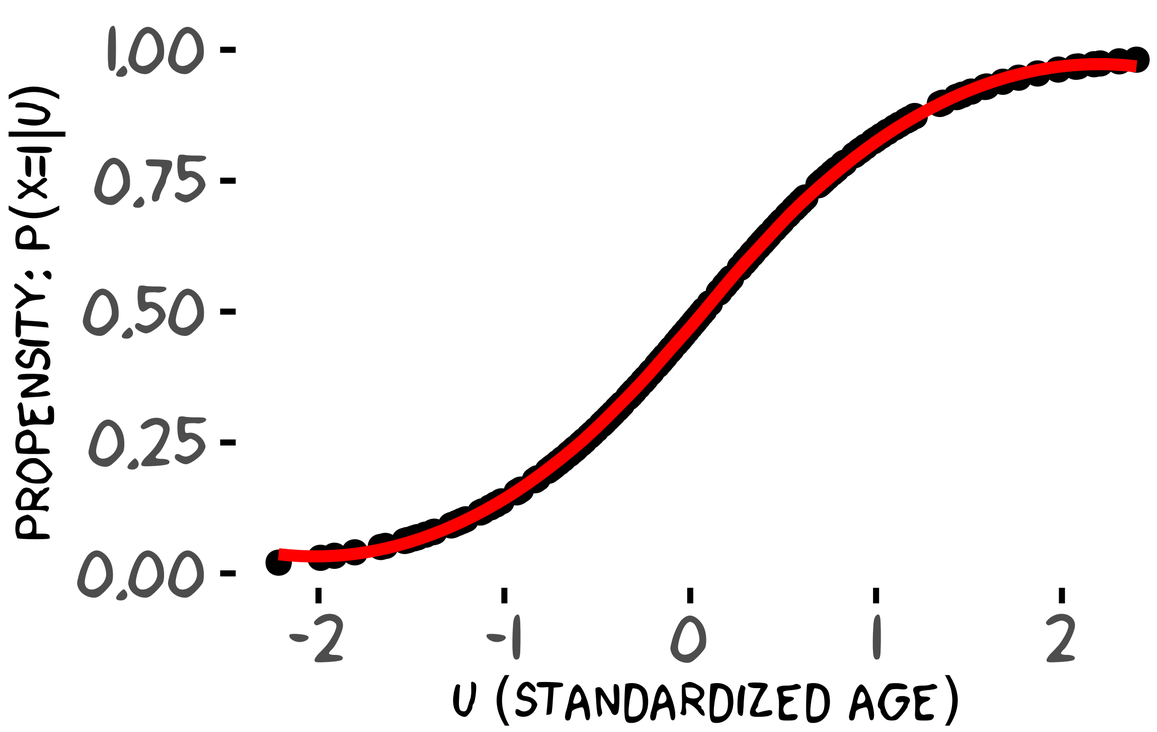
The treated vs. untreated



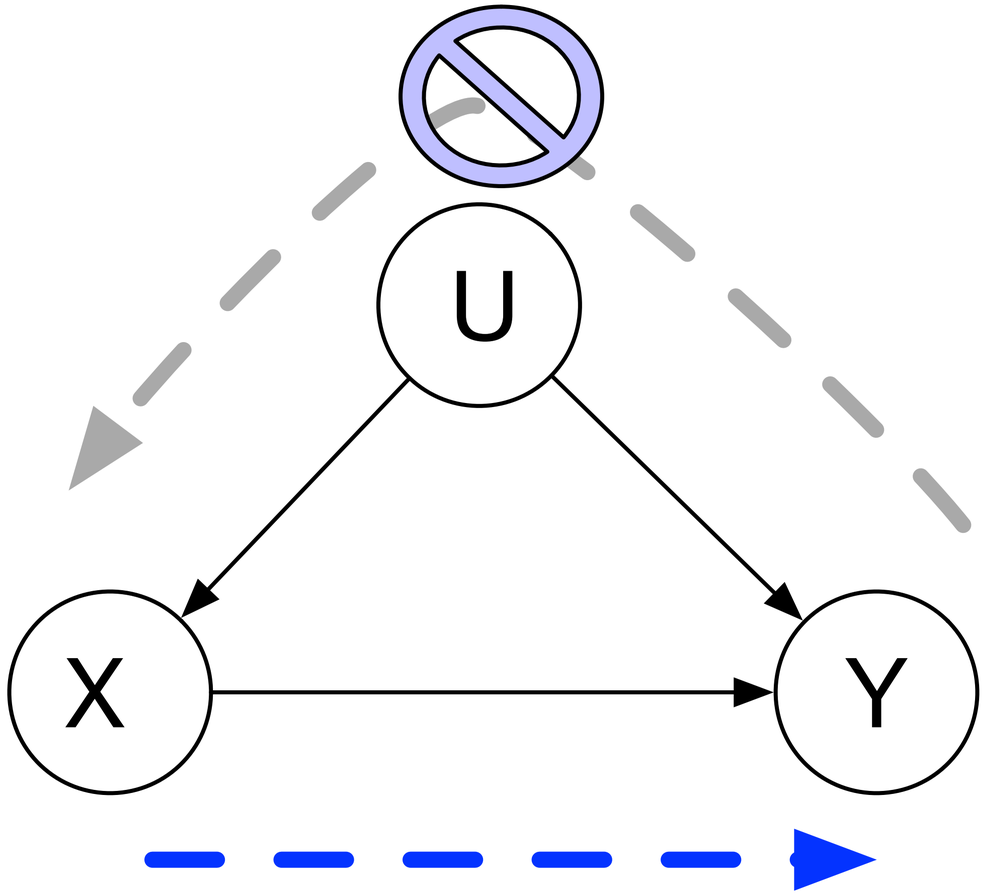
Going back to the toy example: Can we save the study?

Had there been any systematic bias, thou shall adjust it.





Inverse Propensity Weighting “reverse” confounding




```r
p.xu <- glm(x ~ u, family = "binomial") %>%
    predict() %>%
    sigmoid() %>%
    clamp()  # avoid strict 0/1

ww <- x/p.xu + (1 - x)/(1 - p.xu)
```

Propensity: Pr. of assignment $X = 1$:

$$\log \frac{p(X|U)}{1 - p(X|U)} \approx \beta_0 + \beta_1 U$$

Inverse Propensity Weighting “reverse” confounding

$$\hat{Y}_i^{(1)} = \frac{X_i Y_i}{\hat{p}(X_i = 1 | U_i)}$$

$$\hat{Y}_i^{(0)} = \frac{(1 - X_i) Y_i}{1 - \hat{p}(X_i = 1 | U_i)}$$

```r
p.xu <- glm(x ~ u, family = "binomial") %>%
    predict() %>%
    sigmoid() %>%
    clamp()  # avoid strict 0/1

ww <- x/p.xu + (1 - x)/(1 - p.xu)
```

Propensity: Pr. of assignment $X = 1$:

$$\log \frac{p(X|U)}{1 - p(X|U)} \approx \beta_0 + \beta_1 U$$

Inverse Propensity Weighting “reverse” confounding

$$\hat{Y}_i^{(1)} = \frac{X_i Y_i}{\hat{p}(X_i = 1 | U_i)}$$

$$\hat{Y}_i^{(0)} = \frac{(1 - X_i) Y_i}{1 - \hat{p}(X_i = 1 | U_i)}$$

equivalently give weights for $\forall$ i

$$W_i \propto \begin{cases} 1/\hat{p}(X_i = 1 | U_i) & X_i = 1 \\ 1/\hat{p}(X_i = 0 | U_i) & X_i = 0 \end{cases}$$

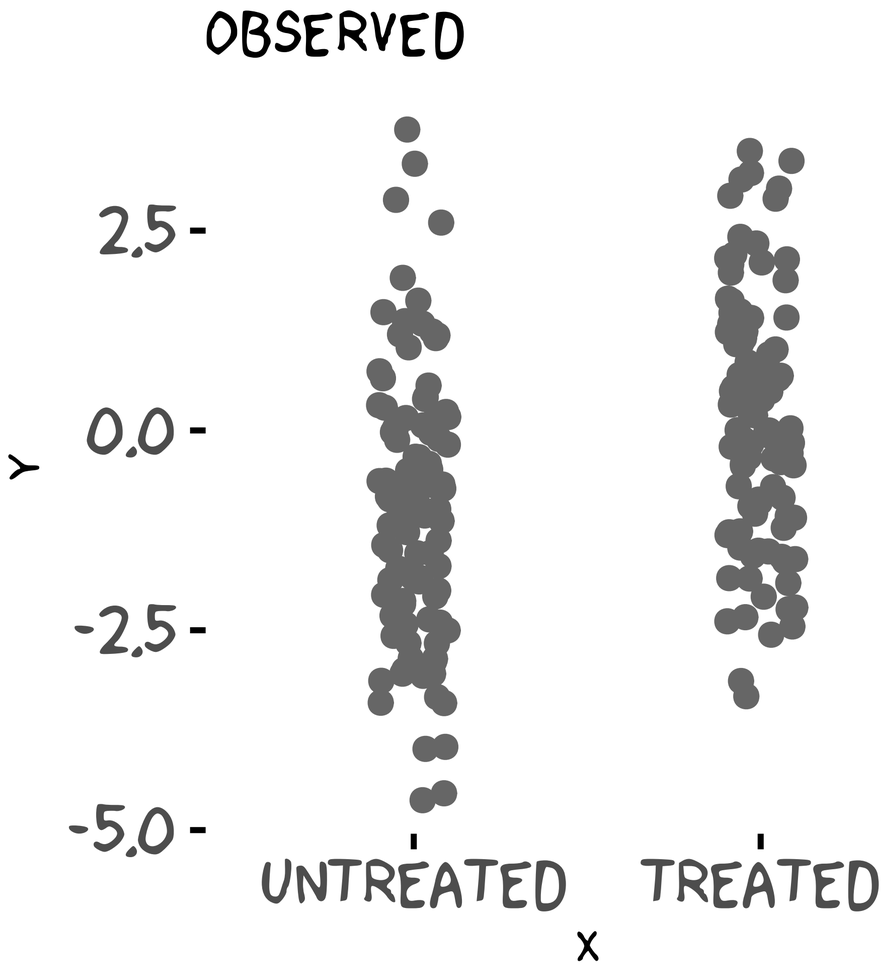
```r
p.xu <- glm(x ~ u, family = "binomial") %>%
    predict() %>%
    sigmoid() %>%
    clamp()  # avoid strict 0/1

ww <- x/p.xu + (1 - x)/(1 - p.xu)
```

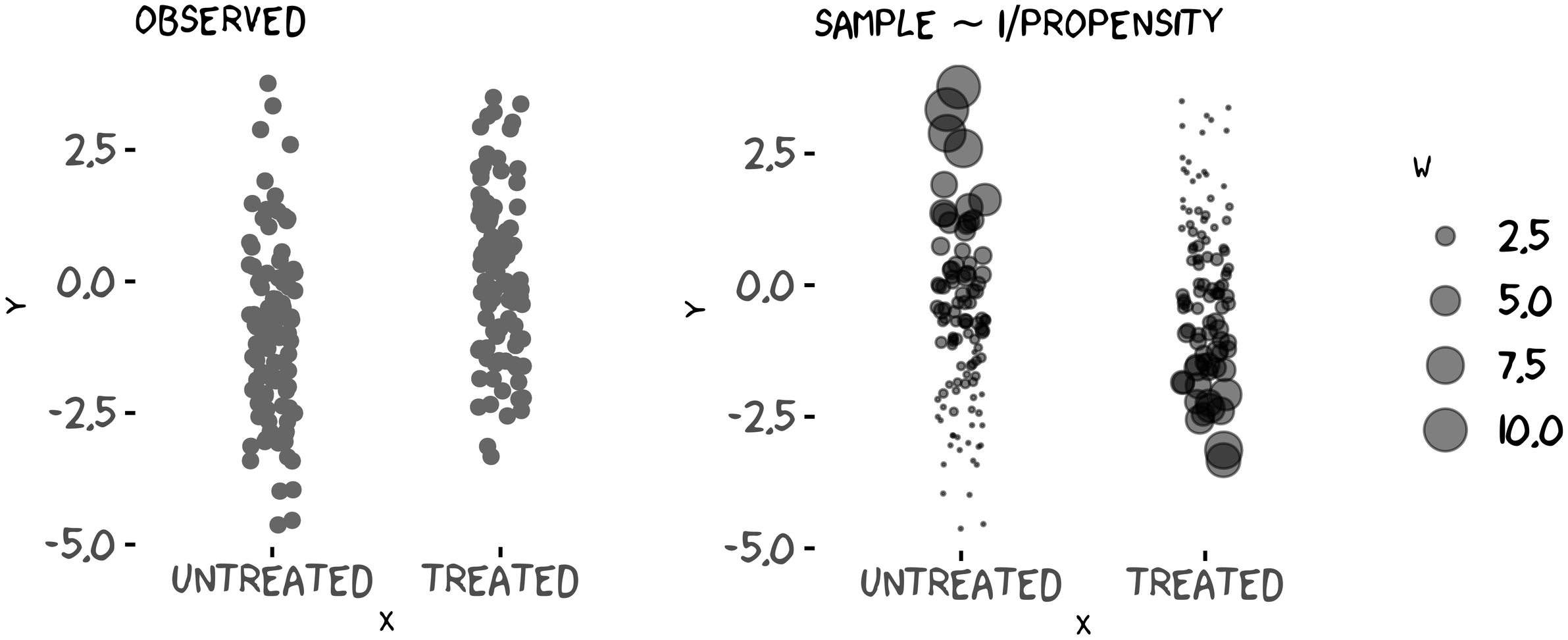
Propensity: Pr. of assignment $X = 1$:

$$\log \frac{p(X|U)}{1 - p(X|U)} \approx \beta_0 + \beta_1 U$$

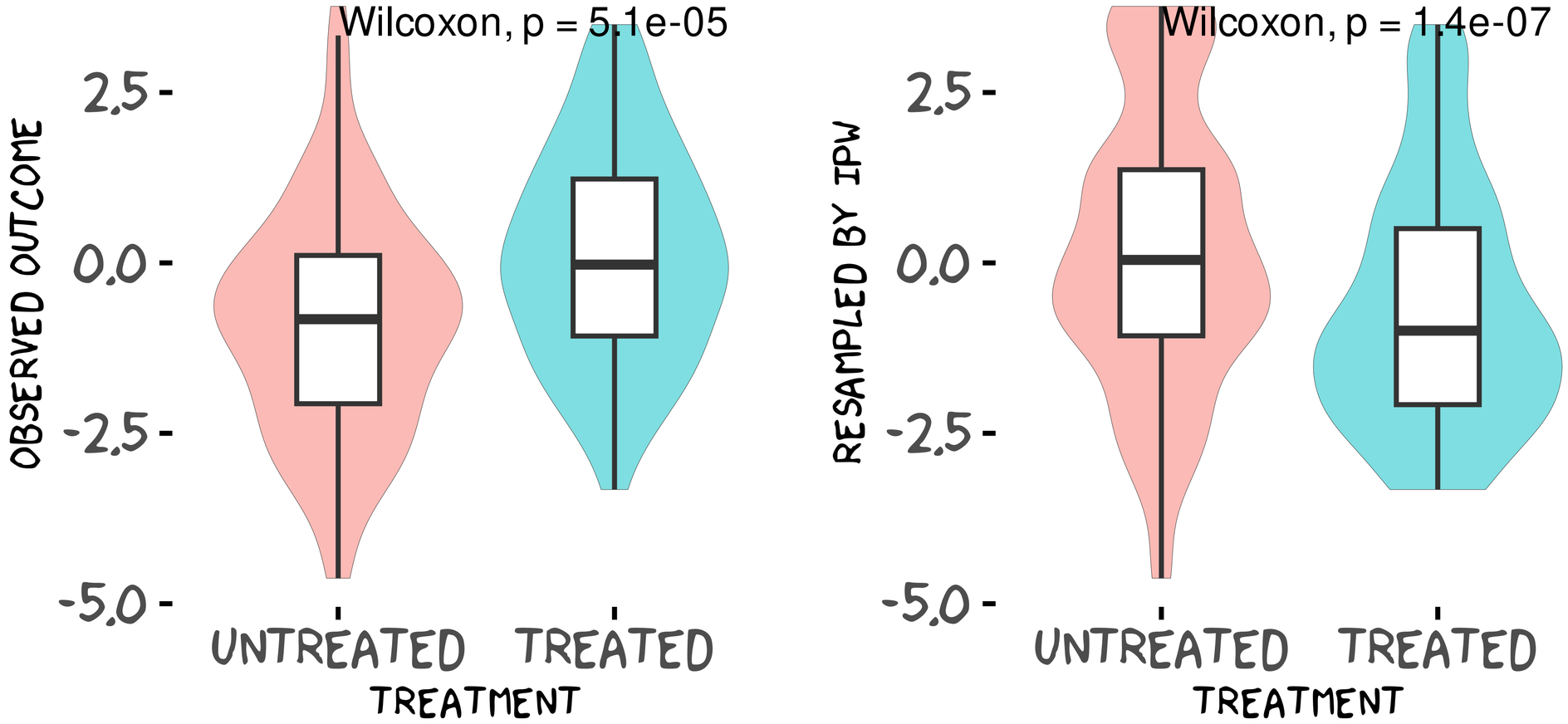
Take samples inversely proportional to propensity



Take samples inversely proportional to propensity



If we bootstrap (sample with replacement)…



Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

we can prove $\mathbb{E}[XY/e(X)] \rightarrow \mathbb{E}\left[Y^{(1)}\right]$

using

- Strong ignorability
- Smoothness
- Stable Unit Treatment (exposure) Variable

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] \quad =$$

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\middle| C_i\right]\right] \quad \text{(law of total expectation)}$$

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\middle| C_i\right]\right] \quad \text{(law of total expectation)}$$

$$\text{(C is sufficient backdoor)} \quad = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i^{(1)}}{e(C_i)}\middle| C_i\right]\right]$$

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\middle| C_i\right]\right] \quad \text{(law of total expectation)}$$

$$\text{(C is sufficient backdoor)} \quad = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i^{(1)}}{e(C_i)}\middle| C_i\right]\right]$$

$$\text{(strong ignorability)} \quad = \mathbb{E}\left[Y_i^{(1)}\mathbb{E}\left[\frac{X_i}{e(C_i)}\middle| C_i\right]\right] \quad Y^{(1)} \perp\!\!\!\perp X|C$$

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\middle| C_i\right]\right] \quad \text{(law of total expectation)}$$

$$\text{(C is sufficient backdoor)} \quad = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i^{(1)}}{e(C_i)}\middle| C_i\right]\right]$$

$$\text{(strong ignorability)} \quad = \mathbb{E}\left[Y_i^{(1)}\mathbb{E}\left[\frac{X_i}{e(C_i)}\middle| C_i\right]\right] \quad Y^{(1)} \perp\!\!\!\perp X|C$$

$$\text{(smoothness)} \quad = \mathbb{E}\left[Y_i^{(1)}\frac{1}{\cancel{e(C_i)}}\cancel{\mathbb{E}[X_i|C_i]}\right] \quad 0 < e(C) < 1$$

Why IPW works? Unbiased estimate potential outcome

Letting $e(z) = \hat{p}(X = 1|C = z)$,

$$\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i}{e(C_i)}\middle| C_i\right]\right] \quad \text{(law of total expectation)}$$

$$\text{(C is sufficient backdoor)} \quad = \mathbb{E}\left[\mathbb{E}\left[\frac{X_i Y_i^{(1)}}{e(C_i)}\middle| C_i\right]\right]$$

$$\text{(strong ignorability)} \quad = \mathbb{E}\left[Y_i^{(1)}\mathbb{E}\left[\frac{X_i}{e(C_i)}\middle| C_i\right]\right] \quad Y^{(1)} \perp\!\!\!\perp X|C$$

$$\text{(smoothness)} \quad = \mathbb{E}\left[Y_i^{(1)}\frac{1}{\cancel{e(C_i)}}\cancel{\mathbb{E}[X_i|C_i]}\right] \quad 0 < e(C) < 1$$

$$= \mathbb{E}\left[Y^{(1)}\right]$$

When do we use IPW to estimate potential outcomes?

- Backdoor variables are sufficiently characterized

When do we use IPW to estimate potential outcomes?

- Backdoor variables are sufficiently characterized
- We have an unbiased way to estimate the propensity model

When do we use IPW to estimate potential outcomes?

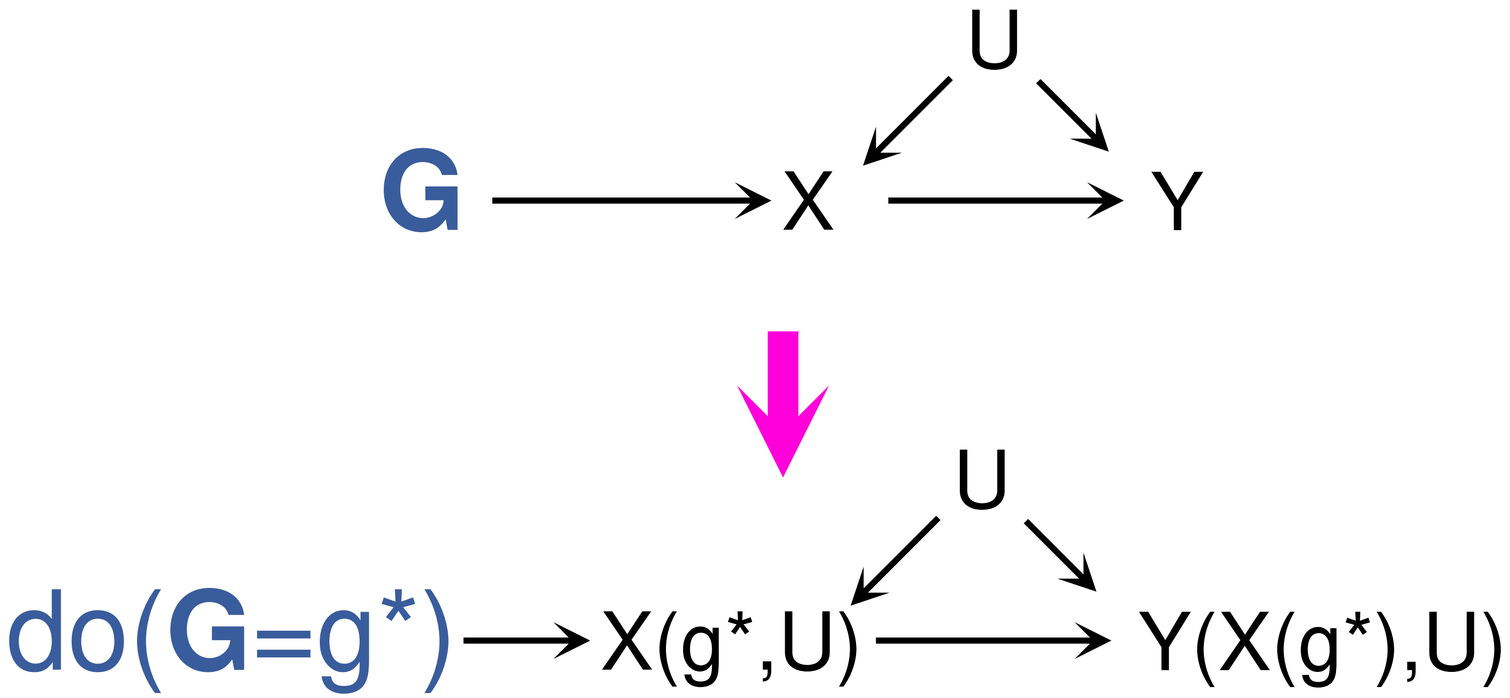
- Backdoor variables are sufficiently characterized
- We have an unbiased way to estimate the propensity model
- Smooth overlap $1 > e(X) > 0$

Today's lecture

1. Summary Statistics-based post-GWAS
2. Intro to causal inference
3. **Mendelian Randomization**

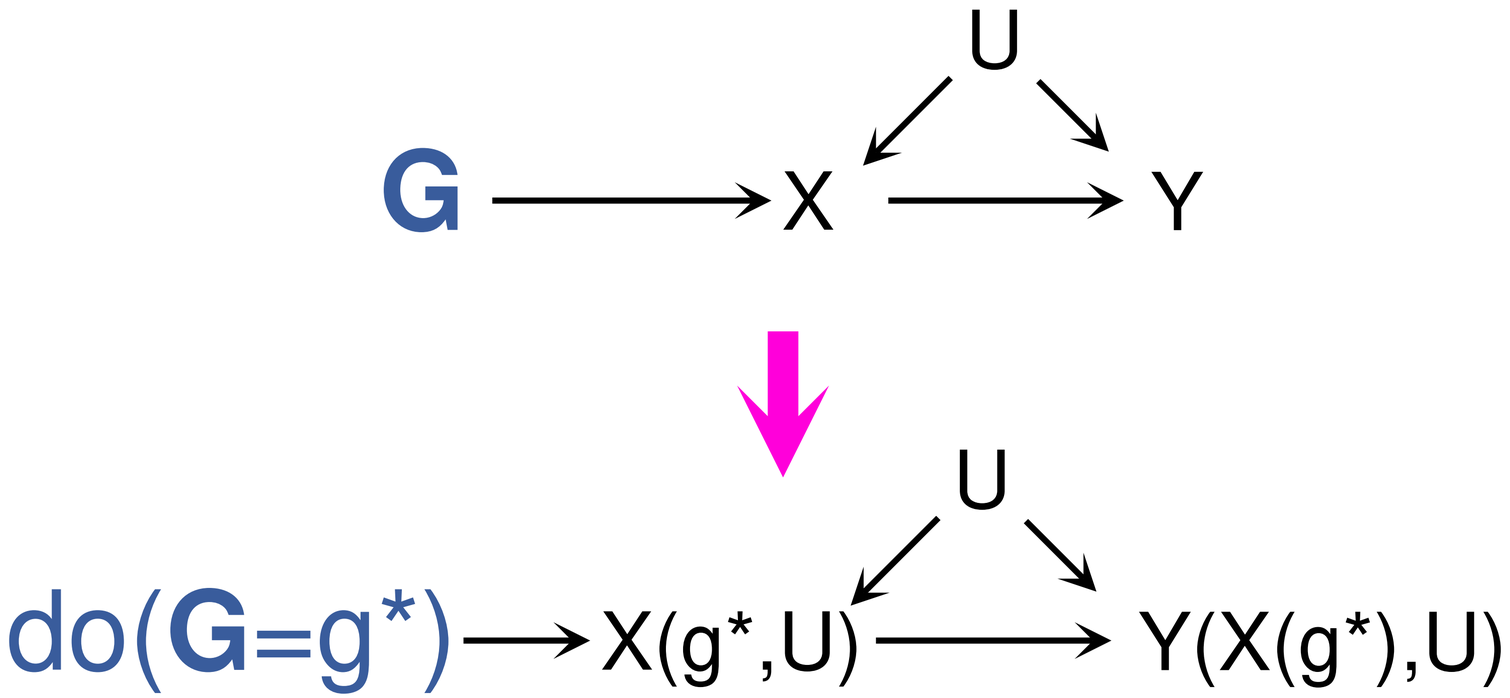
- Can we learn causality from observational data?
- What’s so special about genetic variants?

Mendelian Randomization



We will use a genetic variable G to mimic RCT

Mendelian Randomization



If a genetic variable G is a valid instrumental variable...

MR in modern epidemiology studies

- G: genotype
- X: APOE protein
- Y: cancer

$$G \rightarrow X \rightarrow Y$$

APOLIPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER

SIR,—It is unclear whether the relation between low serum cholesterol levels and cancer[1] is causal. In many studies occult tumour may have depressed cholesterol levels though in others the relation was found when serum cholesterol had been measured many years before the cancer was diagnosed. The relation is probably not explained by diet, because in the Seven Countries Study cohorts with widely different diets and corresponding differences in mean cholesterol levels experienced similar mean cancer rates.[2,3] On the other hand, within each region cancer incidence was higher in men with a serum cholesterol in the lowest part of the cholesterol distribution for that country.[3] Thus, naturally low cholesterol levels are sometimes associated with increased cancer risk.[1,3]

Differences in the aminoacid sequence of apolipoprotein E (apo

Katan, *Lancet*, (1986)

MR in modern epidemiology studies

- G: genotype
- X: APOE protein
- Y: cancer

$$G \rightarrow X \rightarrow Y$$

George Davey Smith

30TH THOMAS FRANCIS JR MEMORIAL LECTURE

'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*

George Davey Smith and Shah Ebrahim

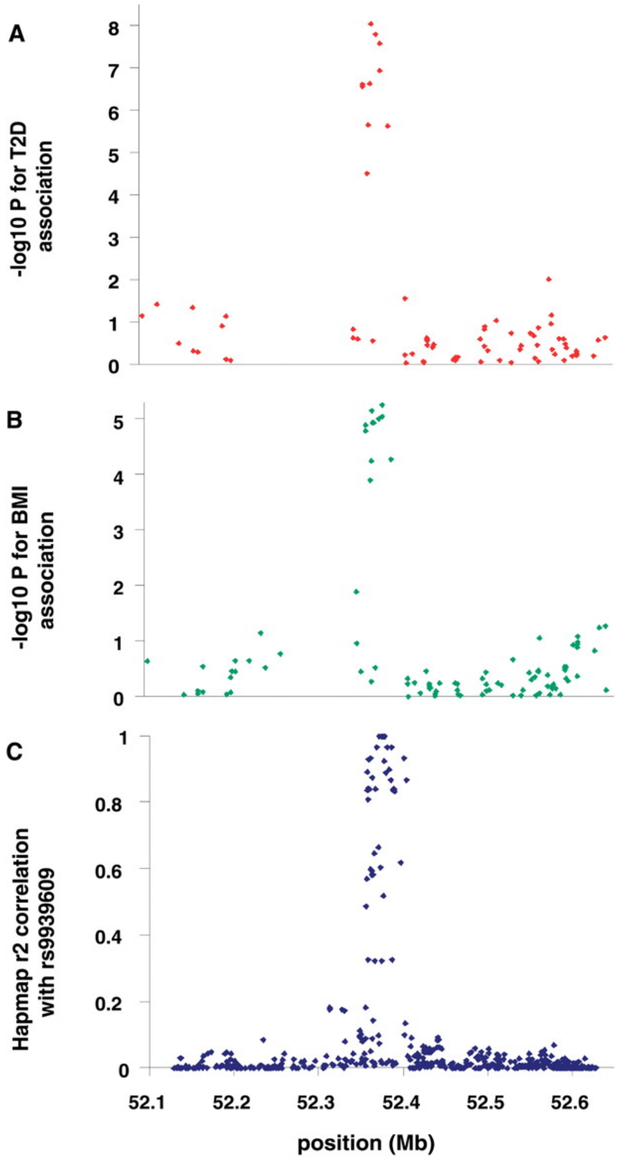
APOLIPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER

SIR,—It is unclear whether the relation between low serum cholesterol levels and cancer[1] is causal. In many studies occult tumour may have depressed cholesterol levels though in others the relation was found when serum cholesterol had been measured many years before the cancer was diagnosed. The relation is probably not explained by diet, because in the Seven Countries Study cohorts with widely different diets and corresponding differences in mean cholesterol levels experienced similar mean cancer rates.[2,3] On the other hand, within each region cancer incidence was higher in men with a serum cholesterol in the lowest part of the cholesterol distribution for that country.[3] Thus, naturally low cholesterol levels are sometimes associated with increased cancer risk.[1,3]

Differences in the aminoacid sequence of apolipoprotein E (apo

Katan, *Lancet*, (1986)

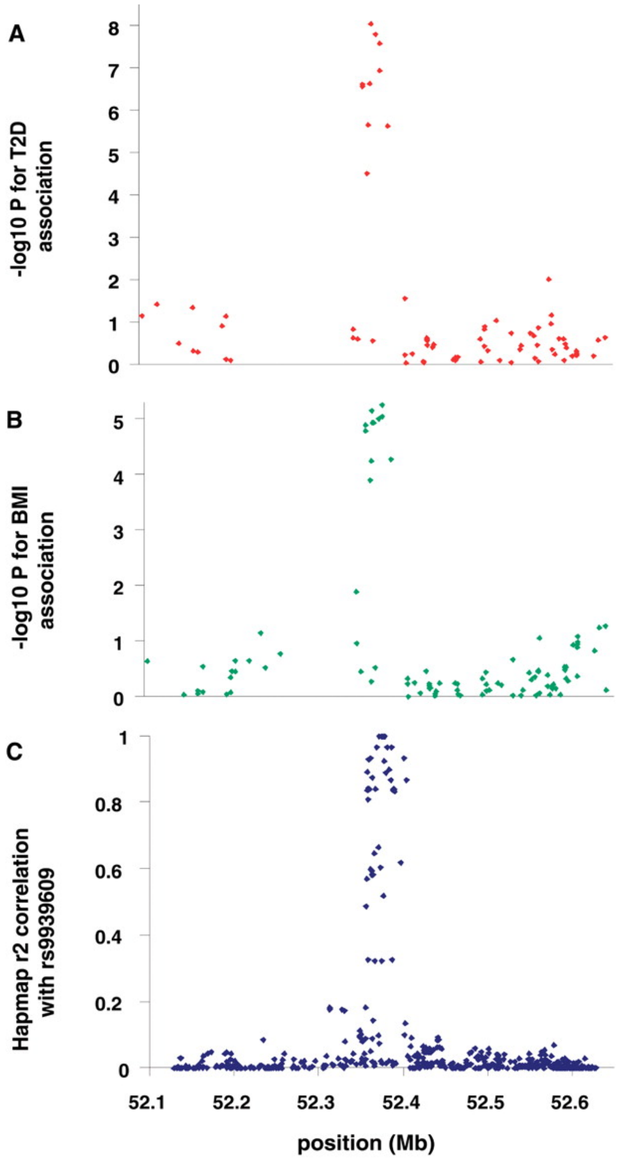
MR in action: *FTO* → fat mass → obesity, diabetes

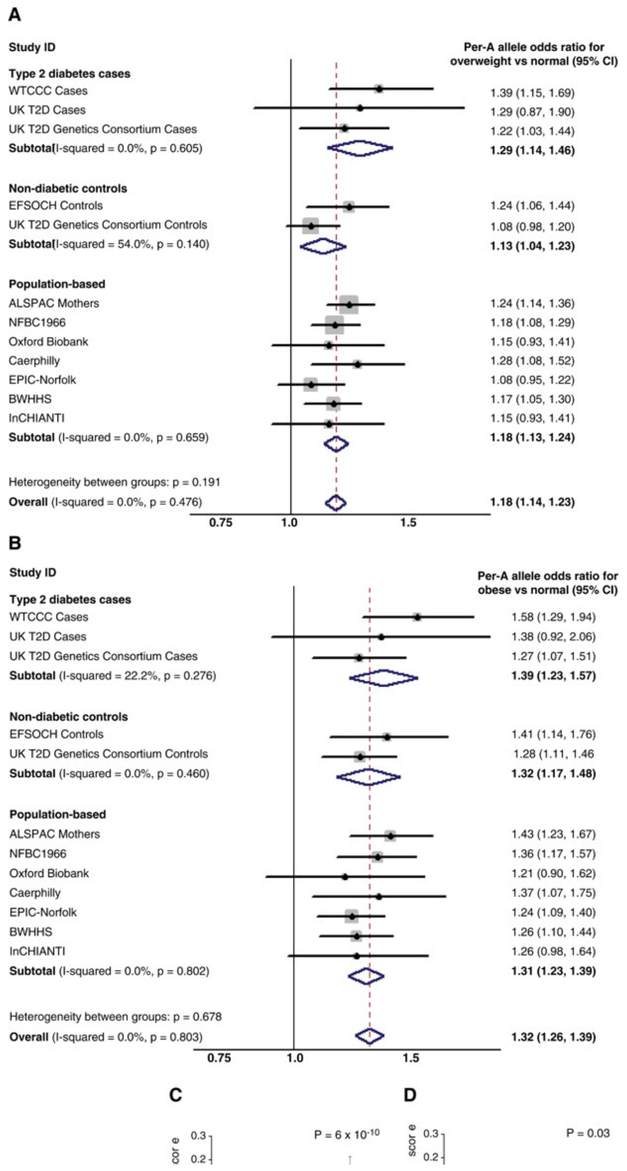


Using genotype as an instrumental variable, we test causality between other exposure variables and downstream phenotypes.

- Genotype in *FTO* locus → T2D

MR in action: *FTO* → fat mass → obesity, diabetes

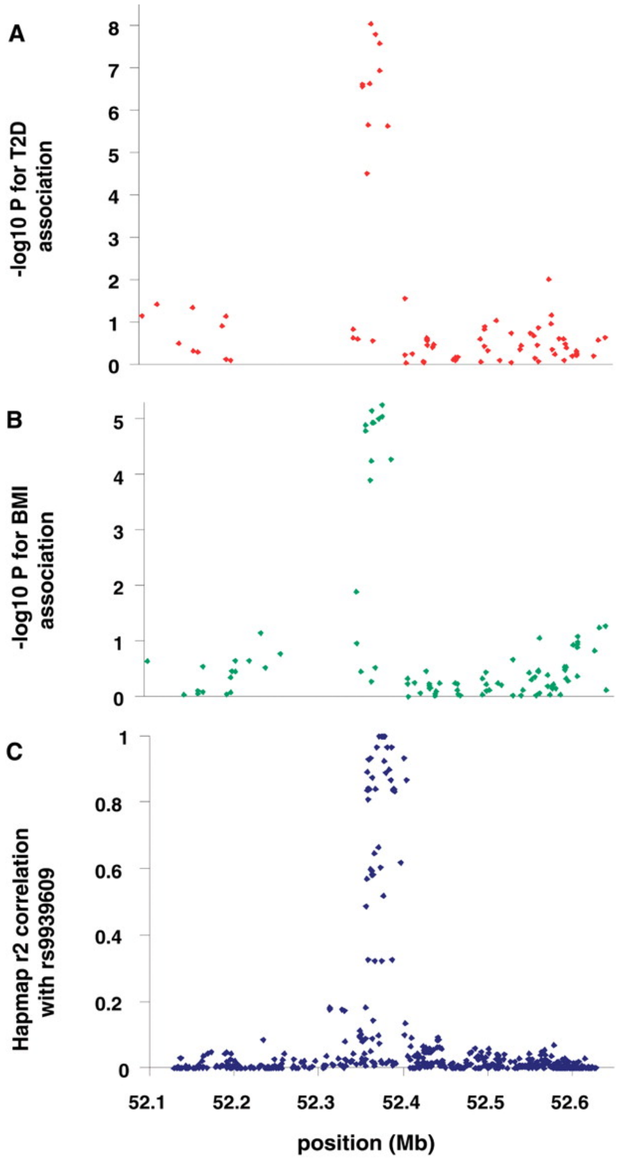


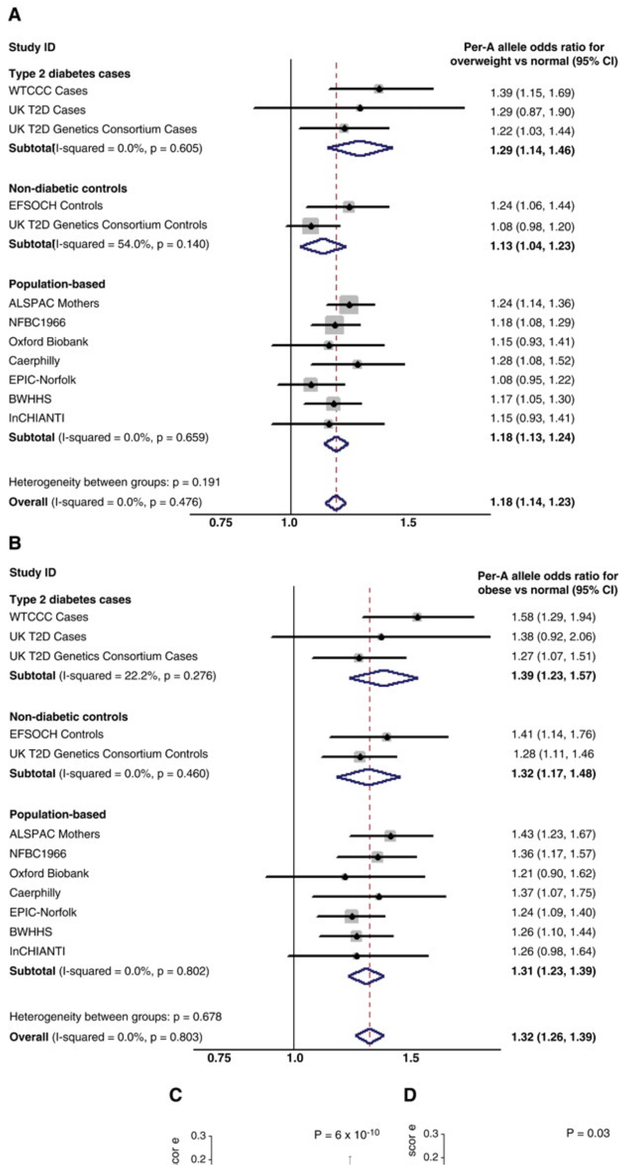


Using genotype as an instrumental variable, we test causality between other exposure variables and downstream phenotypes.

- Genotype in *FTO* locus → T2D
- *FTO* locus → fat mass

MR in action: *FTO* → fat mass → obesity, diabetes

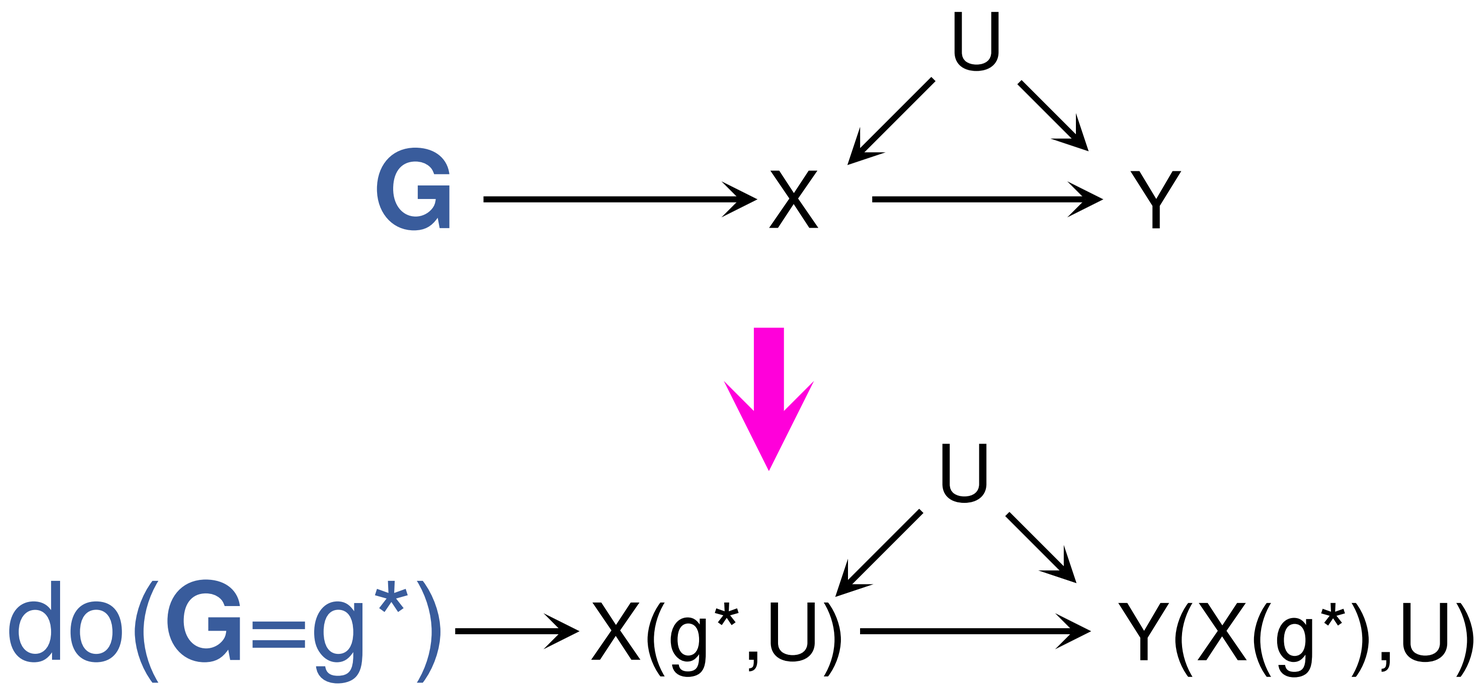




Using genotype as an instrumental variable, we test causality between other exposure variables and downstream phenotypes.

- Genotype in *FTO* locus → T2D
- *FTO* locus → fat mass
- Using *FTO* as ”instrumental variable”, we can ask other MR questions

Why doing Mendelian Randomization?



1. We do not have enough knowledge of U
2. We do not have a way to make interventions on $do(X = x)$

"Genotypes are beautifully randomized" - Fisher (1951)

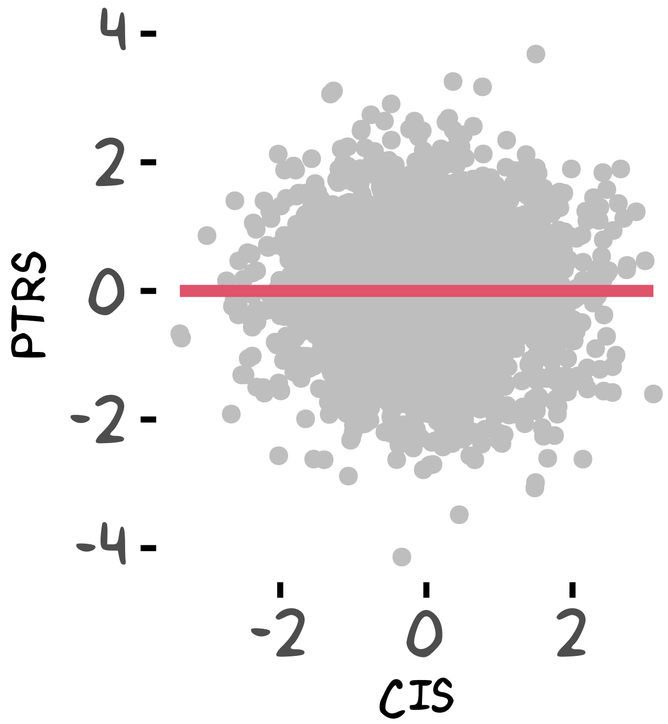
Toy example: How can we learn causality between genetic disorders?

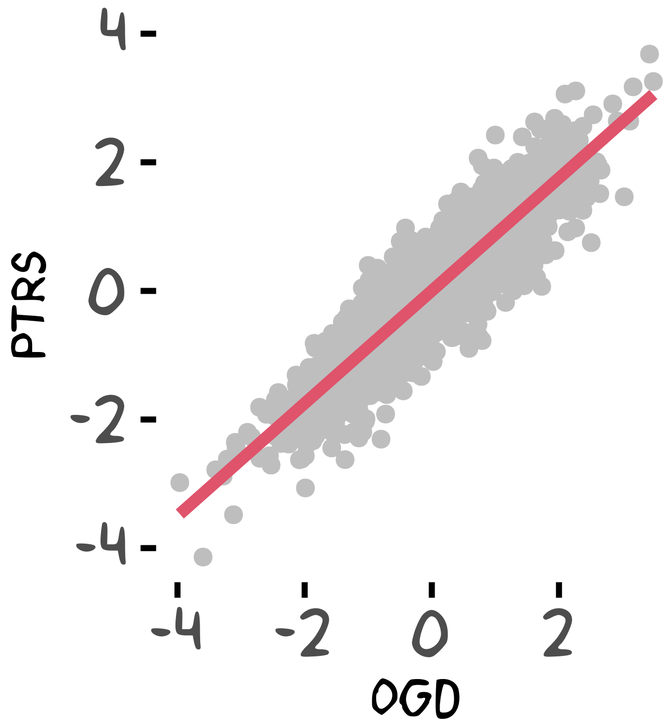
Three disorders (just for demonstration)

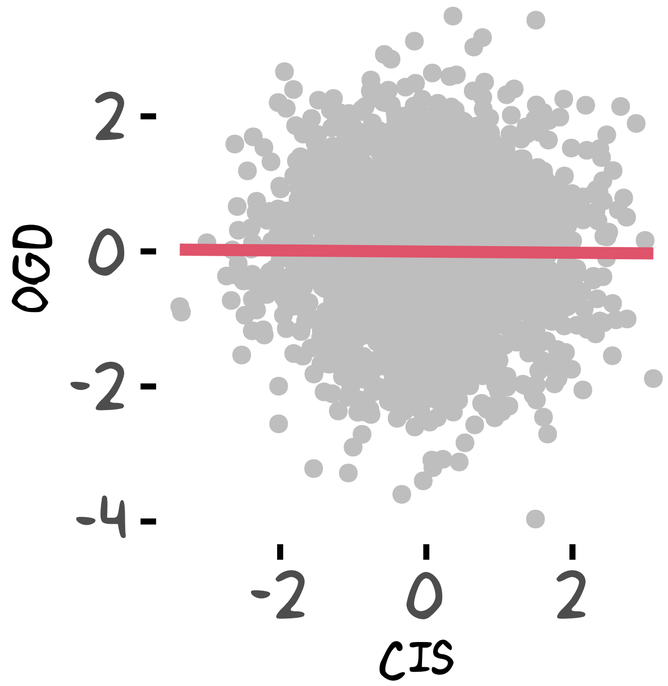
- OGD: Obsessive `ggplot` disorder
- PTRS: Post-traumatic `R` session stress syndrome
- CIS: Chronic indentation syndrome (as a result of `Python` coding)

How can we learn causality between genetic traits?

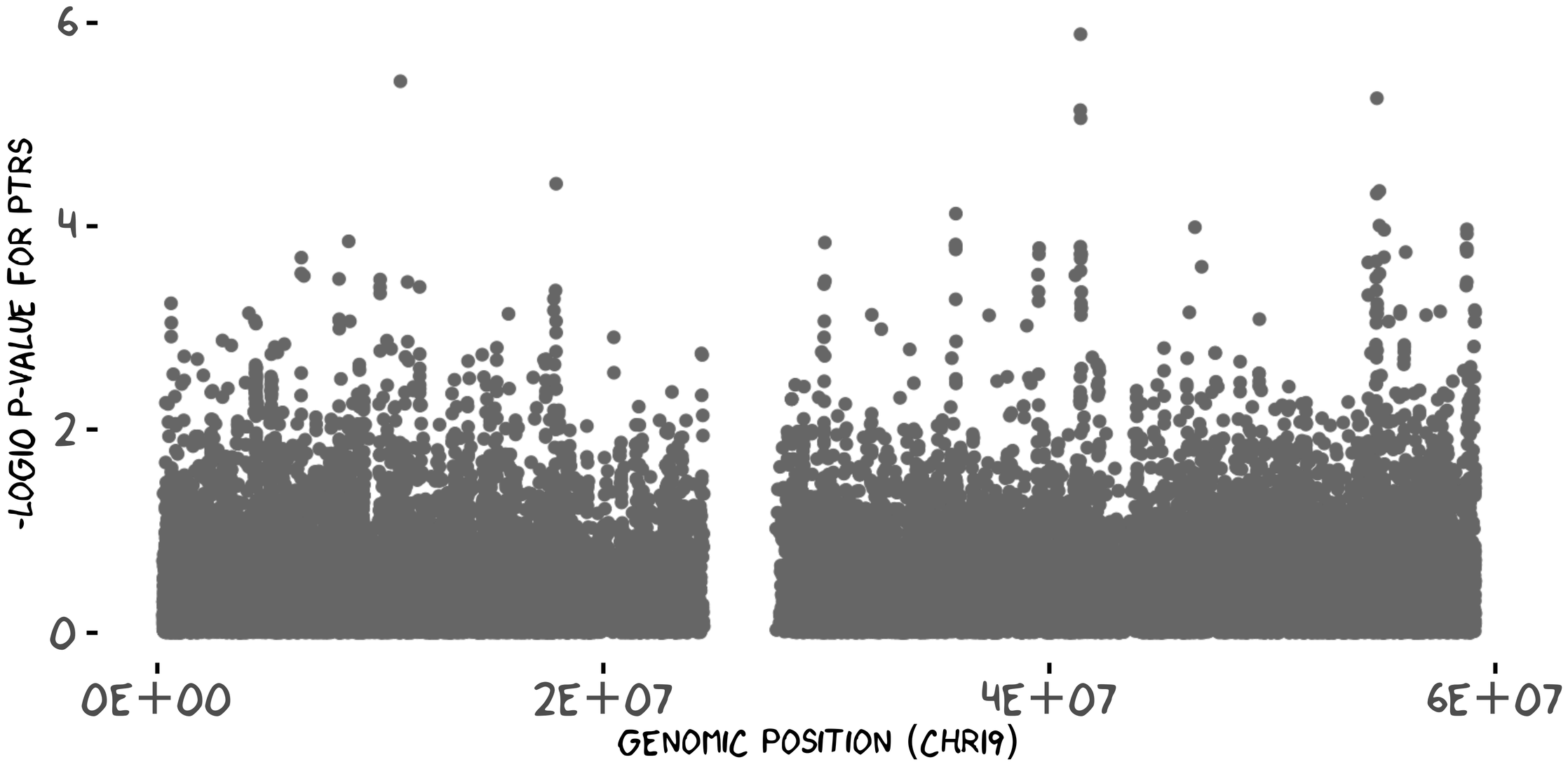
In previous longitudinal studies, we found some relationship.



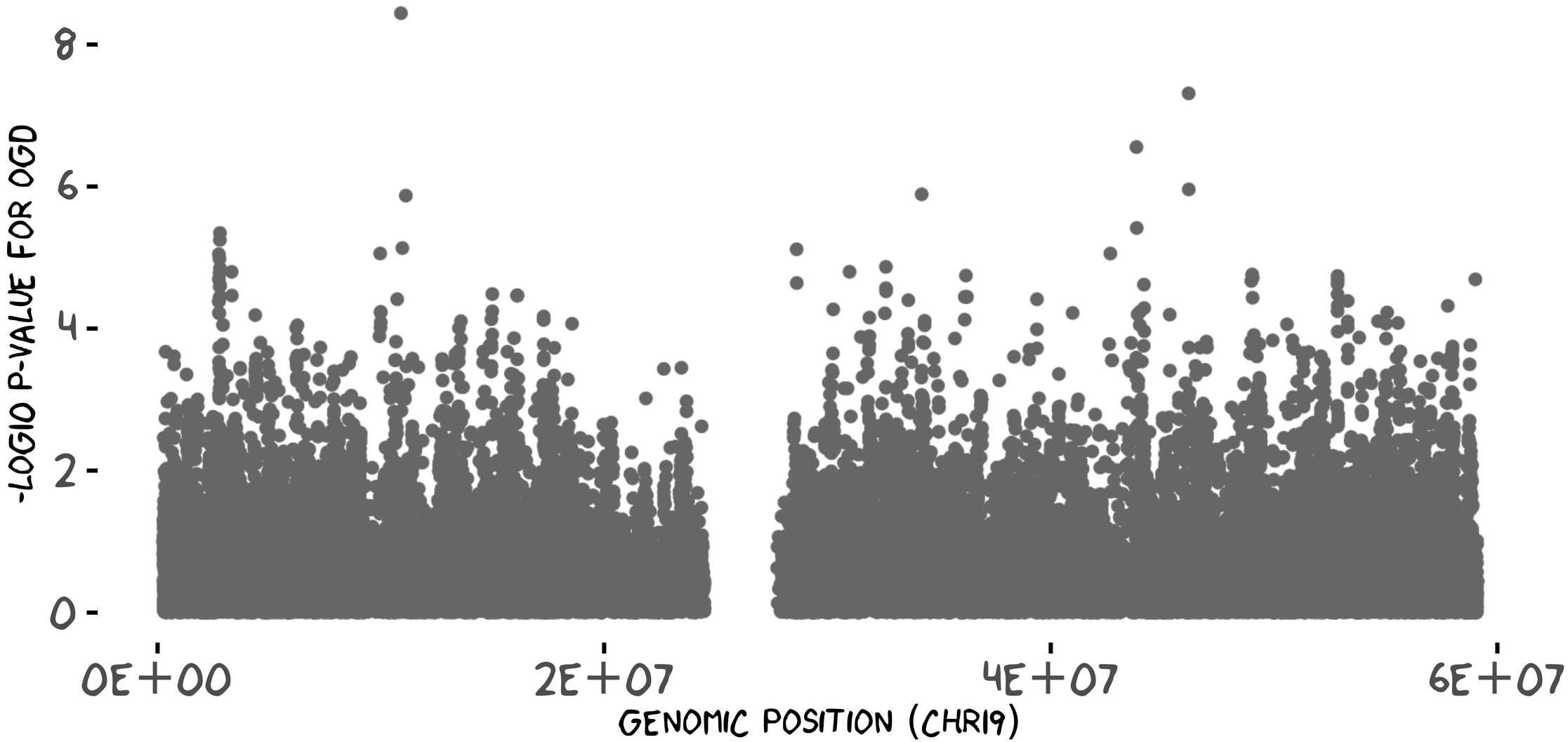




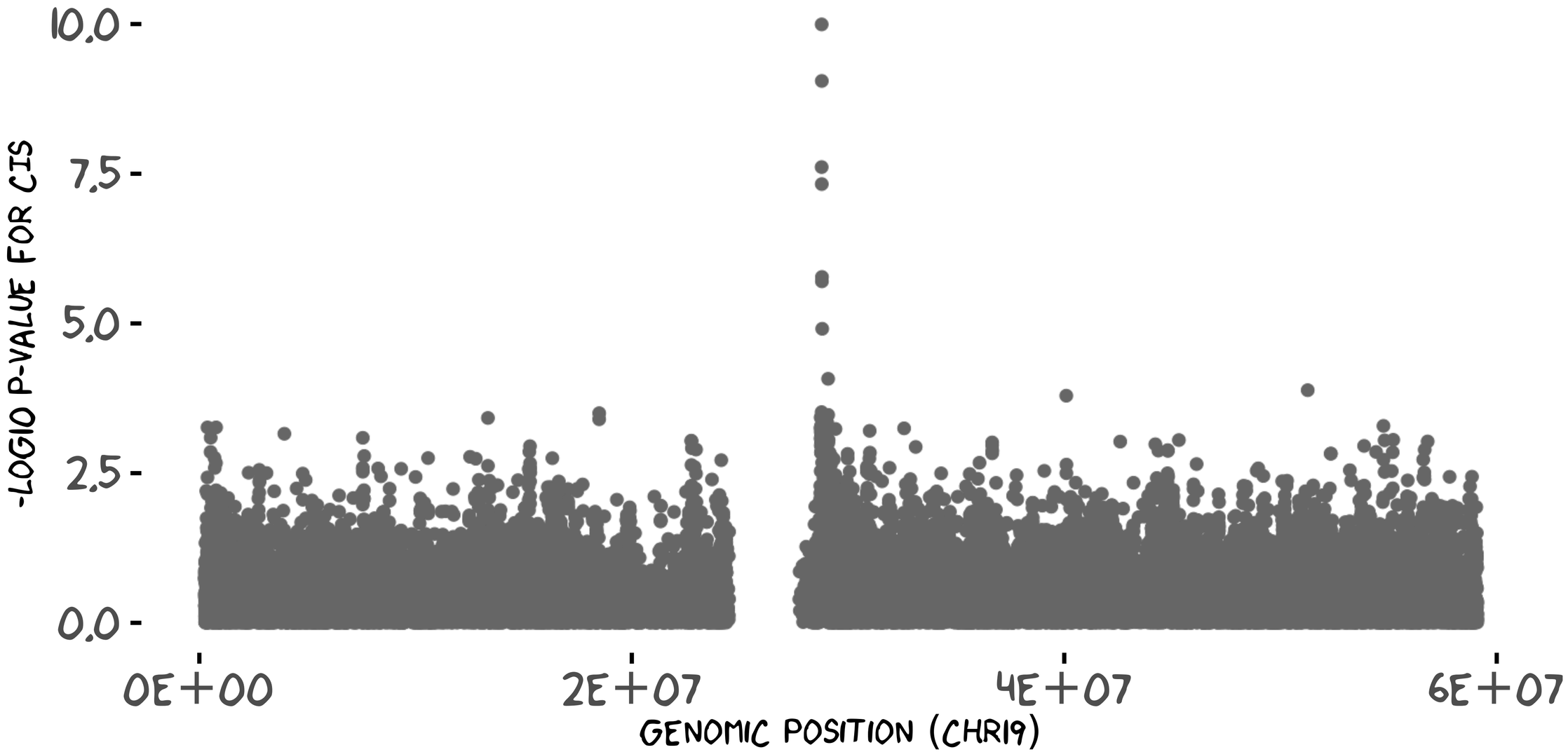
Can we recover such relationships from GWAS data?



Can we recover such relationships from GWAS data?



Can we recover such relationships from GWAS data?



MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Suppose we have estimated

$$\begin{aligned} G &\overset{\alpha}{\rightarrow} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Suppose we have estimated

$$\begin{aligned} G &\overset{\alpha}{\rightarrow} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

and

$$G \overset{\gamma}{\rightarrow} Y$$

MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Goal: What is the causal effect β in $X \overset{\beta}{\to} Y$?

Suppose we have estimated

$$Y = X\beta + \epsilon'$$

$$\begin{aligned} G &\overset{\alpha}{\to} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

and

$$G \overset{\gamma}{\to} Y$$

MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Suppose we have estimated

$$\begin{aligned} G &\overset{\alpha}{\rightarrow} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

and

$$G \overset{\gamma}{\rightarrow} Y$$

Goal: What is the causal effect β in $X \overset{\beta}{\rightarrow} Y$?

$$\begin{aligned} Y &= X\beta + \epsilon' \\ G\hat{\gamma} + \epsilon_Y &= (G\hat{\alpha} + \epsilon_X)\beta + \epsilon' \end{aligned}$$

MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Suppose we have estimated

$$\begin{aligned} G &\xrightarrow{\alpha} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

and

$$G \xrightarrow{\gamma} Y$$

Goal: What is the causal effect β in $X \xrightarrow{\beta} Y$?

$$\begin{aligned} Y &= X\beta + \epsilon' \\ G\hat{\gamma} + \epsilon_Y &= (G\hat{\alpha} + \epsilon_X)\beta + \epsilon' \\ G\hat{\gamma} &= G\hat{\alpha}\beta + \cdots \end{aligned}$$

MR measures mediation effects of X to Y

Given that G is a valid instrumental variable...

Example:

- X: gene expression
- Y: disease phenotype

Suppose we have estimated

$$\begin{aligned} G &\xrightarrow{\alpha} X \\ X &= G\hat{\alpha} + \epsilon_X \end{aligned}$$

and

$$G \xrightarrow{\gamma} Y$$

Goal: What is the causal effect β in $X \xrightarrow{\beta} Y$?

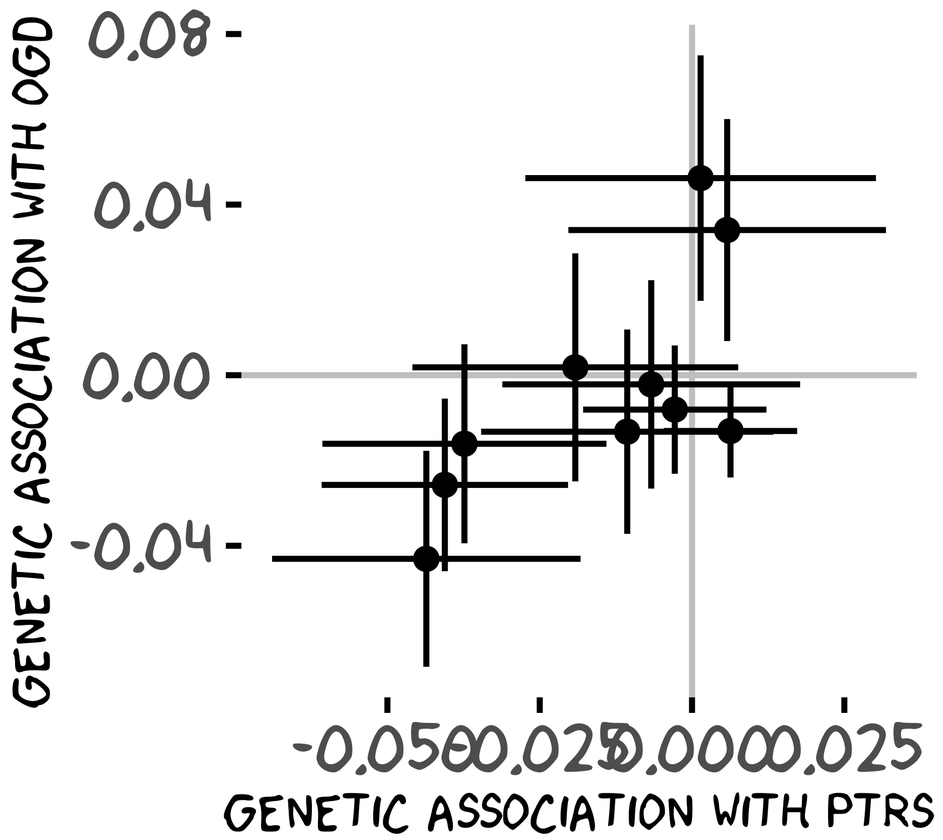
$$\begin{aligned} Y &= X\beta + \epsilon' \\ G\hat{\gamma} + \epsilon_Y &= (G\hat{\alpha} + \epsilon_X)\beta + \epsilon' \\ G\hat{\gamma} &= G\hat{\alpha}\beta + \cdots \end{aligned}$$

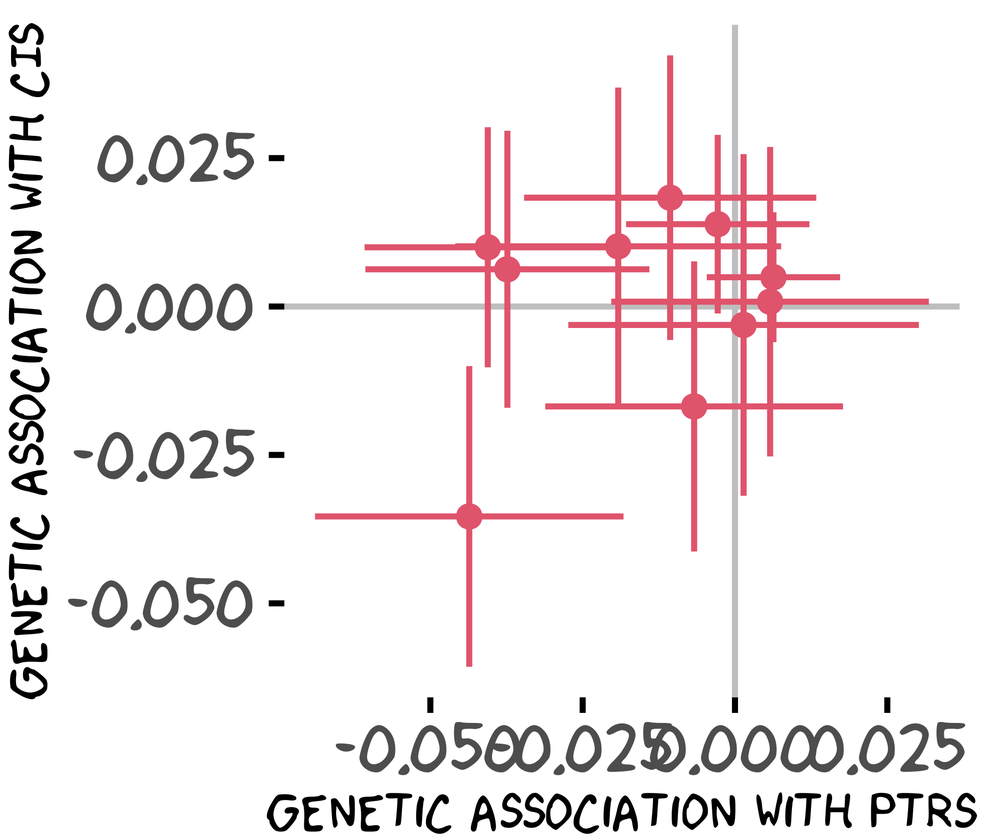
The answer is as simple as

$$\mathbb{E}[\beta] = \frac{\gamma}{\alpha}$$

Will any genetic variable work for MR analysis?

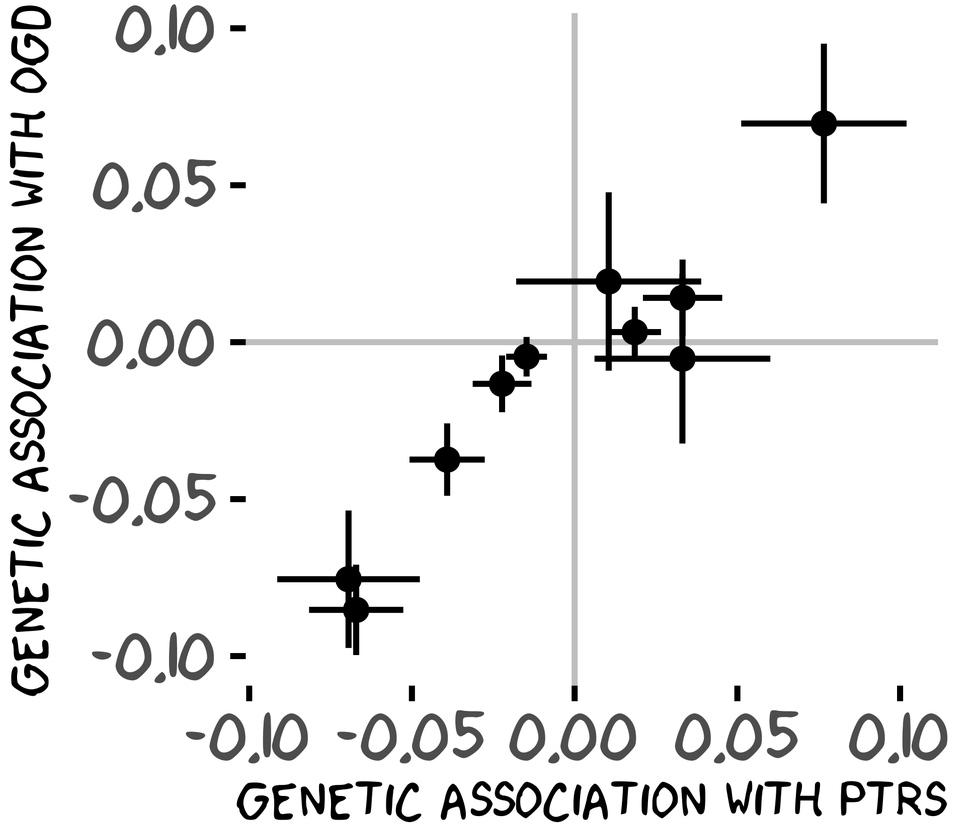
Goal: Is PTRS $\longrightarrow$ OGD or PTRS $\longrightarrow$ CIS?

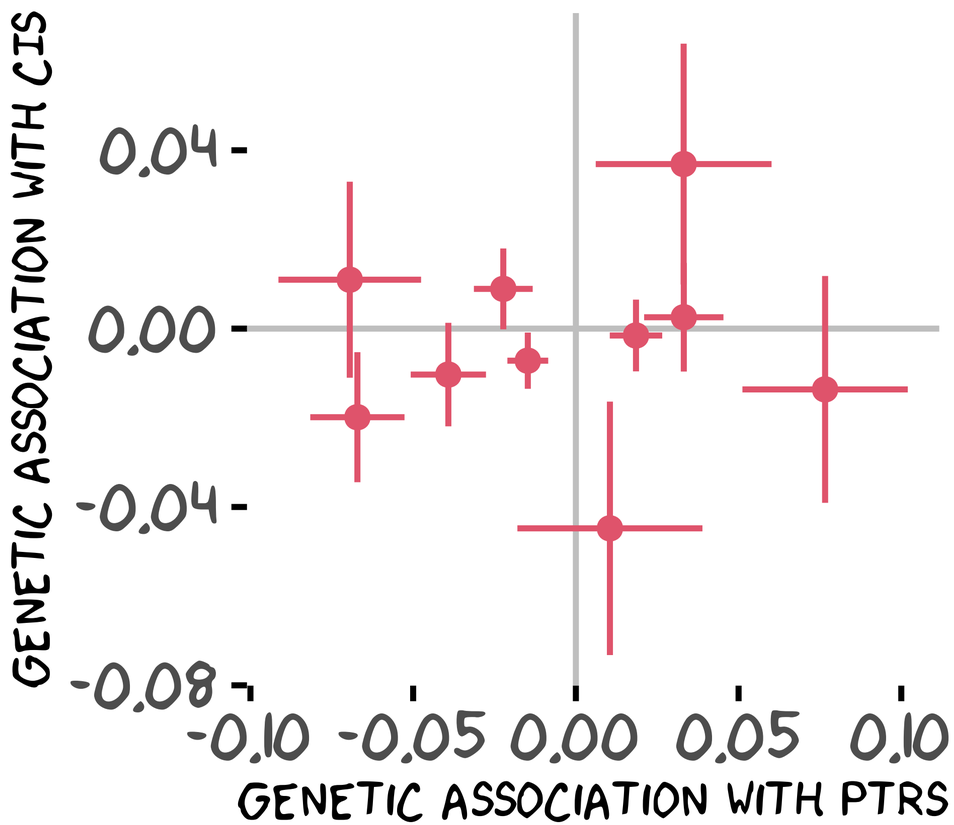




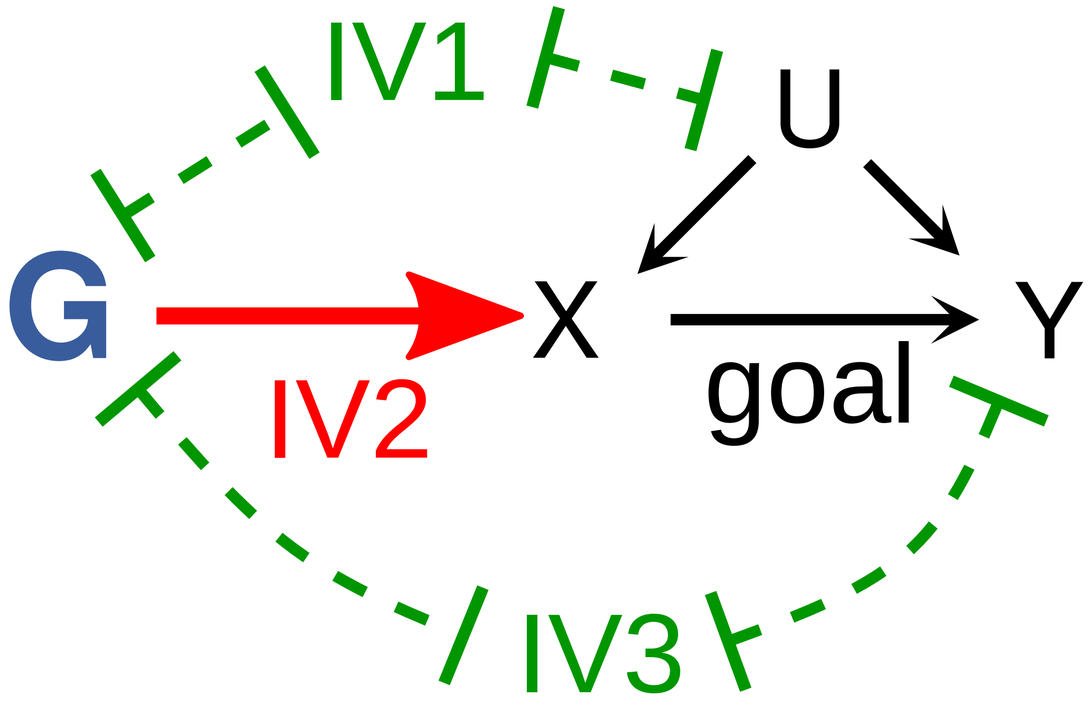
How about selecting causal variants of exposure (PTRS)?

Goal: Is PTRS $\longrightarrow$ OGD or PTRS $\longrightarrow$ CIS?





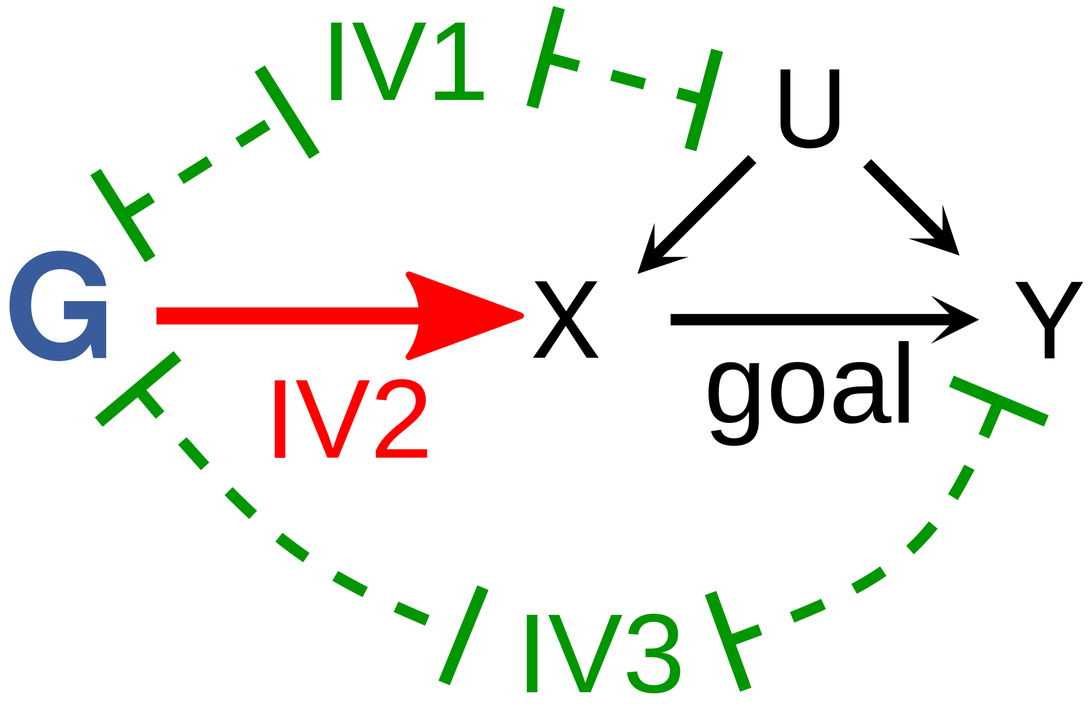
Definition of instrumental variable in MR study



- IV1: The genetic variant G is independent of the potential confounder variable U

Bowden *et al.* (2015)

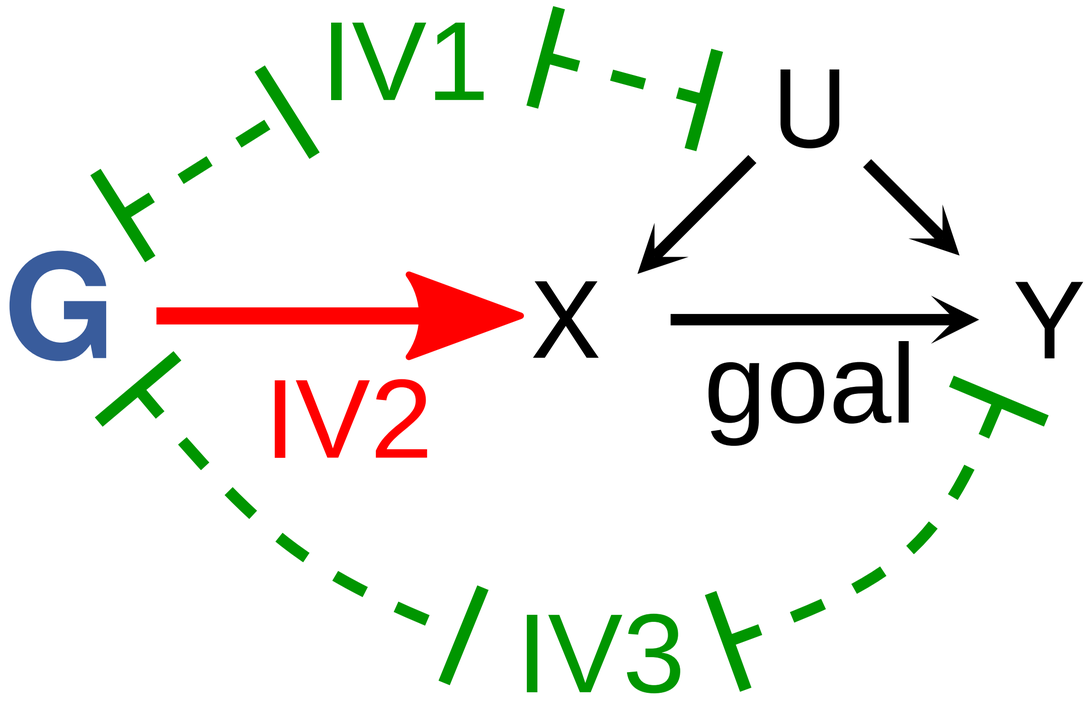
Definition of instrumental variable in MR study



- IV1: The genetic variant G is independent of the potential confounder variable U
- IV2: The genetic variant is associated with the exposure X

Bowden *et al.* (2015)

Definition of instrumental variable in MR study



- IV1: The genetic variant G is independent of the potential confounder variable U
- IV2: The genetic variant is associated with the exposure X
- IV3: The genetic variant is independent of the outcome Y conditioning on X

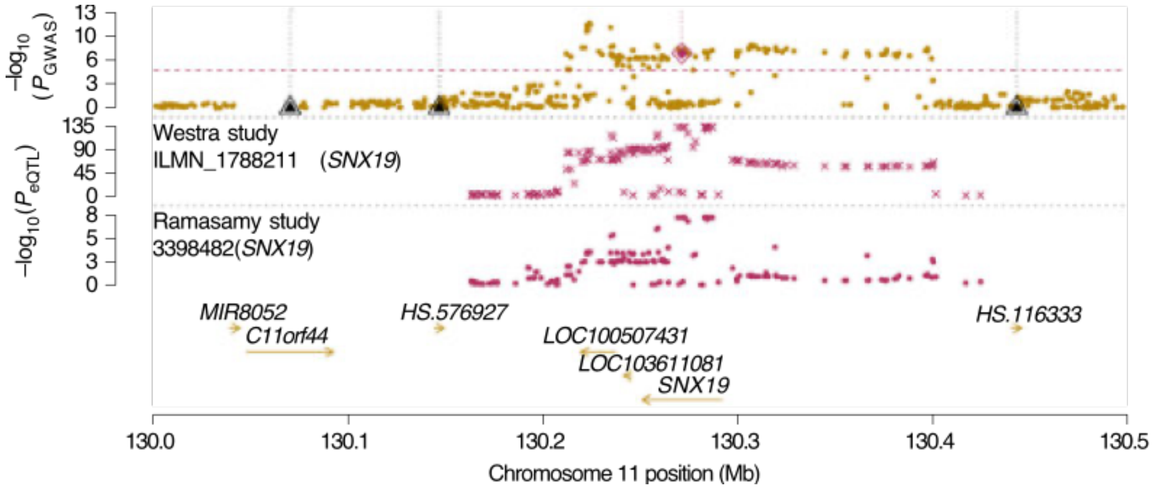
Bowden *et al.* (2015)

Mendelian Randomization (X = a gene)

Input:

$$G \xrightarrow{\gamma} Y$$

$$G \xrightarrow{\alpha} X$$



Goal:

$$G \xrightarrow{\alpha} X \xrightarrow{\beta} Y$$

$$\hat{\beta} = \gamma / \alpha$$

Zhu *et al.* Nature Genetics (2016)

Mendelian Randomization (X = a gene)

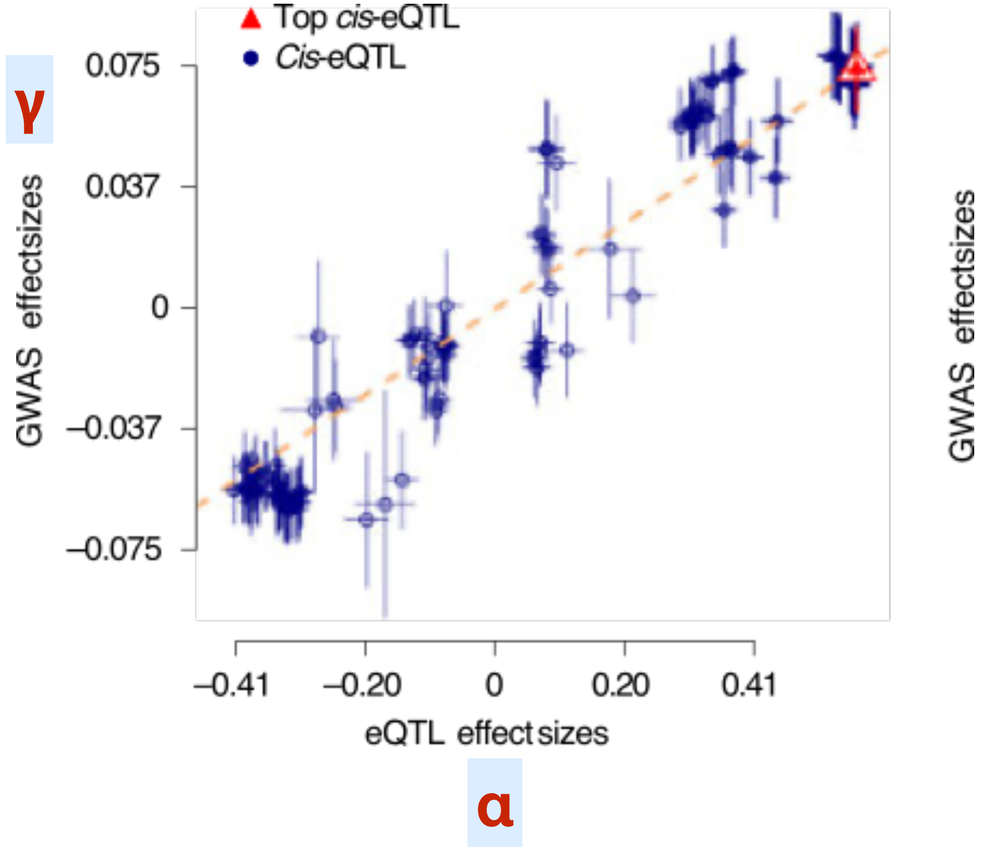
Input: $G \xrightarrow{\gamma} Y$

$G \xrightarrow{\alpha} X$

Goal:

$G \xrightarrow{\alpha} X \xrightarrow{\beta} Y$

$\hat{\beta} = \gamma / \alpha$



Zhu *et al.* Nature Genetics (2016)

Mediation analysis by two-stage regression

Goal: β

$$G \xrightarrow{\alpha} X \xrightarrow{\beta} Y$$

$$G \xrightarrow{\gamma} Y$$

- Step 1: regression Y ~ G
- Step 2: regression X ~ G
- Step 3: regression Y ~ X + G
- Step 4: check if X explained away G on Y
  (no correlation between residuals and X)

Baron & Kenny (1986)

Condition:

- Very strong evidence for $|\gamma| > 0$
- G is a **valid instrumental variable**

Mediation effect (Sobel's test):

$\alpha\beta$ or β ($\approx \gamma/\alpha$)

$$\hat{\text{StdErr}}(\alpha\beta) = \sqrt{\sigma_\alpha^2\beta^2 + \sigma_\beta^2\alpha^2}$$

Sobel (1982)

Today’s lecture

- Summary-based GWAS
- Causal inference methods
- Mendelian Randomization