Selecting Genetic Variants Using Knockoff Under Collider Bias

Qian Zhao

Postdoctoral Scholar, Biomedical Data Science

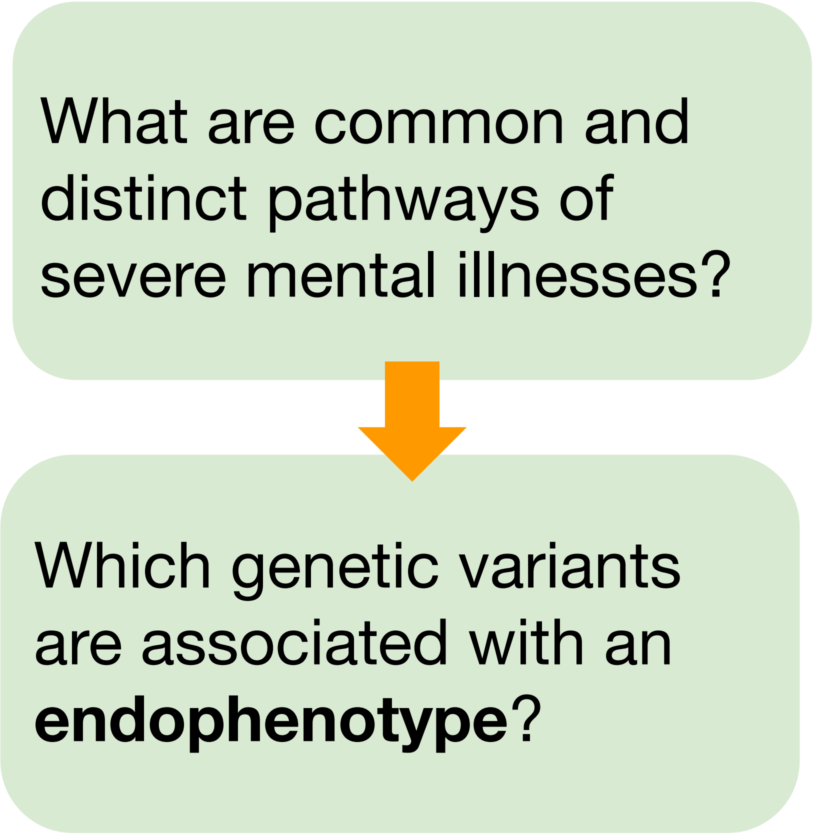
Stanford | Biomedical Data Science

SCHOOL OF MEDICINE

A motivating question

- Severe mental illnesses overlap in symptoms and share some genetic risks

- Leveraging genetic variants can identify biological pathways and thus help psychiatrists define subtypes in a more biologically sound way

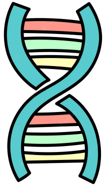


What are common and distinct pathways of severe mental illnesses?

Which genetic variants are associated with an **endophenotype**?

A motivating question

Genotype



Symptoms
& Diagnosis

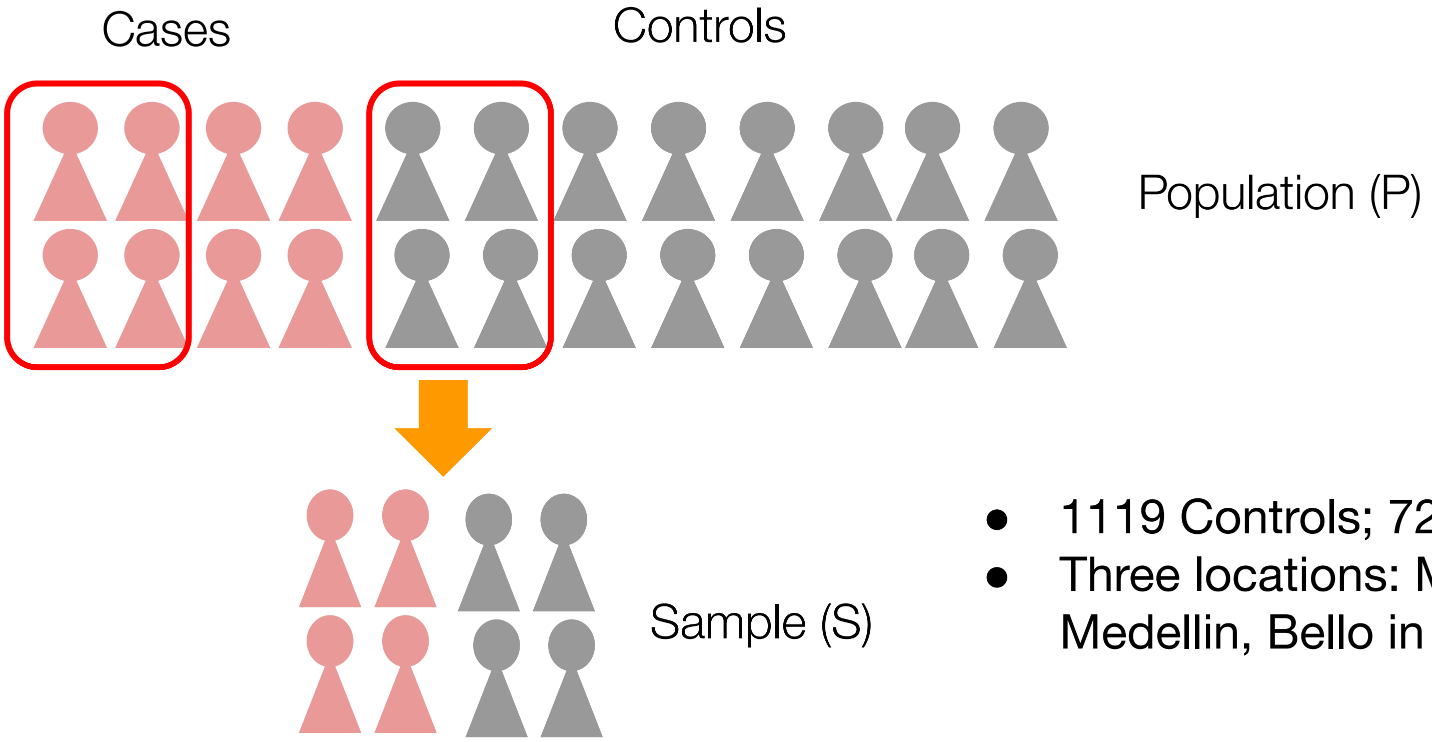
Neurocognitive
function

- Demographics
- Medication
- Substance use
- …

What are common and distinct pathways of severe mental illnesses?

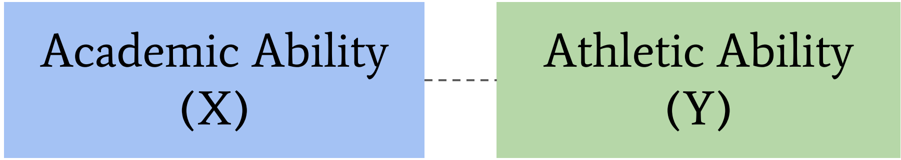
Which genetic variants are associated with an **endophenotype**?

Case-control study design

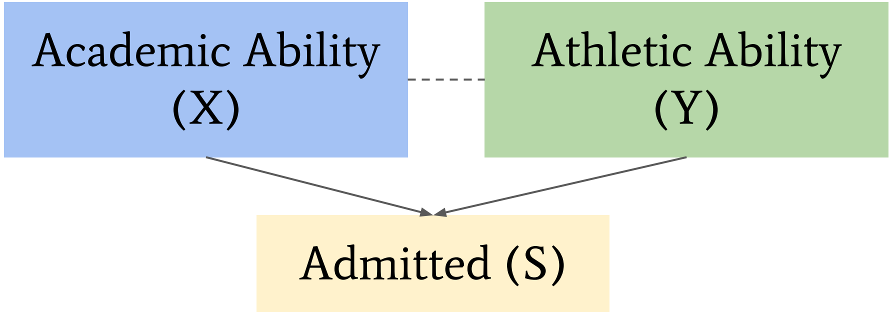


- 1119 Controls; 7246 Cases
- Three locations: Manizales, Medellin, Bello in Columbia

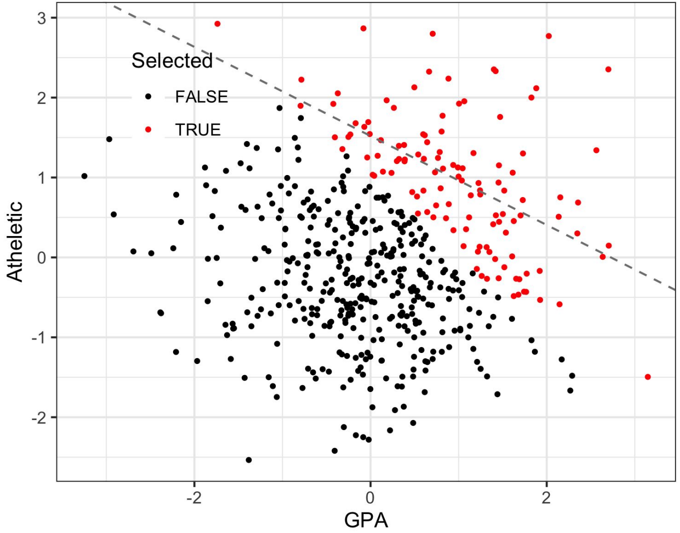
Example: Is academic ability associated with athletic ability?



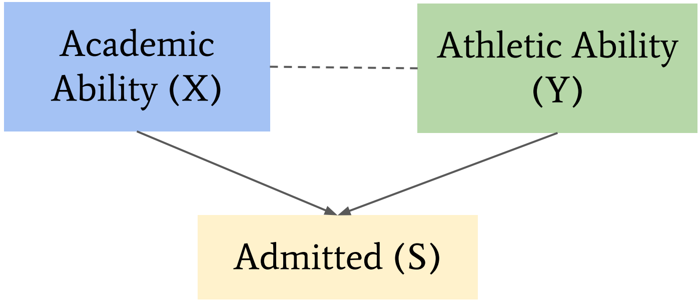
Population P(X,Y)
(Red + Black points)



Sample L(X, Y) = P(X, Y|S=1)
(red points)



Collider Bias



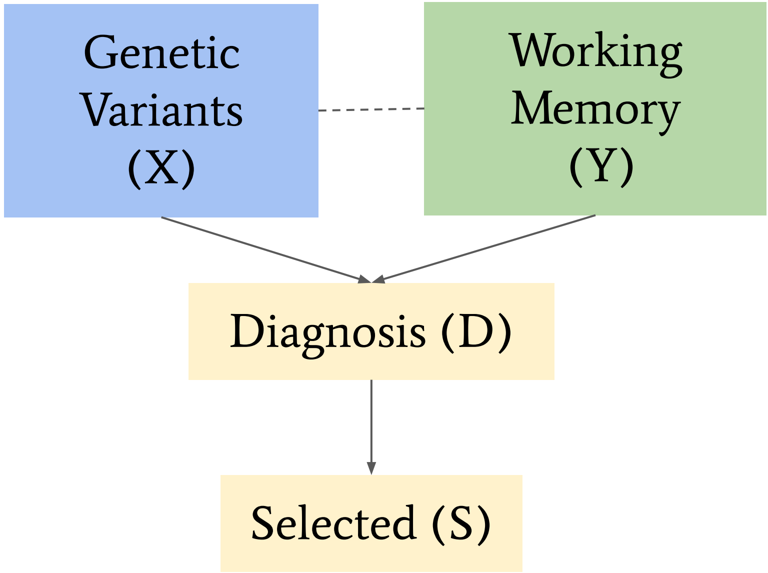
Academic Ability (X)

Athletic Ability (Y)

Admitted (S)

Can we test if X affects Y in the population (P) using case case-control sample (S)?

Case-control study



Genetic Variants (X)

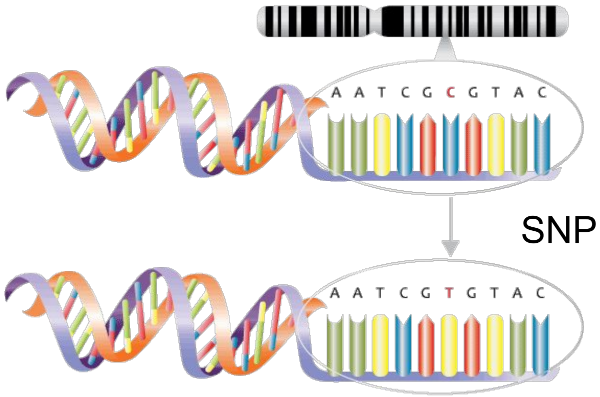
Working Memory (Y)

Diagnosis (D)

Selected (S)

Current approaches usually do not apply when the number of variables is large!

A knockoff approach to select variables in the high-dimensional setting



$$(X_1, X_2, \ldots, X_{10})$$

Figure source

- Single nucleotide polymorphisms (SNP) are alterations in a single nucleotide (A, T, G, C) in the DNA sequence.
- Denote the SNPs as

$$(X_1, X_2, \ldots, X_p)$$

- Denote the endophenotype as Y.
- Identify important SNPs by testing

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}$$

The knockoff method

- The knockoff method is a variable selection procedure that controls the false discovery rate
- For every observation $(X_1, X_2, \ldots, X_p)$
  We construct **knockoff variables** $(\tilde{X}_1, \ldots, \tilde{X}_p)$
  Such that

$$(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p)_{\text{Swap S}} \overset{d}{=} (X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p)$$

For every set S containing only **null** variables

The knockoff method

	X_1	X_2	X_3	X_4	X_5
Sample 1	A	T	G	G	C
Sample 2	A	T	A	G	C
Sample 3	T	T	G	G	A
Sample 4	A	T	A	T	C

$\tilde{X}_1$	$\tilde{X}_2$	$\tilde{X}_3$	$\tilde{X}_4$	$\tilde{X}_5$
T	T	A	G	C
A	T	G	G	A
T	T	A	G	A
A	C	A	T	C

We should not be able to tell between a variable and its knockoff!

Feature importance statistics

$$(Z_1, Z_2, \ldots, Z_5, \tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_5)$$

E.g. LASSO regression coefficients

Scores

$$(W_1, W_2, W_3, W_4, W_5)$$

$$W_j = f(Z_j, \tilde{Z}_j) = -f(\tilde{Z}_j, Z_j)$$

A list of selected variables

How do we construct knockoff variables?

$$(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) \overset{d}{=} (\tilde{X}_j, X_{-j}, X_j, \tilde{X}_{-j}, Y)$$

$$\mathcal{L}(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) = \mathrm{P}(X_j, X_{-j}, Y | S = 1) \boxed{Q(\tilde{X}_j, \tilde{X}_{-j} | X_j, X_{-j}, Y)}$$

User defined!

How do we construct knockoff variables?

$$(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) \overset{d}{=} (\tilde{X}_j, X_{-j}, X_j, \tilde{X}_{-j}, Y)$$

$$\mathcal{L}(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) = \boxed{\mathrm{P}(X_j, X_{-j}, Y | S = 1)} Q(\tilde{X}_j, \tilde{X}_{-j} | X_j, X_{-j}, Y)$$

We don’t know what is
P(X, Y)!

How do we construct knockoff variables?

$$(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) \overset{d}{=} (\tilde{X}_j, X_{-j}, X_j, \tilde{X}_{-j}, Y)$$

$$\mathcal{L}(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}, Y) = \mathrm{P}(X_j, X_{-j}, Y | S = 1) Q(\tilde{X}_j, \tilde{X}_{-j} | X_j, X_{-j}, Y)$$

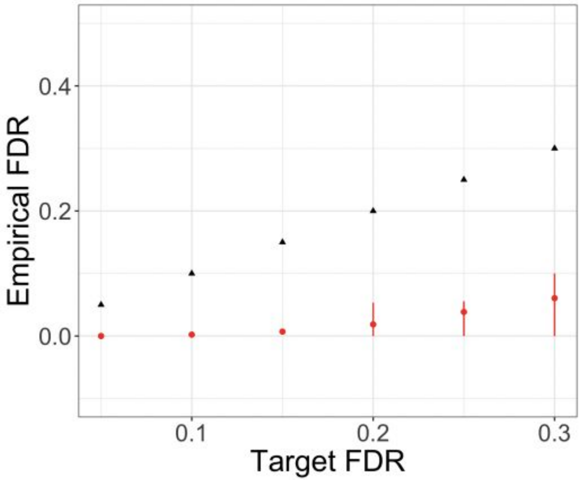
$$\mathrm{P}(X_j, X_{-j}) \, \mathrm{P}(S = 1 | X_j, X_{-j}, Y)$$

We can use this distribution instead!

(Selection probability)

$$\propto \mathrm{P}(X_j | X_{-j}, Y, S = 1)$$

How do we construct knockoff variables?



If

$$\mathrm{P}(X_j, X_{-j})\,\mathrm{P}(S = 1|X_j, X_{-j}, Y)$$

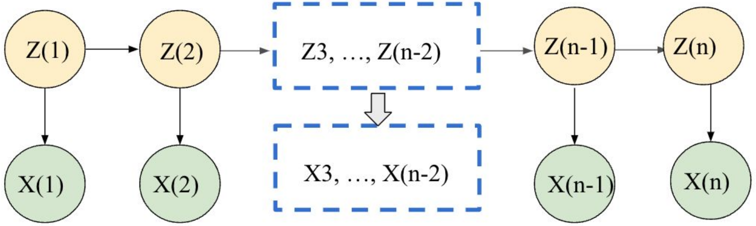
Is multivariate Gaussian, then we can sample knockoffs from another multivariate Gaussian.

Number of cases = Number of controls = 1,000; Number of variables = 200; X is from a multivariate Gaussian distribution where the covariance matrix is block diagonal (block size = 10); Y is from a linear model where 10% of the variables are non-nulls;

$\mathrm{P}(D = 1|X, Y) = e^{-v^2/2}$, $v = \gamma_0 Y + X^\top \gamma$.

Conclusion and remaining challenges

- Collider bias can occur in many studies!
- We describe one way to adjust the knockoff sampling procedure to select variables that controls the false discovery rate (1) in high-dimensions (2) under collider bias
- It’s unclear how to define $Q(\tilde{X}_j, \tilde{X}_{-j} | X_j, X_{-j}, Y)$
- One idea: approximate $\mathrm{P}(X_j, X_{-j})\,\mathrm{P}(S = 1 | X_j, X_{-j}, Y)$ by another distribution for which we know Q



Thank you! Questions?