

Polygenic Scores

Patrick Turley

RSF Summer Institute

14 June 2019

Outline

- 1. *SNP Heritability***
2. Polygenic Scores
3. Predictive Power of Polygenic Scores
4. Constructing Polygenic Scores
5. Potential Uses and Limitations
6. Applications

Consider the additive model (ignoring confounds for simplicity):

$$\tilde{y}_i = A(\mathbf{x}_i) + \epsilon_i = \sum_{j=1}^J \beta_j x_{ij} + \epsilon_i.$$

Narrow heritability: $h_A^2 \equiv \frac{\text{Var}(A(\mathbf{x}_i))}{\text{Var}(\tilde{y}_i)}$.

In genome-wide data, we only measure $K \ll J$ genetic variants, all SNPs.

Without loss of generality, index these as variants $j = 1, 2, \dots, K$.

Define:

$$A_{\text{SNP}}(\mathbf{x}_i) \equiv \sum_{j=1}^K b_j x_{ij},$$

where $A_{\text{SNP}}(\mathbf{x}_i)$ is the best linear predictor using just the first K SNPs.

We can thus re-express the model:

$$\tilde{y}_i = A_{\text{SNP}}(\mathbf{x}_i) + \epsilon_{\text{SNP},i} = \sum_{j=1}^K b_j x_{ij} + \epsilon_{\text{SNP},i}.$$

$$\text{SNP heritability} : h_{\text{SNP}}^2 \equiv \frac{\text{Var}(A_{\text{SNP}}(\mathbf{x}_i))}{\text{Var}(\tilde{y}_i)}.$$

By definition, $h_{\text{SNP}}^2 \leq h_A^2$.

Define:

$$A_{\text{SNP}}(\mathbf{x}_i) \equiv \sum_{j=1}^K b_j x_{ij},$$

where $A_{\text{SNP}}(\mathbf{x}_i)$ is the best linear predictor using just the first K SNPs.

We can thus re-express the model:

$$\tilde{y}_i = A_{\text{SNP}}(\mathbf{x}_i) + \epsilon_{\text{SNP},i} = \sum_{j=1}^K b_j x_{ij} + \epsilon_{\text{SNP},i}.$$

$$\text{SNP heritability} : h_{\text{SNP}}^2 \equiv \frac{\text{Var}(A_{\text{SNP}}(\mathbf{x}_i))}{\text{Var}(\tilde{y}_i)}.$$

By definition, $h_{\text{SNP}}^2 \leq h_A^2$.

SNP heritability can be estimated from individual-level genome-wide data or with GWAS summary statistics.

It is an upper bound on the total predictive power of the SNPs identified by GWAS.

It is a lower bound on narrow heritability.

In the limit of sequencing data—with every genetic variant measured—SNP heritability will converge to narrow heritability.

Estimating SNP Heritability

Two main types of methods:

1. Molecular-level data

- E.g., GREML (Yang et al., 2010), LDAK (Speed et al. 2010)

2. Summary-statistic-level data

- E.g., Linkage-disequilibrium (LD) score regression (Bulik-Sullivan et al., 2015), SumHer (Speed and Balding 2019)

Some shared assumptions, some distinct.

Some Estimated SNP Heritabilities (in European-descent populations)

| Phenotype | Paper | GWAS N | h^2_{SNP} |
|--------------------|------------------------|----------|--------------------|
| Height | Wood et al. (2014) | 253,288 | 50% |
| BMI | Locke et al. (2015) | 339,224 | 21% |
| EA | Rietveld et al. (2013) | 101,069 | 22% |
| EA | Okbay et al. (2016a) | 394,769 | 22% |
| SWB | Okbay et al. (2016b) | 298,420 | 9%* |
| Age at first birth | Barban et al. (2016) | 238,064 | 15% |

* From Rietveld et al. (2013, *PNAS*).

SNP vs. Narrow Heritability

Estimates of h_{SNP}^2 tends to be one-half to one-third as large as estimates of h_A^2 . Why?

- h_{SNP}^2 does not include the effects of all SNPs.
 - In particular, it misses the effects of rare variants (those with $\text{MAF} < 0.01$), which are not well tagged by SNPs measured in genome-wide data.
- h_A^2 may be overestimated.
 - As we've seen, estimates of h_A^2 from twin studies are biased upward.
 - This is a result of model misspecification by assuming away non-additive effects
- On the other hand, h_{SNP}^2 may be biased upward
 - Assortative mating and indirect parental effects

SNP vs. Narrow Heritability

Estimates of h_{SNP}^2 tends to be one-half to one-third as large as estimates of h_A^2 . Why?

- h_{SNP}^2 does not include the effects of all SNPs.
 - In particular, it misses the effects of rare variants (those with $\text{MAF} < 0.01$), which are not well tagged by SNPs measured in genome-wide data.
- h_A^2 may be overestimated.
 - Estimates of h_A^2 from twin studies can be biased upward.
 - This is a result of model misspecification by assuming away non-additive effects
- On the other hand, h_{SNP}^2 may be biased upward
 - Assortative mating and indirect parental effects

SNP vs. Narrow Heritability

Estimates of h_{SNP}^2 tends to be one-half to one-third as large as estimates of h_A^2 . Why?

- h_{SNP}^2 does not include the effects of all SNPs.
 - In particular, it misses the effects of rare variants (those with $\text{MAF} < 0.01$), which are not well tagged by SNPs measured in genome-wide data.
- h_A^2 may be overestimated.
 - As we've seen, estimates of h_A^2 from twin studies are biased upward.
 - This is a result of model misspecification by assuming away non-additive effects
- On the other hand, h_{SNP}^2 may be biased upward
 - Assortative mating and indirect parental effects

Outline

1. *SNP Heritability*
2. **Polygenic Scores**
3. Predictive Power of Polygenic Scores
4. Constructing Polygenic Scores
5. Potential Uses and Limitations
6. Applications

Start with additive model using measured SNPs:

$$\tilde{y}_i = A_{\text{SNP},i}(\mathbf{x}_i) + \epsilon_{i,\text{SNP}} = \sum_{j=1}^K b_j x_{ij} + \epsilon_{i,\text{SNP}}.$$

True polygenic score : $A_{\text{SNP},i} \equiv \sum_{j=1}^K x_{ij} b_j$.

Variance explained by the true polygenic score is the SNP heritability, h_{SNP}^2 .

Polygenic score (PGS) : Linear predictor of y_i using *estimated* coefficients (Meuwissen et al., 2001; Purcell et al., 2009):

$$\hat{A}_{\text{SNP},i} \equiv \sum_{j=1}^K x_{ij} \hat{b}_j .$$

PGS = True PGS + error

If $\hat{\mathbf{b}}$ is an unbiased estimate of \mathbf{b} , then $\hat{\mathbf{b}} = \mathbf{b} + \mathbf{u}$, where \mathbf{u} is mean-zero estimation error uncorrelated with $\boldsymbol{\beta}$.

$$\hat{A}_{SNP,i} = \sum_{j=1}^K x_{ij} \hat{b}_j = \sum_{j=1}^K x_{ij} (b_j + u_j) = A_{SNP,i} + U_i,$$

where $U_i \equiv \sum_{j=1}^K x_{ij} u_j$ is mean-zero measurement error.

Therefore,

$$E(\hat{A}_i | A_i) = A_i.$$

Outline

1. *SNP Heritability*
2. Polygenic Scores
3. **Predictive Power of Polygenic Scores**
4. Constructing Polygenic Scores
5. Potential Uses and Limitations
6. Applications

Predictive Power of the PGS

If we regress \tilde{y}_i on $\hat{A}_{SNP,i}$, we get an OLS coefficient of

$$\begin{aligned}\alpha &= \frac{\text{Cov}(\hat{A}_{SNP,i}, \tilde{y}_i)}{\text{Var}(\hat{A}_{SNP,i})} \\ &= \frac{\text{Cov}(A_{SNP,i} + U_i, A_{SNP,i} + \epsilon_{i,SNP})}{\text{Var}(A_{SNP,i} + U_i)} \\ &= \frac{\text{Var}(A_{SNP,i})}{\text{Var}(A_{SNP,i}) + \text{Var}(U_i)}\end{aligned}$$

Predictive Power of the PGS

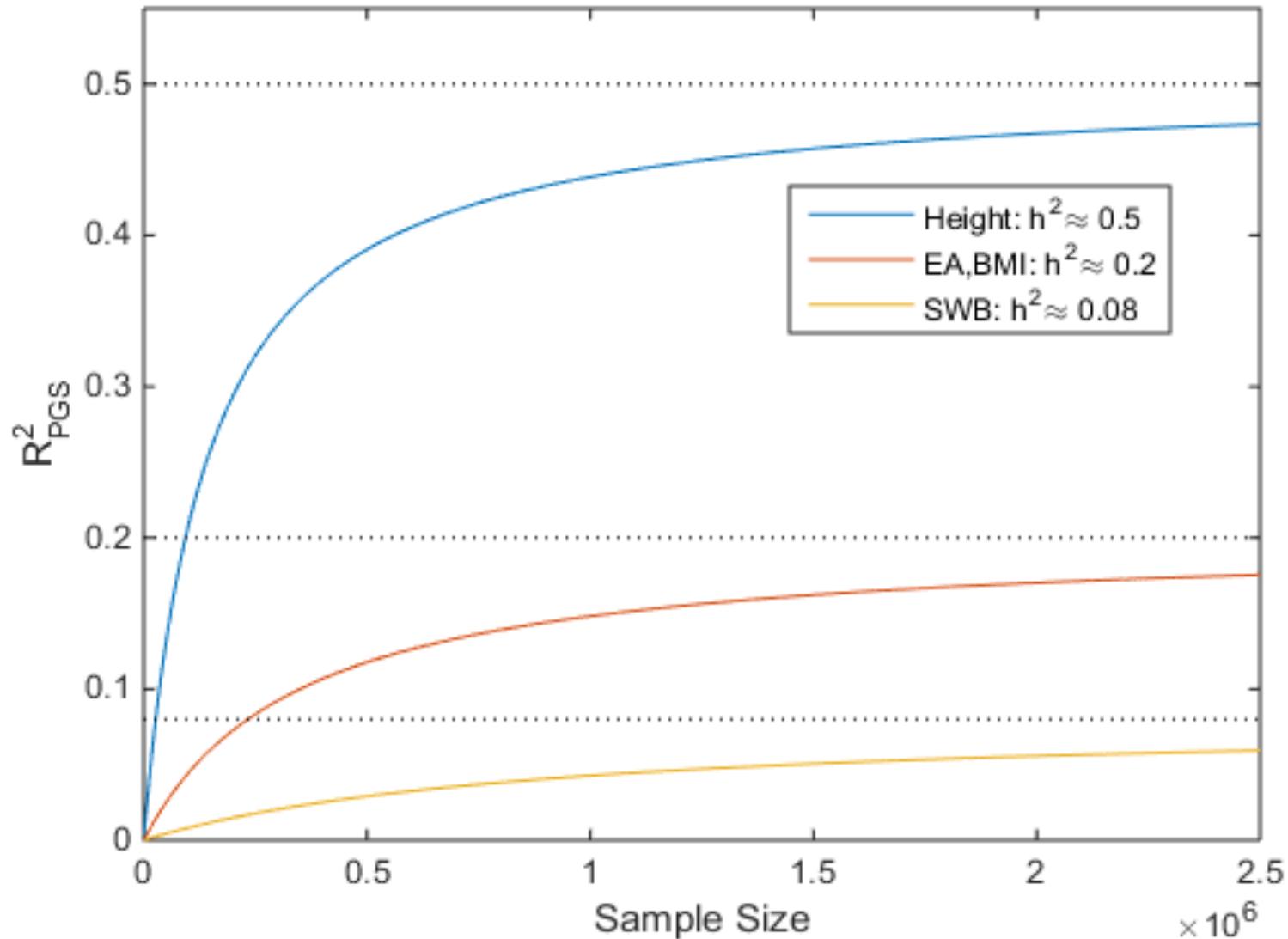
And predictive power is

$$\begin{aligned} E(R^2) &= \frac{a^2 \text{Var}(\hat{A}_{SNP,i})}{\text{Var}(\tilde{y}_i)} \\ &= \frac{\text{Var}(A_{SNP,i})^2 / \text{Var}(\tilde{y}_i)}{\text{Var}(A_{SNP,i}) + \text{Var}(U_i)} \\ &= \frac{[\text{Var}(A_{SNP,i}) / \text{Var}(\tilde{y}_i)]^2}{\text{Var}(A_{SNP,i}) / \text{Var}(\tilde{y}_i) + \text{Var}(U_i) / \text{Var}(\tilde{y}_i)} \\ &= \frac{(h_{SNP}^2)^2}{h_{SNP}^2 + M_e / N} \end{aligned}$$

This is sometimes called the Daetwyler formula (Daetwyler et al. 2008)

- M_e is the effective number of SNPs in the PGS
- M_e is between 50k-70k in genome-wide data (Wray et al. 2013)

Theoretical Projection for R_{PGS}^2



Predictive Power and Heterogeneity

What if we are predicting into a cohort where the genetic architecture isn't the same? (Mostafavi et al. 2019)

Let $A_{SNP,i}^*$ be the additive genetic factor for phenotype \tilde{y}_i^* in the prediction cohort

- So $A_{SNP,i}^* \neq A_{SNP,i} \rightarrow h_{SNP}^{2*} \equiv \text{Var}(A_{SNP,i}^*) / \text{Var}(\tilde{y}_i^*) \neq h_{SNP}^2$
- Define the genetic correlation to be $r_g \equiv \text{Corr}(A_{SNP,i}^*, A_{SNP,i})$

In this setting, we have

$$E(R^2) = \frac{r_g^2 h_{SNP}^2 h_{SNP}^{2*}}{h_{SNP}^2 + M_e/N}$$

(De Vlaming et al. 2016)

Note that this formula will hold even if \tilde{y}_i^* is a different phenotype

Predictive Power and Heterogeneity

What if we are predicting into a cohort where the genetic architecture isn't the same? (Mostafavi et al. 2019)

Let $A_{SNP,i}^*$ be the additive genetic factor for phenotype \tilde{y}_i^* in the prediction cohort

- So $A_{SNP,i}^* \neq A_{SNP,i} \rightarrow h_{SNP}^{2*} \equiv \text{Var}(A_{SNP,i}^*) / \text{Var}(\tilde{y}_i^*) \neq h_{SNP}^2$
- Define the genetic correlation to be $r_g \equiv \text{Corr}(A_{SNP,i}^*, A_{SNP,i})$

In this setting, we have

$$E(R^2) = \frac{r_g^2 h_{SNP}^2 h_{SNP}^{2*}}{h_{SNP}^2 + M_e/N}$$

(De Vlaming et al. 2016)

Note that this formula will hold even if \tilde{y}_i^* is a different phenotype

Measurement Error Correction for Effect of PGS

- Typical application runs a regression of some outcome w_i on the PGS $\hat{A}_{\text{SNP},i}$.
- Effect sizes normally reported in units of SD of the PGS.
 - But interpretation of these units isn't always clear.
 - Scaling of the PGS puts true PGS on wrong scale
 - Attenuation bias varies by GWAS for PGS→PGSs not comparable
- More meaningful to report effect sizes from regression of w_i on the *true* PGS $A_{\text{SNP},i}$.
 - Can adjust estimates to obtain this quantity

Measurement Error Correction for Effect of PGS

- Recall that

$$E(R^2) = \frac{r_g h_{SNP}^2 h_{SNP}^{2*}}{h_{SNP}^2 + M_e/N}$$

- The term, M_e/N , is the term that is a result of measurement error which leads to attenuation bias
- If we know h_{SNP}^2 , M_e , and N , we can get the value of R^2 without attenuation by multiplying by

$$\frac{h_{SNP}^2 + M_e/N}{h_{SNP}^2}$$

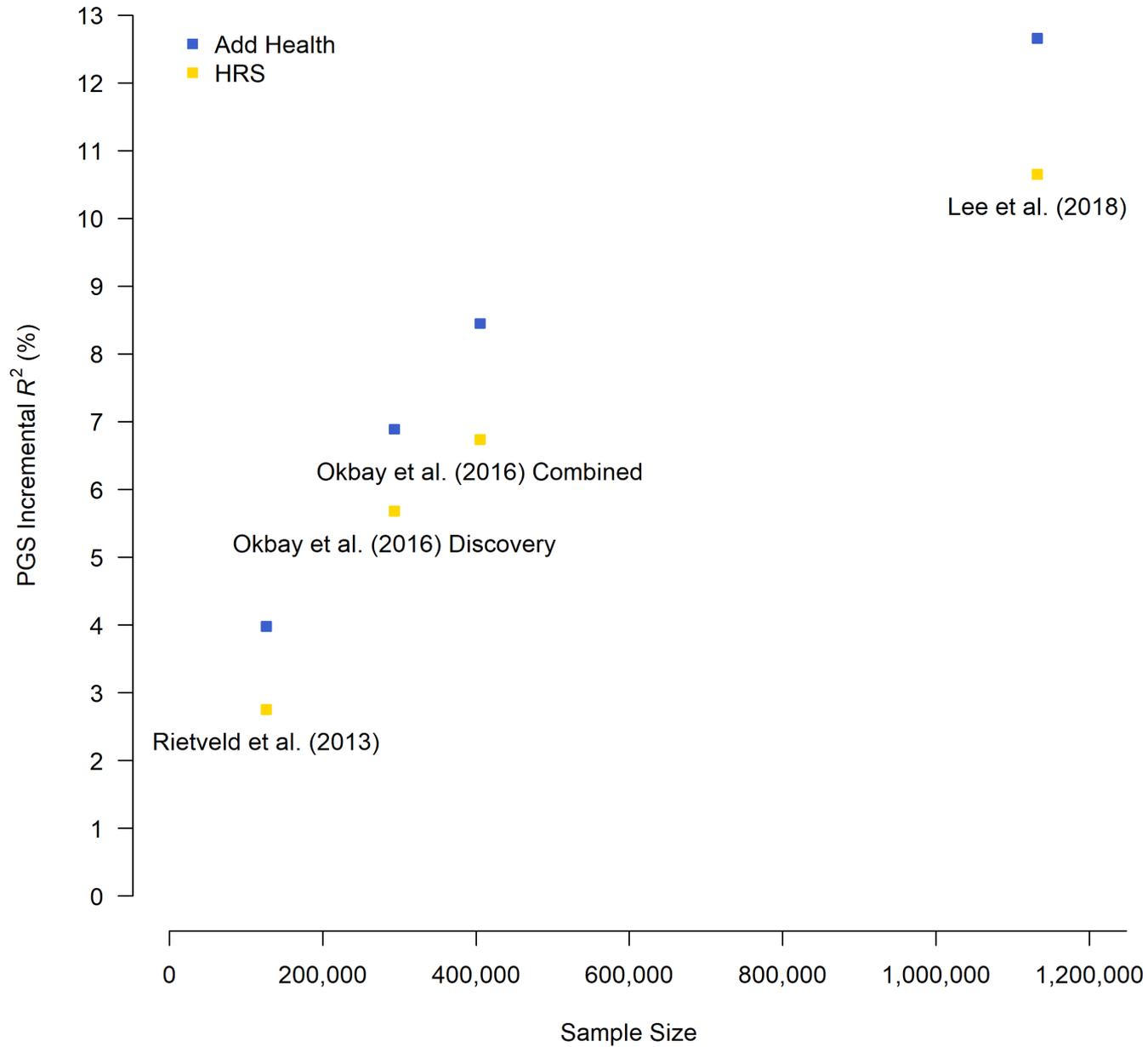
- Similar principles can be used to correct beta estimates in simple and multi-variate regressions

- Major advantage of PGS over specific genetic variants: can have much greater predictive power, especially if J is large.

Some R^2_{PGS} from Published GWAS (in European-descent populations)

| Phenotype | Paper | GWAS N | h^2_{SNP} | R^2_{PGS} |
|--------------------|------------------------|----------|--------------------|--------------------|
| Height | Wood et al. (2014) | 253,288 | 50% | 14% |
| BMI | Locke et al. (2015) | 339,224 | 21% | 7% |
| EA | Rietveld et al. (2013) | 101,069 | 22% | 2% |
| EA | Okbay et al. (2016a) | 394,769 | 22% | 7% |
| SWB | Okbay et al. (2016b) | 298,420 | 9%* | 1% |
| Age at first birth | Barban et al. (2016) | 238,064 | 15% | 1% |

* From Rietveld et al. (2013, *PNAS*).



Source: Lee et al. (2018).

- Major advantage of PGS over specific genetic variants: can have much greater predictive power, especially if J is large.
- Can estimate $\{\hat{b}_j\}$ in large- N GWAS, but then construct PGS in other samples that have measured the J genetic variants.
 - Can be well-powered to study PGS in much smaller samples.
 - E.g., if $R_{\text{PGS}}^2 = 0.07$, then 80% to detect its effect in a sample of size ~ 110 individuals. If $R_{\text{PGS}}^2 = 0.09$, then ~ 85 individuals.
 - Can study PGS in datasets containing high quality measures of outcomes, mediators, and covariates.

Outline

1. *SNP Heritability*
2. Polygenic Scores
3. Predictive Power of Polygenic Scores
4. **Constructing Polygenic Scores**
5. Potential Uses and Limitations
6. Application

Constructing PGSs

Problem: GWAS results give us \hat{b}_j^{GWAS} , not \hat{b}_j .

Why is this a problem?

If we construct $\sum_{j=1}^K x_{ij} \hat{b}_j^{GWAS}$, we will overweight (“double count”) SNPs with high LD scores.

Constructing PGSs

Problem: GWAS results give us \hat{b}_j^{GWAS} , not \hat{b}_j .

Why is this a problem?

If we construct $\sum_{j=1}^K x_{ij} \hat{b}_j^{GWAS}$, we will overweight (“double count”) SNPs with high LD scores.

Three solutions:

1. Pruning: Include only a single SNP from each LD block, the most strongly associated SNP. (Purcell et al., 2009)
 - Reduces R_{PGS}^2 because it makes K smaller.
2. Conditional-joint analysis (COJO): Infer \hat{b}_j 's from \hat{b}_j^{GWAS} 's and LD estimates from a reference panel. (Yang et al., 2012)
 - Key idea: For LD matrix D , $\hat{b}^{GWAS} = D\hat{b}$. Hence we can calculate $\hat{b} = D^{-1}\hat{b}^{GWAS}$.
3. Bayesian approaches accounting for LD
 - LDpred (Vilhjálmsón 2015) or PRS-CS (Ge et al. 2019)

Three solutions:

1. Pruning: Include only a single SNP from each LD block, the most strongly associated SNP. (Purcell et al., 2009)
 - Reduces R_{PGS}^2 because it makes K smaller.
2. Conditional-joint analysis (COJO): Infer \hat{b}_j 's from \hat{b}_j^{GWAS} 's and LD estimates from a reference panel. (Yang et al., 2012)
 - Key idea: For LD matrix \mathbf{D} , $\hat{\mathbf{b}}^{\text{GWAS}} = \mathbf{D}\hat{\mathbf{b}}$. Hence we can calculate $\hat{\mathbf{b}} = \mathbf{D}^{-1}\hat{\mathbf{b}}^{\text{GWAS}}$.
3. Bayesian approaches accounting for LD
 - LDpred (Vilhjálmsón 2015) or PRS-CS (Ge et al. 2019)

Three solutions:

1. Pruning: Include only a single SNP from each LD block, the most strongly associated SNP. (Purcell et al., 2009)
 - Reduces R_{PGS}^2 because it makes K smaller.
2. Conditional-joint analysis (COJO): Infer \hat{b}_j 's from \hat{b}_j^{GWAS} 's and LD estimates from a reference panel. (Yang et al., 2012)
 - Key idea: For LD matrix \mathbf{D} , $\hat{\mathbf{b}}^{GWAS} = \mathbf{D}\hat{\mathbf{b}}$. Hence we can calculate $\hat{\mathbf{b}} = \mathbf{D}^{-1}\hat{\mathbf{b}}^{GWAS}$.
3. Bayesian approaches accounting for LD
 - LDpred (Vilhjálmsón 2015) or PRS-CS (Ge et al. 2019)

Three solutions:

1. Pruning: Include only a single SNP from each LD block, the most strongly associated SNP. (Purcell et al., 2009)
 - Reduces R_{PGS}^2 because it makes K smaller.
2. Conditional-joint analysis (COJO): Infer \hat{b}_j 's from \hat{b}_j^{GWAS} 's and LD estimates from a reference panel. (Yang et al., 2012)
 - Key idea: For LD matrix \mathbf{D} , $\hat{\mathbf{b}}^{\text{GWAS}} = \mathbf{D}\hat{\mathbf{b}}$. Hence we can calculate $\hat{\mathbf{b}} = \mathbf{D}^{-1}\hat{\mathbf{b}}^{\text{GWAS}}$.
3. Bayesian approaches accounting for LD
 - LDpred (Vilhjálmsón 2015) or PRS-CS (Ge et al. 2019)

Bayesian/Shrinkage Approaches

Uses as weights

$$E(\mathbf{b} \mid \hat{\mathbf{b}}^{GWAS}, \mathbf{D})$$

By Bayes' Rule

$$f(\mathbf{b} \mid \hat{\mathbf{b}}^{GWAS}, \mathbf{D}) = \frac{f(\hat{\mathbf{b}}^{GWAS} \mid \mathbf{b}, \mathbf{D})f(\mathbf{b} \mid \mathbf{D})}{f(\hat{\mathbf{b}}^{GWAS} \mid \mathbf{D})}$$

Shrinkage depends on the prior.

Bayesian/Shrinkage Approaches

Potential prior distributions

1. LDpred: Gaussian or Spike-and-Slab

$$(b_j | \mathbf{D}) = b_j \sim \begin{cases} 0 & \text{with probability } \pi \\ N(0, \tau^2) & \text{with probability } 1 - \pi \end{cases}$$

2. PRS-CS: “Continuous shrinkage”

$$(b_j | \mathbf{D}) = b_j \sim N(0, \phi \psi_j)$$

$$\psi_j \sim \text{Gamma}(a, \delta_j)$$

$$\delta_j \sim \text{Gamma}(b, 1)$$

- Parameters a and b determine how aggressively to shrink small estimates and how much you don't shrink large ones
- Complicated prior, but much simpler computationally

Outline

1. *SNP Heritability*
2. Polygenic Scores
3. Predictive Power of Polygenic Scores
4. Constructing Polygenic Scores
5. **Potential Uses and Limitations**
6. Application

Some Uses of PGSs

- Identify correlates of genetic factors
 - Educational attainment PGS predicts early speech acquisition and is mediated by cognitive ability (Belsky et al., 2016).
 - Educational attainment PGS predicts school achievement (e.g., Ward et al., 2014) and the quality of mid-life financial decision making (Barth, Papageorge, and Thom, 2016).
 - Educational attainment PGS substitutes for SES in preventing high school dropout, but complements SES in facilitating college attendance (Papageorge and Thom, 2016).
 - Twin studies could address these questions, but PGSs enable addressing them in other samples and with different assumptions.
- Identify causal effects of genetic factors
 - PGS is randomly assigned conditional on parents.
 - Sibling data and family fixed effects → causal effect of PGS.

- Study treatment effect heterogeneity by genotype
 - Increase of compulsory schooling age in U.K. reduces BMI only among those with a high-BMI PGS (Barcellos, Carvalho, and Turley 2016).
 - Swedish comprehensive schooling reform increased completion of compulsory schooling among those with low-EA PGS, and completion of higher degrees among those with high-EA PGS (Beauchamp, Okbay, Oskarsson, and Thom, 2016).
- Use as control variable
 - To control for confounding genetic factors.
 - To increase statistical power for estimating the effect of a randomized treatment.
 - Rietveld et al. (2013) calculate: if incremental R_{PGS}^2 is 15%, then power increase is equivalent to 17% increase in sample size.
 - Especially useful when treatment is expensive (e.g., providing free preschool).
 - As long as genetic variants are measured, could construct and control for many PGSs.
- Use for balance tests of randomization
 - PGSs should be identically distributed in treatment and control groups (Davies et al. 2016, Barcellos, Carvalho, and Turley 2016).
- Identify at-risk individuals
 - Still speculative, but especially useful at young ages before accurate testing is possible (e.g., dyslexia).

Limitation #1

- *Mechanisms* affecting behavior are poorly understood.
 - Hard to specify what is captured by PGS.
 - Use as structural parameter or IV is difficult to justify.
- Often a tradeoff: Including many genetic variants...
 - Increases predictive power.
 - But requires including many genetic variants with unknown function.
- Will learn more as PGSs are studied.
- Potential for functionally-partitioned PGS in future.

Limitation #2

- Current polygenic scores far less predictive in non-European-descent samples.
- For example, for the EA2 score in HRS:
 - $R^2 \approx 7\%$ for European-ancestry individuals.
 - $R^2 \approx 1\%$ for African-ancestry individuals.
 - $R^2 \approx 1\%$ for self-classified Hispanics.
- Optimal $\hat{\beta}_j$'s may differ for many reasons, including:
 - Different true β_j 's.
 - Different correlations between measured x_{ij} 's and unmeasured x_{ik} 's that are proxied for.
- With larger non-European samples or improved methods, will be able to estimate appropriate $\hat{\beta}_j$'s.

Limitation #3

- Communication is difficult
 - All the critiques of interpreting heritability correspond to polygenic scores
 - What does it mean for a person to have a high polygenic score?
- Even the word “score” may carry value judgments to lay-people
 - If I have a “higher score,” does that make me a better person?

Limitation #4

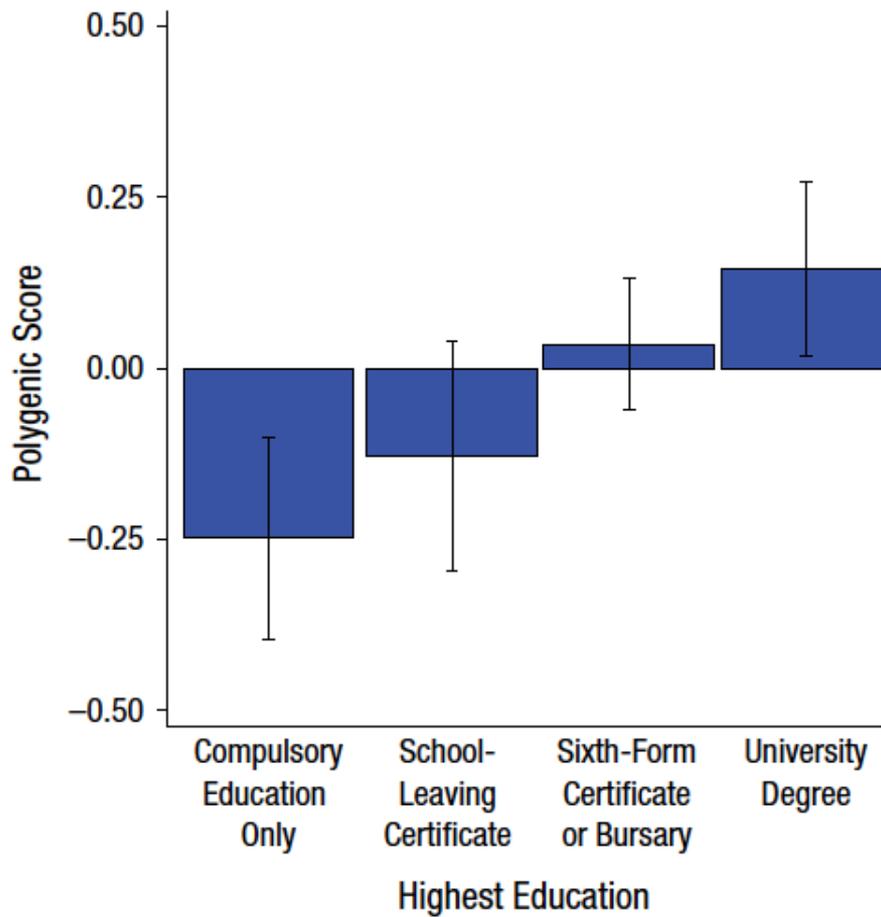
- Two sources of population stratification
- In the discovery phase
 - Leads to bias in the GWAS estimates, so the polygenic score may give more weight to SNPs that just correspond to ancestry
- In the prediction phase
 - If the prediction sample is stratified, this can lead to bias in our PGS-based analyses even if SNP-weights are unbiased
- Interaction of bias in both phases
 - The combination of these two interact so group differences are strongly exaggerated

Outline

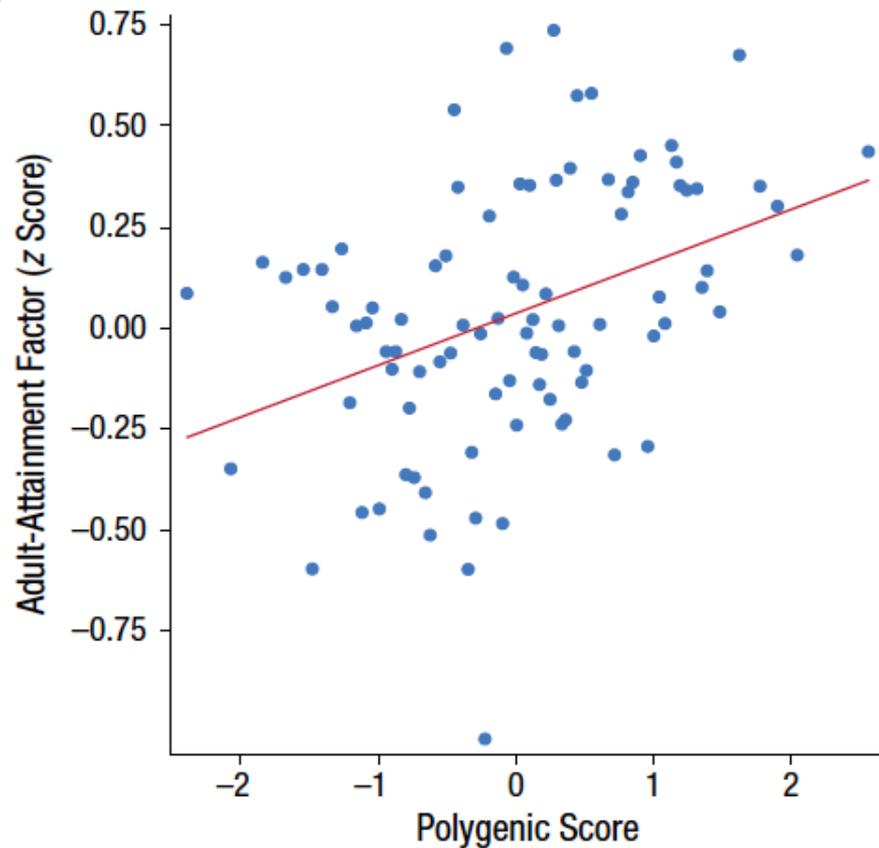
1. *SNP Heritability*
2. Polygenic Scores
3. Predictive Power of Polygenic Scores
4. Constructing Polygenic Scores
5. Potential Uses and Limitations
6. **Application**

Example: Belsky et al. (2016)

- How does the additive genetic component of EA relate to life-course development?
- Used longitudinal Dunedin Study ($N \approx 1,000$).
 - Born 1972-1973 in Dunedin, New Zealand.
 - Rich data from ages 3 through 38.
- Constructed PGS for EA based on EA1 (Rietveld et al, 2013), $R_{\text{PGS}}^2 \approx 2.3\%$.
- “Adult attainment factor”: Constructed from factor analysis of occupation, income, wealth, financial situation, and credit score.

a

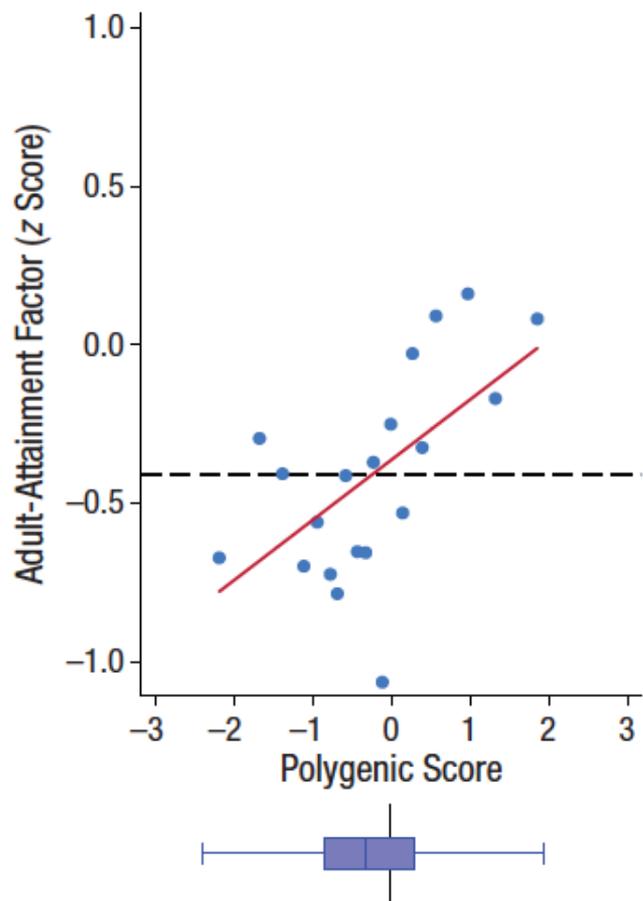
$r = 0.15$ ($p < 0.001$)

b

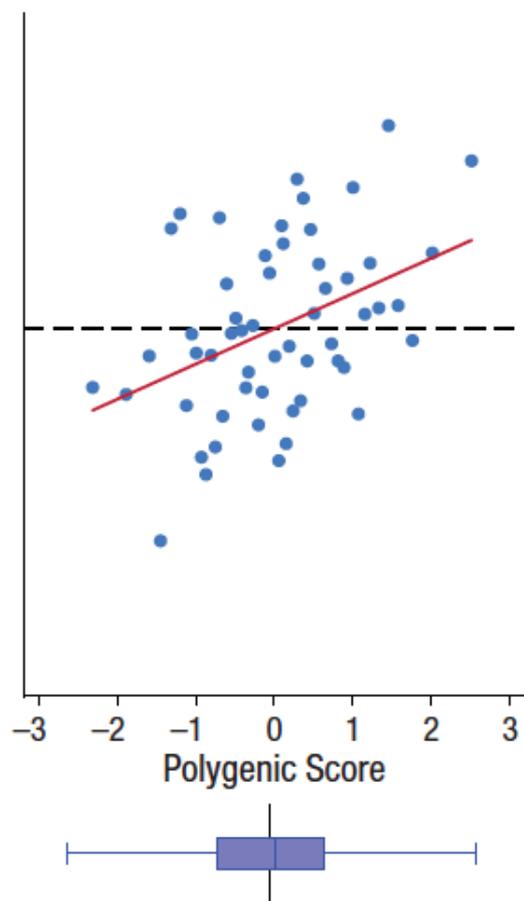
$r = 0.13$ ($p < 0.001$)

Controlling for EA: $r = 0.07$ ($p = 0.035$)

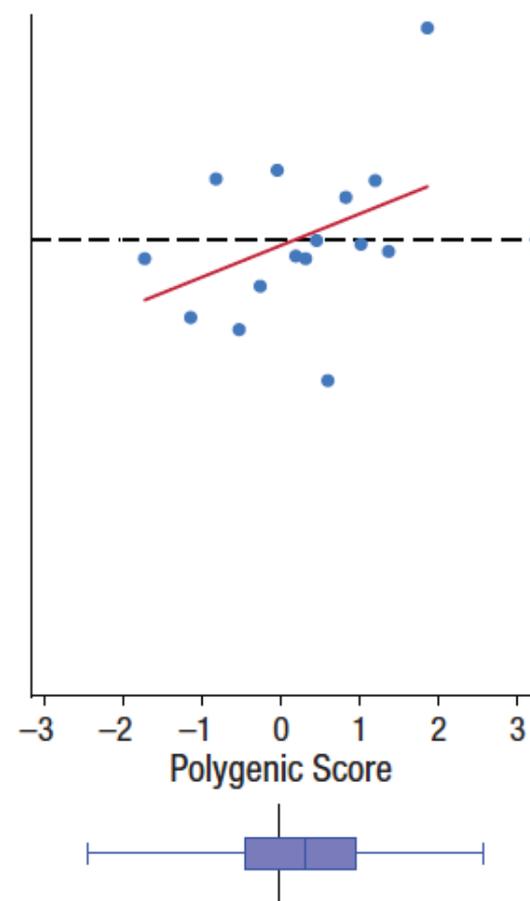
Low-SES Families
($n = 175$)

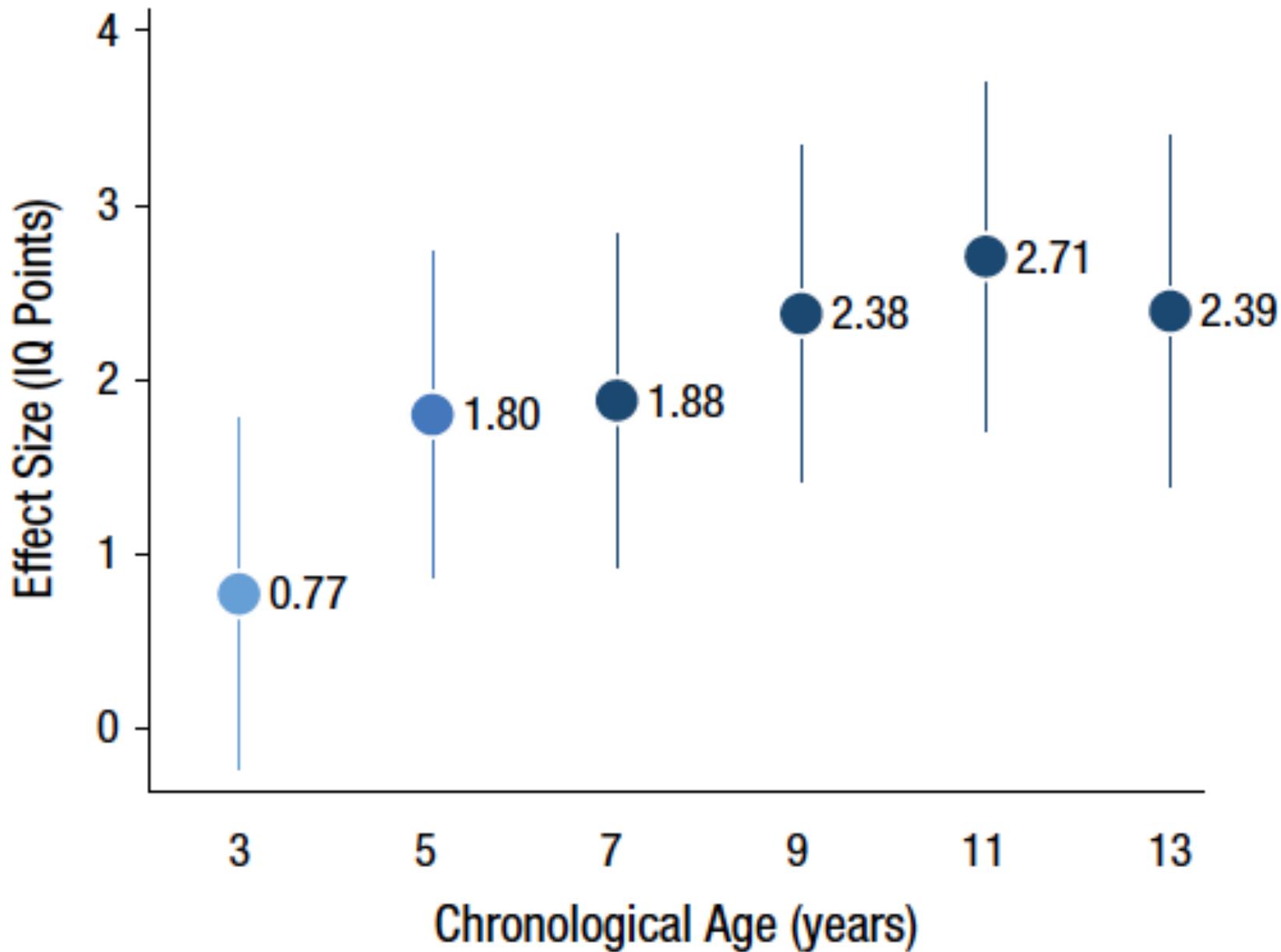


Middle-SES Families
($n = 570$)



High-SES Families
($n = 152$)





Note: Effect of 1 SD increase in PGS.

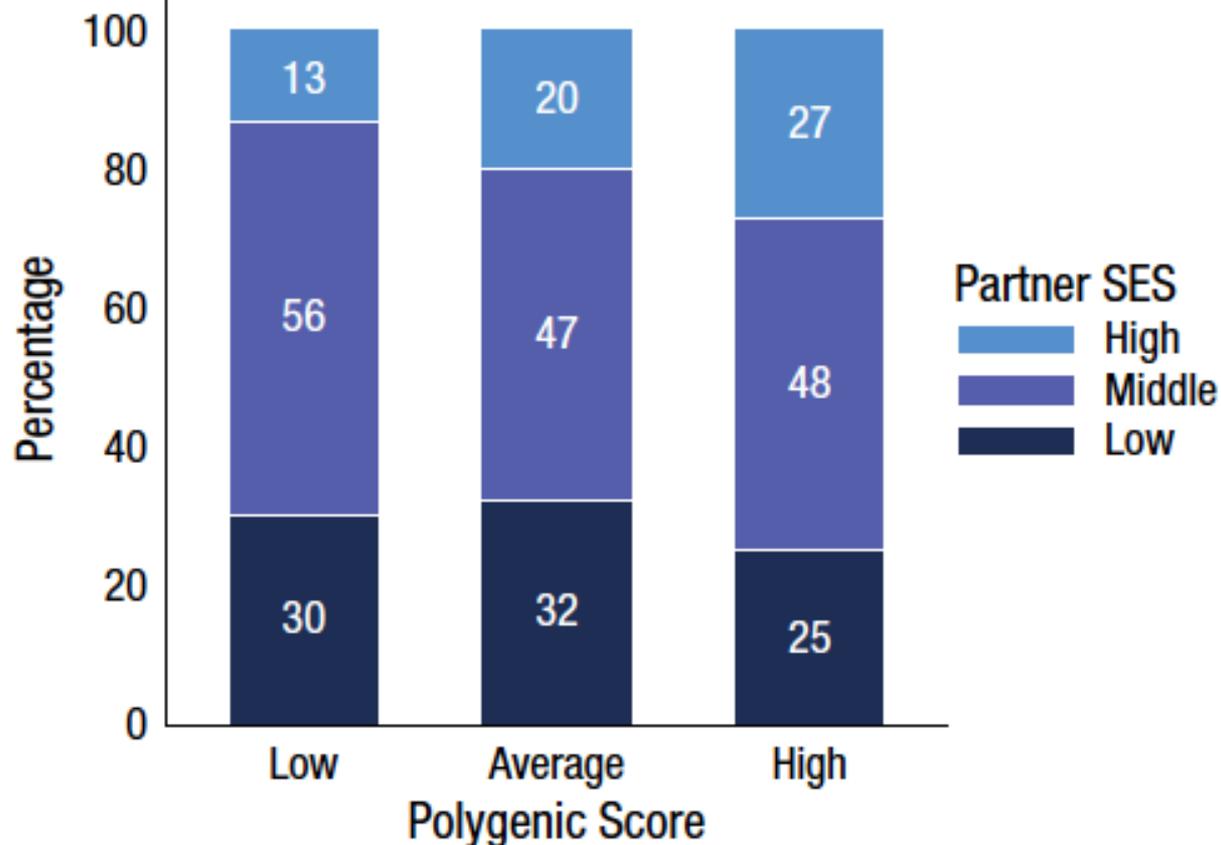


Fig. 5. Association between polygenic score and partner's socioeconomic status (SES). The graph shows the percentages (inside bars) of members who had low-, middle-, and high-SES partners, separately for Dunedin Study members with low polygenic scores (≥ 1 *SD* below the mean; $n = 119$), average polygenic scores (within 1 *SD* of the mean; $n = 504$), and high polygenic scores (≥ 1 *SD* above the mean; $n = 136$). Partners' SES was defined according to whether they had completed a university degree and whether their income was above the national sex-specific median: High-SES partners had a university education and an above-median income, middle-SES partners met only one of these criteria, and low-SES partners met neither criterion.