

(Epi-)genetic Risk Scores

Andrea Allegrini & Alexander Neumann

Paradise Meeting 2023.05.23

Outline

- 1. Fundamentals of Psychiatric Genetics
- 2. Polygenic Risk Scores
- 3. Methylation Risk scores

Nature or Nurture

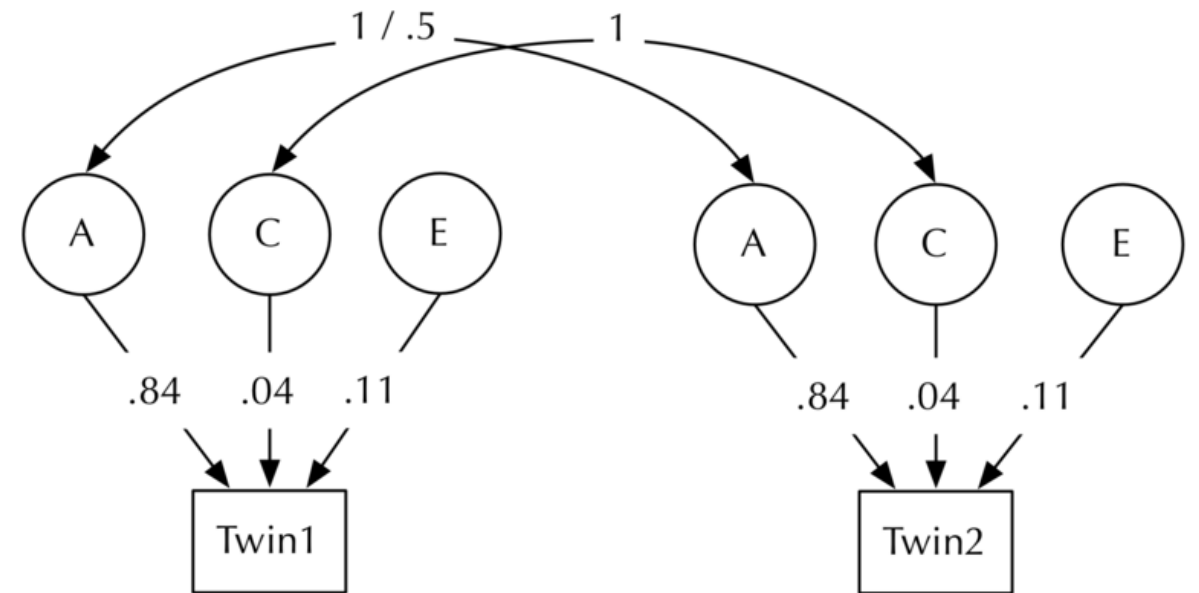
- Arguably the most general etiological question is to which degree a disorder is determined by genetics (heritability) or environment

Question

- How heritable are psychiatric disorders on average?
 - ▶ I.e. how much can the variability in disorder occurrence be explained by genetic factors?
- A: 10%
- B: 20%
- C: 30%
- D :40%
- E: 50%
- F: 70%
- G: 80%
- H: 90%

Twin studies: ACE Model


- Idea:
 - ▶ Compare whether monozygotic twin pairs (~100% identical genetics variants) are more similar than dizygotic twin pairs (on avg. ~ 50% identical genetic variants)
 - The more similar monozygotic twins pairs, compared to similarity within dizygotic twin pairs, the more heritable a trait
 - Latent variable modeling allows estimation of A (Additive genetics), C (shared environment) and (E) non-shared Environment



https://commons.wikimedia.org/wiki/File:Twin_Study_Structural_ACE_model_STD.png

MaTCH database


- Meta-analysis of all twin studies published until 2012
- ICD-10 classification
- <https://match.ctglab.nl/>





MaTCH Meta-Analysis of Twin Correlations and Heritability


This website provides a resource for the heritability of all human traits that have been investigated with the classical twin design. The traits have been classified into 28 broad trait domains, as well as according to the standard classification schemes of the International Classification of Functioning, Disability and Health (ICF) or the International Classification of Diseases and Related Health Problems (ICD-10). Currently the database includes information from 2748 papers, published between 1958 and 2012, reporting on 17804 traits on a total of 14,558,903 twin pairs. Have Fun!

Please refer to the original paper: Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D. Meta-Analysis of the Heritability of Human Traits based on Fifty Years of Twin Studies. *Nature Genetics*, 2015 Jul;47(7):702-9 doi:10.1038/ng.3285, [published online May 18, 2015](#)

 **Analysis**
Specific Traits

 **Analysis**
Multiple Traits

 **Overview**
[What's in here](#)

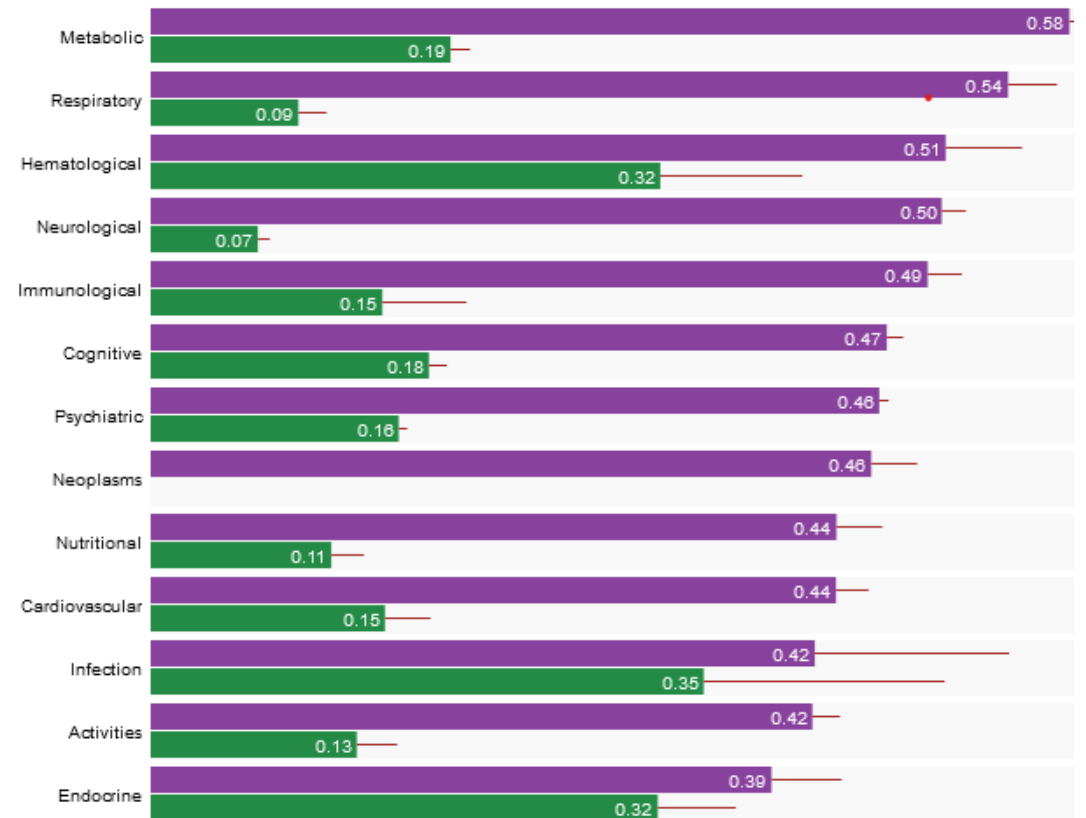
 **About**
How it works

MaTCH demonstration

- If you want to follow along, go to <https://match.ctglab.nl/>

Heritability of Psychiatric Disorders

- Psychiatric problems overall have a twin heritability of 46% [95% CI: 45-47%], with shared environment contributing 16% [95% CI: 15-17%]



Heritability of Psychiatric Disorders

- Less common ($\approx 1\%$ prevalence) psychiatric disorders tend to have the highest heritability
 - ▶ Schizophrenia: 79% [95% CI: 65-99%] (Hilker et al., 2018)
 - ▶ Bipolar disorder: 68% [95% CI: 64-72%]
 - ▶ Pervasive developmental disorder: 60% [95% CI: 54-66%]
- Exception: ADHD, common (5-7.5% prevalence), but highly heritable
 - ▶ ADHD: 72-88% (Larsson et al., 2014)
- Internalizing disorders tend to have lower estimates:
 - ▶ Depressive episode: 34% [95% CI: 31-37%]
 - ▶ Other anxiety disorder: 40% [95% CI: 37-43%]

Heritability Misconceptions

- ▶ ACE estimates importance of genome across populations, not individuals.
 - A 46% heritability does not imply that the cause is genetic for 46% of people and environmental for 54%
 - For some individuals causes will be more genetic or environmental, but on average we can expect the influence of both
- ▶ High heritability does not imply fate
 - Current or future prevention/therapy could help even for highly heritable disorders
 - Individuals with high genetic risk may never show symptoms
 - Example:
 - Environmental influences in ADHD likely not more than 30%
 - This implies that factors like parenting are not as important as parent's genes in population
 - But for certain individuals, exceptional parenting can make all the difference, even if parenting differences in the population at large may comparatively be less important

Heritability Misconceptions

- ▶ E is non-shared environment, but is not only representing stable, long-lasting environmental factors (e.g. traumatic events)
 - Should be more thought as any variance not attributable to A or C
 - Stochastic processes/chance events and measurement error are also included
- ▶ Environment can be heritable, too (Gene-environment correlation)
- ▶ A is genetic, but not necessarily stable or unchangeable
 - A and C estimates are not absolute values, but proportions
 - So if E becomes more important, A and C could also change, even if absolute genetic effects stay the same
 - Genetic effects can be also differently expressed depending on age
 - Next slide: example of changing heritability due to age

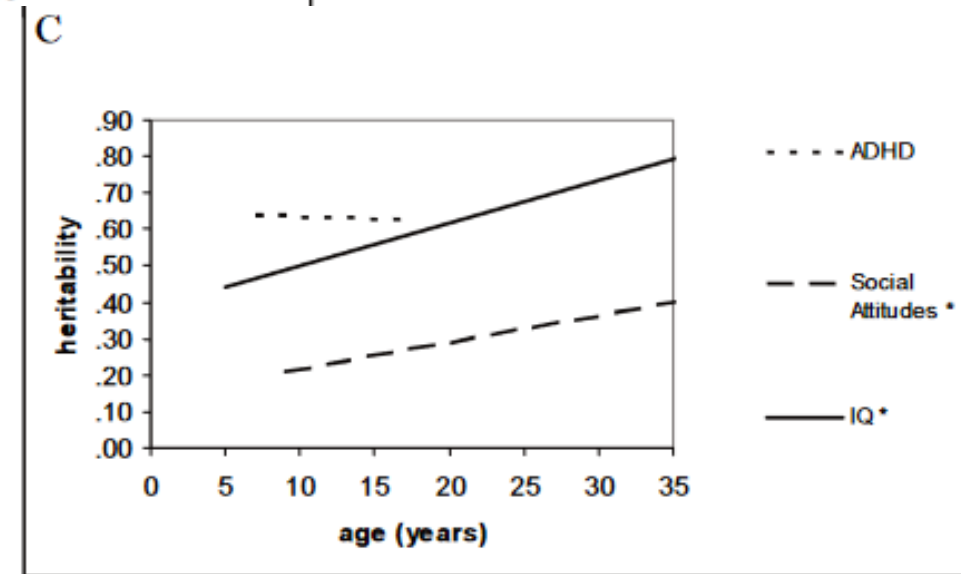
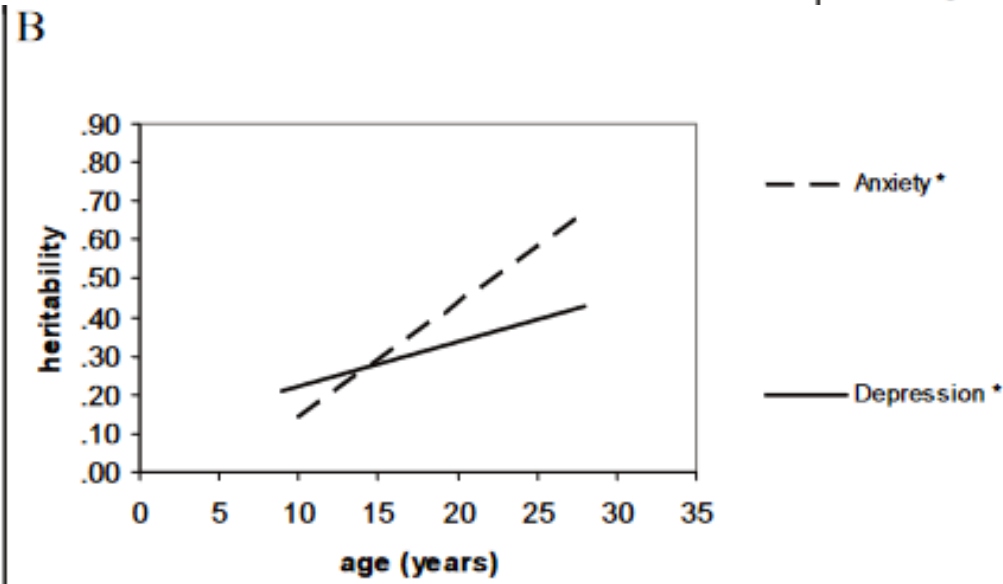
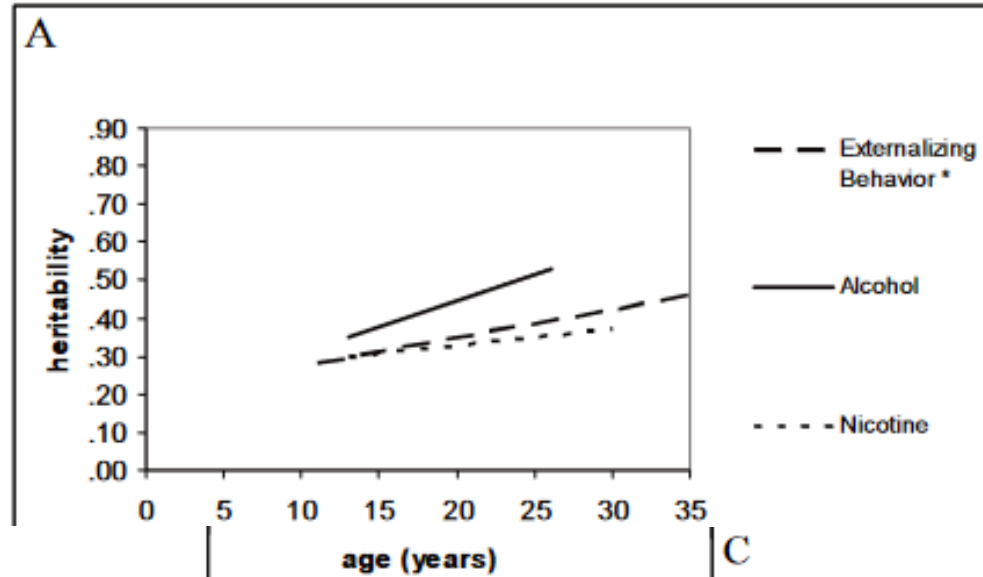
Poll

- How does the heritability of psychiatric disorders change with age?
 - ▶ Increase
 - ▶ Decrease
 - ▶ Depends on disorder

Heritability changes across lifespan

Typically 1-3% increase per year from childhood to adulthood

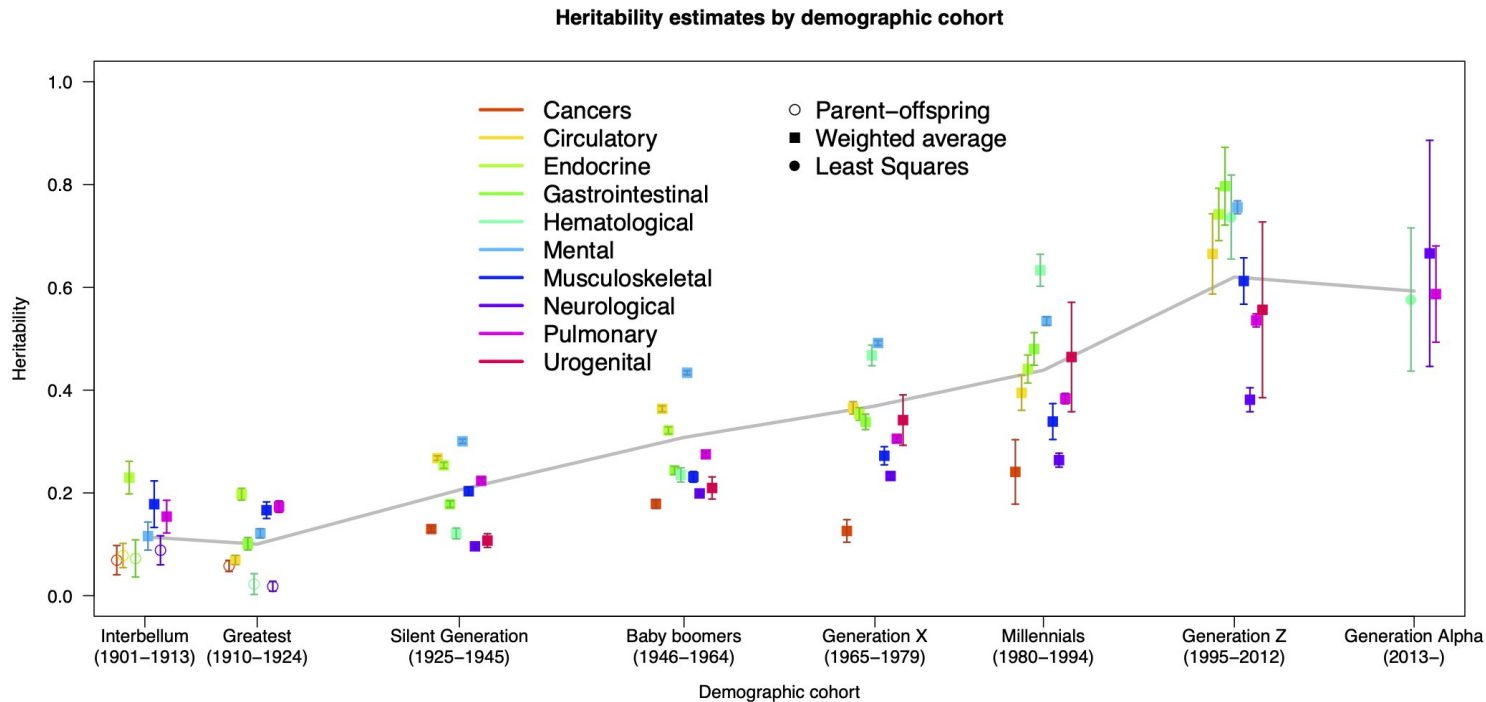
Bergen, Sarah E., Charles O. Gardner, and Kenneth S. Kendler. "Age-related changes in heritability of behavioral phenotypes over adolescence and young adulthood: a meta-analysis." *Twin Research and Human Genetics* (2007)



Possible Explanations

1. Active Gene-Environment correlation
 - Child is genetically predisposed towards hostility -> makes friends with other children, who are hostile -> these influence the child to show even more aggression over time
2. Accumulation of stable genetic effects
 - As opposed to potentially more inconsistent and changing environmental effects
3. Decrease in influence of shared-environment
 - E.g. decreasing influence of parenting in adolescence compare to childhood
4. Measurement error reductions
 - Children become better at introspection and are able to express problems better, especially internalizing problems
 - Reduction of E would then lead to increase in A

Heritability changes across generations



- Danish registry data
- Higher h^2 with newer generations
- Early vs late onset of disorders?
- Accumulation of environmental factors?

- Athanasiadis, Georgios, et al. "A comprehensive map of genetic relationships among diagnostic categories based on 48.6 million relative pairs from the Danish genealogy." *Proceedings of the National Academy of Sciences* 119.6 (2022): e2118688119.

Genetic Architecture of Psychiatric Disorders

- Twin studies give important insights into the contribution of genetic effects in a broad sense
- However, twin heritability does not provide insights into what kind of genetic effects contribute to this heritability:
 - ▶ Common or rare variants?
 - ▶ Which genetic pathways?
 - ▶ Single nucleotide polymorphism (SNP) or structural variants?
- To gain these insights, we need molecular genetic studies, which have directly measured genotypes

SNP Heritability

- Basic idea: genotype a set of variants and estimate how much they jointly contribute to psychiatric problems
- Most frequent application: Estimate the variance explained of a psychiatric outcome by the joint contribution of common autosomal SNPs = SNP Heritability
 - ▶ Same set of variants, which is most often used for polygenic risk score calculations
 - ▶ Thus SNP h^2 provides upper limit of how much a PRS could predict an outcome
- However, other sets also possible
 - ▶ E.g. What is the contribution of SNPs, which are involved in immune function?
- Popular methods: GREML and LD score regression

SNP H2 database resources

- GWAS Atlas (<https://atlas.ctglab.nl/>)
- Look-up of SNP H2 in UKBB
https://nealelab.github.io/UKBB_Idsc/h2_browser.html
- Look-up of rG in UKBB
- https://ukbb-rg.hail.is/rg_browser/

Polygenic Risk Scores

- Heritability estimates provide population description on the contribution of genetics towards psychiatric disorders
- How do we estimate individual genetic risk?
- Presentation on PRS by Andrea

PRS challenges

- PRS are used very often in research settings
- However, genetics are (barely) used in current clinical settings
- Let us discuss some limitations preventing clinical adoption

Explanatory power

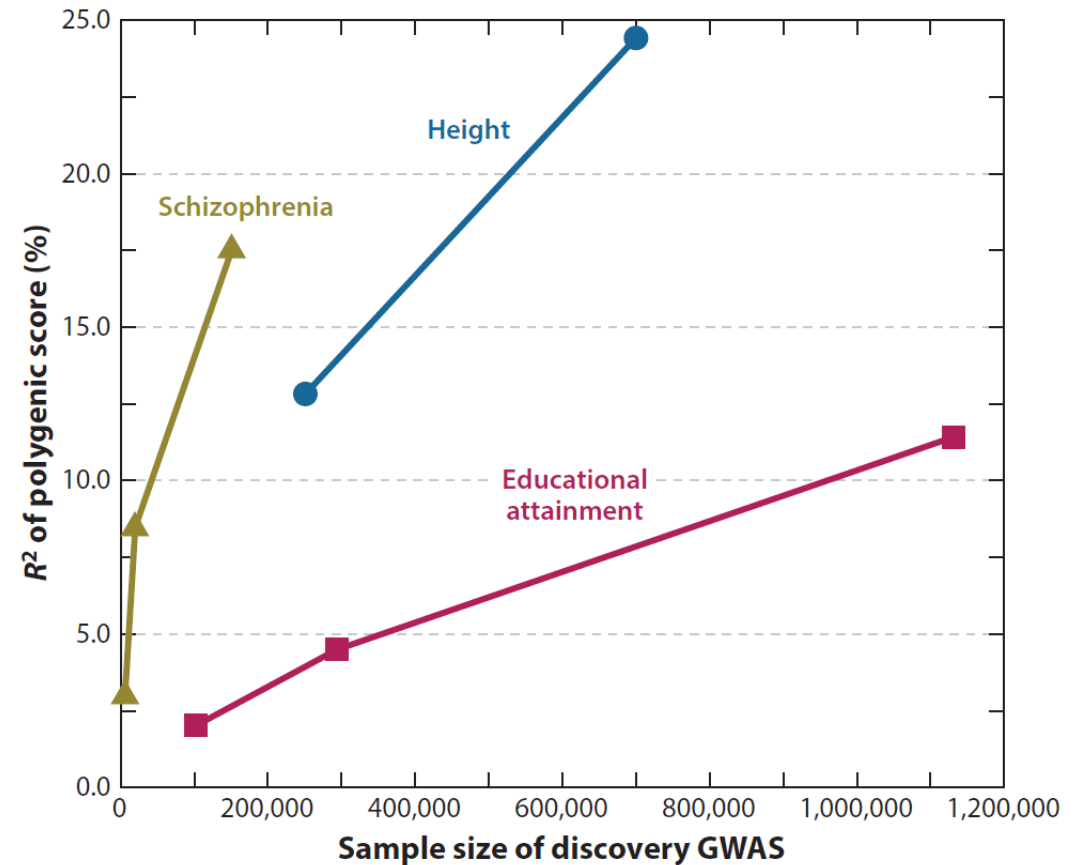
- Example: ADHD PRS score based on Demontis et al. (2019) GWAS
- Explains 4.0% of ADHD variance according to meta-analysis (Li et al, 2021)
- Not sufficient to reliably predict ADHD development in general population

Speculative current uses

- Fullerton & Nurnberger (2019) discuss some clinical uses, which perhaps could be implemented today/near future
- Identification of participants with extreme genetic predisposition:
 - Top 10% PRS score = 3x Schizophrenia risk, 2.5x major depression risk

PRS quality is dependent on discovery GWAS quality

- The quality of a PRS is based on estimates of the discovery GWAS
- Thus, the more precise the GWAS estimates, the better prediction of a PRS
- How do we achieve higher precision?
 - ▶ Higher Sample Size
 - ▶ Better outcome measures

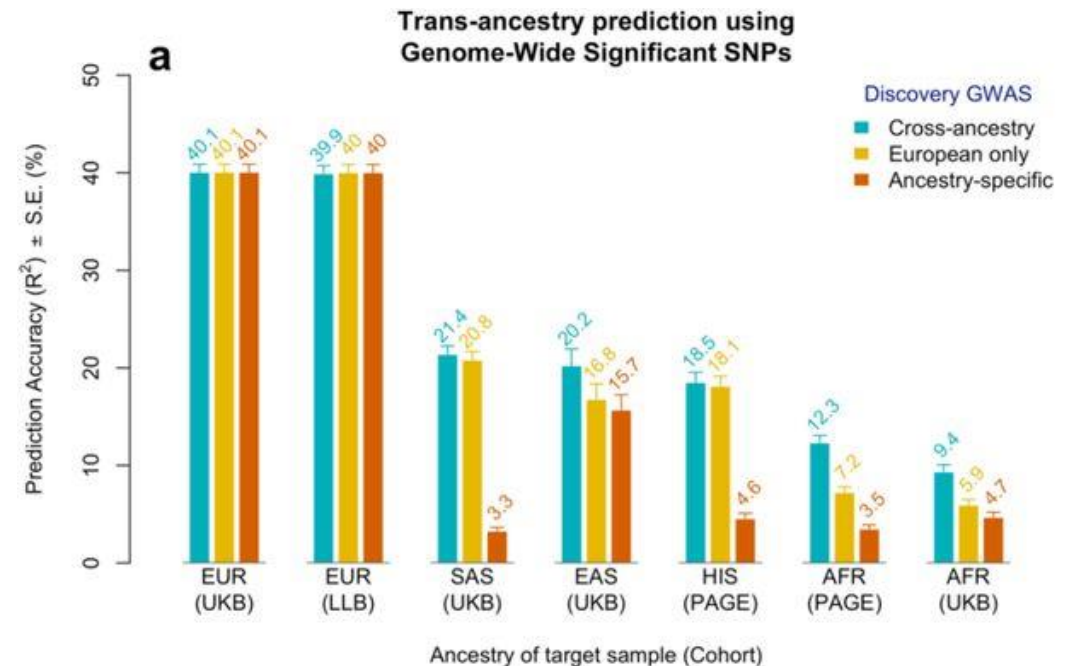


Raffington, Laurel, Travis Mallard, and K. Paige Harden. "Polygenic Scores in Developmental Psychology: Invite Genetics In, Leave Biodeterminism Behind." *Annual Review of Developmental Psychology* 2 (2020): 389-411.

Will we ever have enough sample size?

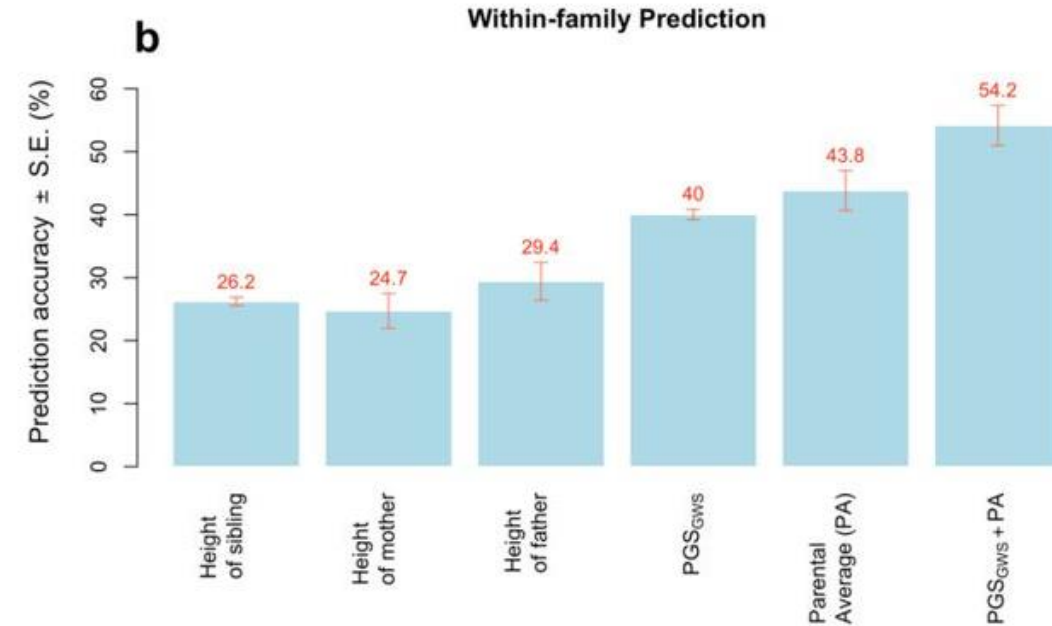
- 5.4 million participants necessary to identify most relevant common variant effects underlying height
- PRS explain 80% of total SNP h^2 (40% vs 50%)
- But poor performance for non-Europeans (more diverse GWAS necessary)

- Yengo, Loic, et al. "A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries." *bioRxiv* (2022).



Will we ever have enough sample size?

- Combination of parental height and PRS showed best performance (54.2)
- Higher measurement error in psychiatry likely implies sample sizes higher than 5 million participants, but likely within reach in the future
- Potential for PRS to improve prediction beyond family history



Do we *really* need good measures of psychopathology?

- Obviously, we want to measure psychopathology as accurate and precise as possible
- But better measures usually result in lower sample size due to higher costs or time requirements

Example Scenario

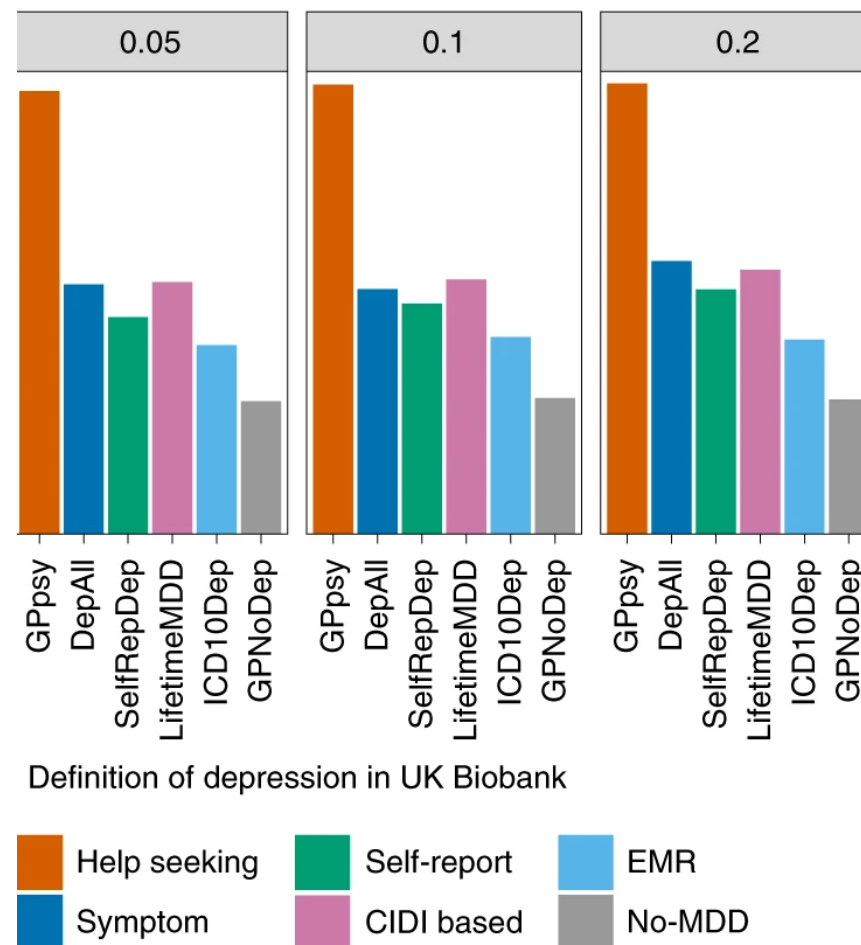
- You want to predict the occurrence of depression based on common genetic variants
- To create this polygenic risk score, you need estimates of genetic associations based on a GWAS of depression
- You have two GWAS to choose, which one do you pick?
- Reference: Cai et al., Minimal phenotyping yields genome-wide association signals of low specificity for major depression, Nature Genetics (2020)

Which discovery GWAS?

- Help-seeking
- “Seen doctor for nerves, anxiety, tension or depression”
- Yes: 113,262
- No: 219,360
- CIDI-based
- Extensive self-report questionnaire analogous to clinical interview
- Depression: 16,301
- No depression: 50,870

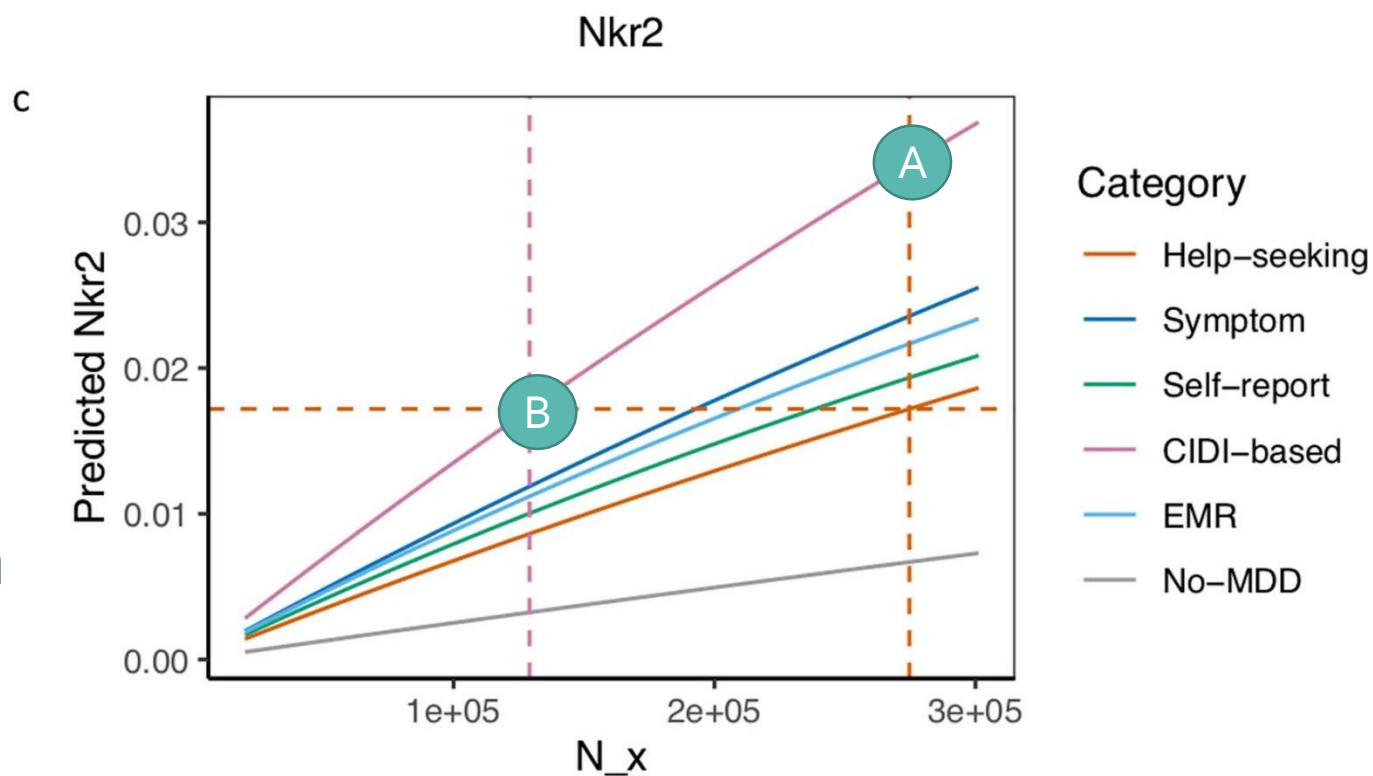
Results

- Discovery GWAS: UK Biobank
- Testing set: PGC29-MDD case-control study
- Y-axis: Variance explained by PRS in independent replication sample
- PRS based on help seeking GWAS (orange) predicts MDD best



Simulated Data

- Y-Axis: Variance explained
- X-Axis: Sample size
- B: CIDI-based GWAS equal performance to help-seeking GWAS at $n=130,000$
- A: CIDI-based GWAS much better performance at equal sample size ($n=330,000$)



Drawbacks of minimal phenotyping

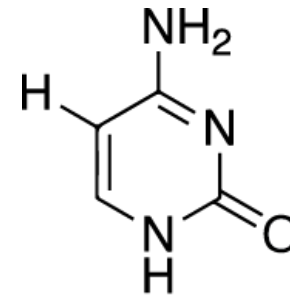
- Conclusion: Always use the cheapest, shortest measure available and get as many participants as possible?
- Not necessarily, Cai et al. argue that the minimal approach leads to results, which are less specific to depression and capture other related traits, such as neuroticism
- So GWAS using minimal phenotyping might miss more specific genetic effects

Epigenetics

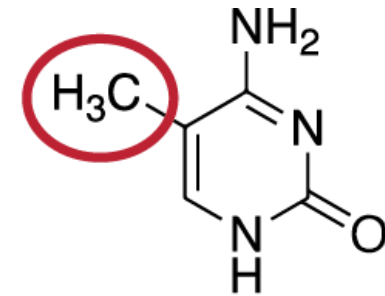
- Study of changes in gene function, which are not due to structural changes in DNA
 - ▶ Note, some definitions include heritable, but this mostly refers to mitosis, not meiosis
- One of the most well-studied mechanism: DNA Methylation

DNA methylation

- Addition of DNA methylation to C
- Typically found at CpG sites (e.g. sequences like this CGCGCG...)
- Usually inhibits gene expression (e.g. DNA methylation in promoter regions)



Cytosine



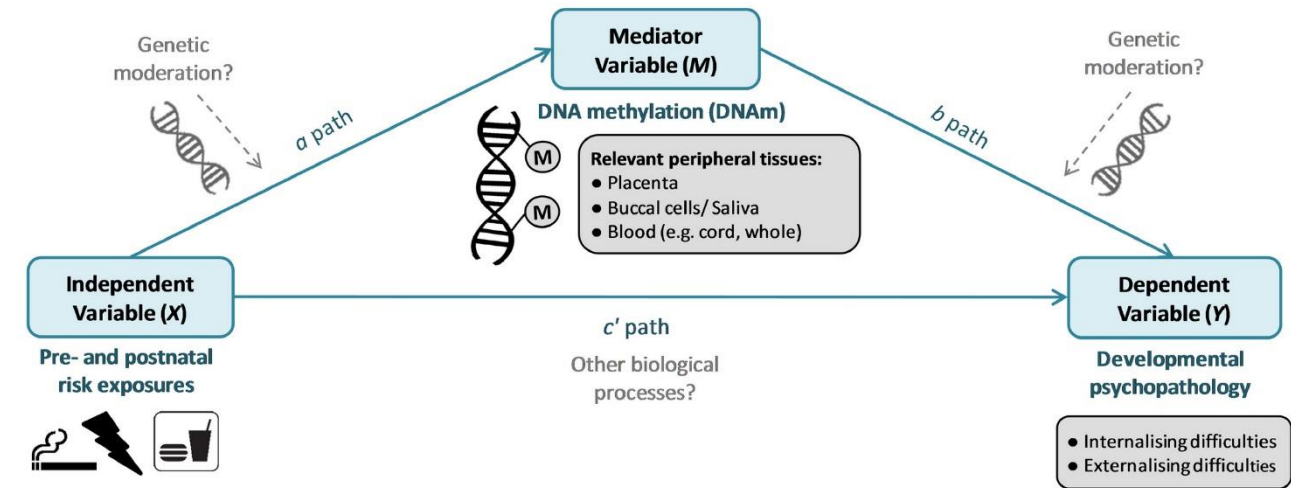
methylated Cytosine

https://commons.wikimedia.org/wiki/File:DNA_methylation.png

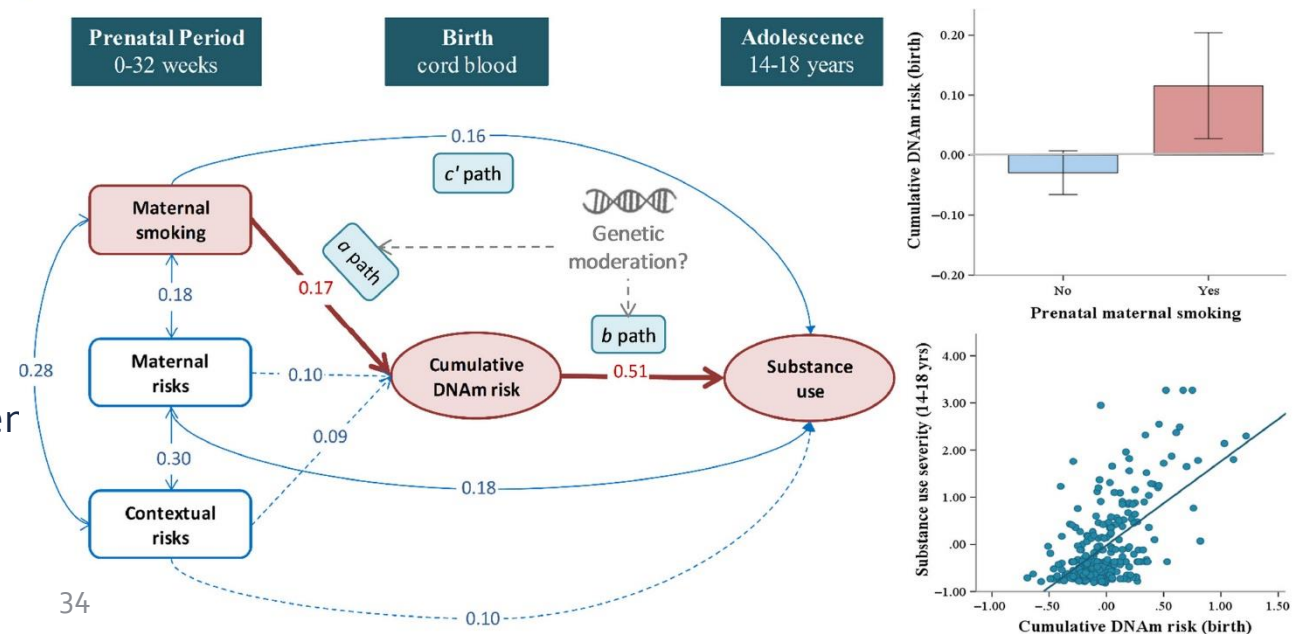
Mediation

- DNA methylation is affected by both genetics and environment
- Thus a potential mediator or marker of environmental risk factors
 - ▶ But can be also a (partial) marker for genetic effects
- Review: Barker, Edward D., Esther Walton, and Charlotte AM Cecil. "Annual Research Review: DNA methylation as a mediator in the association between risk exposure and child and adolescent psychopathology." *Journal of Child Psychology and Psychiatry* 59.4 (2018): 303-322.

(A) Conceptual model



(B) Empirical example



EWAS of ADHD (Neumann et al. 2020)

Table 2 EWAS results.

| CpG | Gene | Chr | Position | Birth methylation | | | | | School-age methylation | | | | |
|------------|------------|-----|-------------|----------------------|------|-------|------|----------|------------------------|------|-------|------|------|
| | | | | n_{studies} | n | B | SE | p | n_{studies} | n | B | SE | p |
| cg25520701 | CREB5 | 7 | 28,800,657 | 6 | 2450 | -3.53 | 0.60 | 4.95E-09 | 5 | 2279 | -0.13 | 1.09 | 0.94 |
| cg24838839 | Intergenic | 5 | 61,031,569 | 6 | 2468 | -4.15 | 1.79 | 3.95E-08 | 5 | 2287 | 1.52 | 1.38 | 0.33 |
| cg22997238 | Intergenic | 7 | 36,014,218 | 6 | 2465 | -1.63 | 0.30 | 8.81E-08 | 5 | 2291 | -0.06 | 0.47 | 0.94 |
| cg21600027 | Intergenic | 4 | 124,443,502 | 6 | 2464 | -3.04 | 0.81 | 2.64E-08 | 5 | 2281 | 0.98 | 0.89 | 0.33 |
| cg17876201 | ZBTB38 | 3 | 141,139,991 | 6 | 2457 | -4.41 | 1.20 | 7.58E-09 | 4 | 2066 | 0.56 | 1.32 | 0.73 |
| cg11251614 | PPIL1 | 6 | 36,839,846 | 6 | 2451 | -3.43 | 0.68 | 3.89E-08 | 5 | 2276 | 0.77 | 1.52 | 0.68 |
| cg09762907 | TRERF1 | 6 | 42,290,256 | 6 | 2460 | -2.11 | 0.39 | 8.76E-08 | 5 | 2284 | -0.55 | 0.64 | 0.46 |
| cg09158638 | Intergenic | 16 | 62,309,996 | 6 | 2470 | -2.55 | 1.40 | 1.89E-08 | 5 | 2270 | -0.33 | 1.04 | 0.80 |
| cg01271805 | ERC2 | 3 | 55,694,954 | 6 | 2469 | -2.86 | 1.71 | 5.24E-08 | 5 | 2289 | 0.28 | 0.73 | 0.76 |

Chr chromosome, n_{studies} number of studies, n number of participants, B regression coefficient, SE standard error.

Genetics

vs

Epigenetics

- No reverse causality
- Confounding possible (population stratification, gene-environment correlation)
 - ▶ but more limited relative to most observational studies
- Assessment time irrelevant
- Tissue independent

- Reverse causality possible
- Various sources of possible genetic and environmental confounding factors
- DNA methylation changes over time, so assessment age important
- Tissue-specific

Polygenic Scores = Methylation Scores?

- Can we apply the same PGS methods to DNAm data?
- In principle yes, in practice one big obstacle:
 - ▶ Correlation structure between CpG sites depends on tissue and result of dynamic influences
 - e.g. time and environmental exposures
 - ▶ No well-defined static LD structure

EWAS vs Methylation Score

- EWAS:
 - $y \sim \beta_1 * CpG_1 + Cov;$
 - $y \sim \beta_2 * CpG_2 + Cov;$
 - $y \sim \beta_3 * CpG_2 + Cov;$
- What we want:
 - $y \sim \beta_1 * CpG_1 + \beta_2 * CpG_2 + \beta_3 * CpG_3 + Cov$

EWAS to Methylation Score

| | CpG1 | CpG2 | CpG3 | CpG | β |
|------|------|------|------|------|---------|
| CpG1 | 1 | 0.8 | -0.1 | CpG1 | 0.4 |
| CpG2 | 0.8 | 1 | -0.2 | CpG2 | 0.3 |
| CpG3 | -0.1 | -0.2 | 1 | CpG3 | -0.6 |

$$y \sim 0.4 * \text{CpG1} + 0.3 * \text{CpG2} - 0.6 * \text{CpG3} + \text{Cov} \quad ?$$

EWAS to Methylation Score

| | CpG1 | CpG2 | CpG3 | CpG | β |
|------|------|------|------|------|---------|
| CpG1 | 1 | 0.8 | -0.1 | CpG1 | 0.4 |
| CpG2 | 0.8 | 1 | -0.2 | CpG2 | 0.3 |
| CpG3 | -0.1 | -0.2 | 1 | CpG3 | -0.6 |

$$y \sim 0.4 * CpG1 + 0.3 * CpG2 - 0.6 * CpG3 + Cov$$

Independent CpG Effects

- Instead of finding ways to translate marginal single CpG situations to obtain conditional effects, independent of all others, why not directly fit all CpG sites in one regression?
- Problem:
 - ▶ More predictors (400,000-800,000) than n
 - ▶ Risk for overfitting

Solution: Elastic Net Regression

- Elastic Net is a regularized/shrinkage approach:
 - ▶ Non-important predictors have 0 coefficients
 - ▶ Other predictors have shrunken coefficients to account for overfitting
- Lambda: Amount of Shrinkage
- Alpha: Between 0 and 1
 - ▶ 1: Sparse/Parsimonious model, out of correlated coefficients only one selected
 - ▶ 0: Keep correlated coefficient and assign similar coefficient
 - ▶ 0-1: In-between values possible

Elastic Net example

| | CpG1 | CpG2 | CpG3 | CpG | β |
|------|------|------|------|------|---------|
| CpG1 | 1 | 0.8 | -0.1 | CpG1 | 0.4 |
| CpG2 | 0.8 | 1 | -0.2 | CpG2 | 0.3 |
| CpG3 | -0.1 | -0.2 | 1 | CpG3 | -0.6 |

Alpha 1: $y \sim 0.3 * CpG1 + 0 * CpG2 - 0.5 * CpG3 + Cov$

Alpha 0: $y \sim 0.16 * CpG1 + 0.14 * CpG2 - 0.5 * CpG3 + Cov$

Cross-validation to determine best alpha and lambda

$n = 12$
 $k = 3$



Test



Train

Data



<https://commons.wikimedia.org/wiki/File:KfoldCV.gif>

Elastic-Net demo

- https://github.com/inDEPTHlab/PRS/blob/main/presentations/elastic_net_demo.pdf

Summary

- Genetic risk scores are widely used in research, but not in clinical use
- Advances in methods and increases in sample size should increase utility of PRS in both research and clinical settings
- Methylation risk scores are only now starting to be utilized in psychological and psychiatric research
- Dynamic nature of DNA methylation both a major source of potential (think treatment monitoring or environmental exposome marker), but also major methodological challenge
- Integration of genetic and epigenetic information necessary to understand intergenerational transmission