Evaluation, visualization, and Pitfalls of Polygenic Prediction

Jian Zeng

j.zeng@uq.edu.au

Slide courtesy: Naomi Wray

PGS evaluation in quantitative traits

Prediction accuracy

Squared correlation between phenotype and PGS in the validation sample

- The proportion of phenotypic variance explained by PGS (prediction R^2)
- The SNP-based heritability is its upper bound

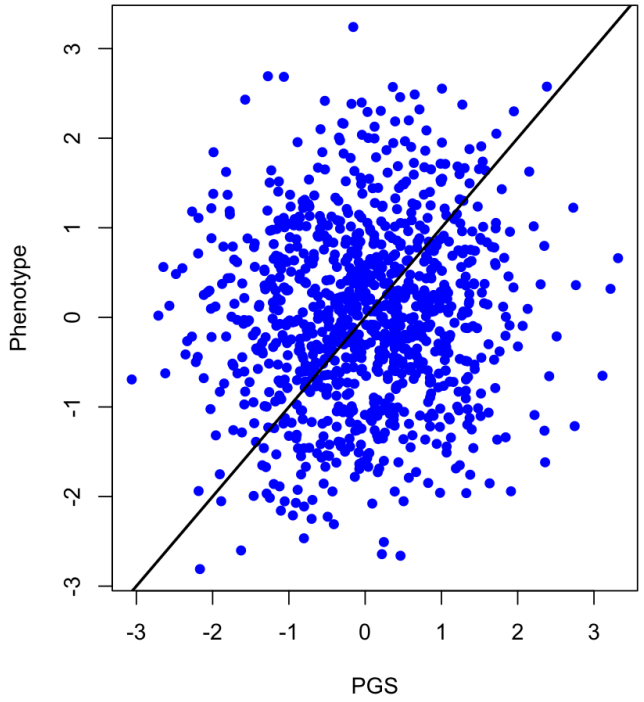
It's common to adjust for covariates (sex, age, top 10 PCs, etc)

- Null model: y = covariates + e
- Full model: y = covariates + PGS + e
- Incremental R^2: $R^2_{Full} - R^2_{Null}$

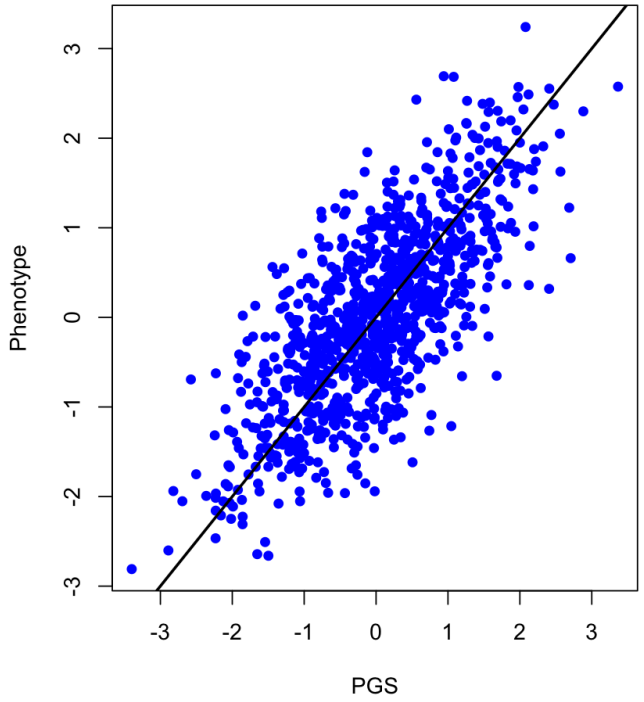
PGS evaluation in quantitative traits

Prediction accuracy

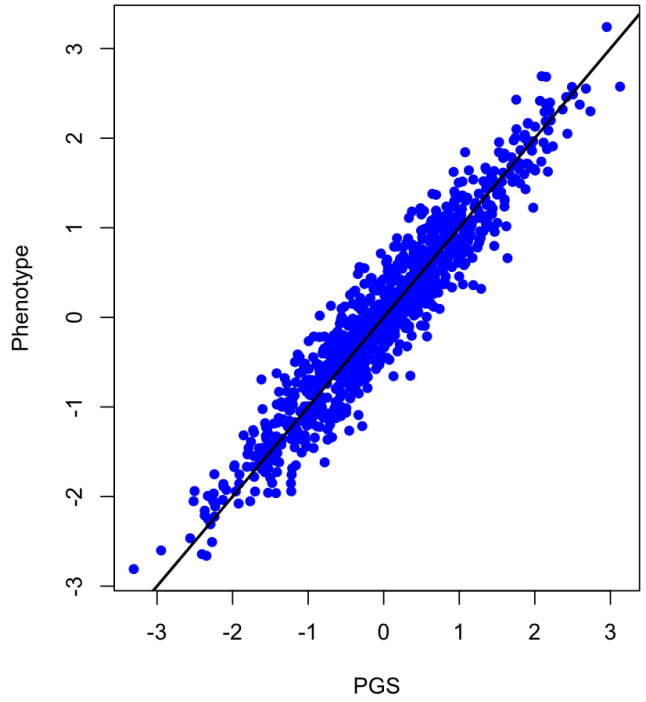
$R^2 = 0$



$R^2 = 0.5$



$R^2 = 0.9$



PGS evaluation in quantitative traits

Prediction bias

The slope of regression of phenotypes on PGS in the validation sample is expected to be 1.

- 1 unit increase in PGS leads to 1 unit increase in phenotype
- The PGS are unbiased

If the slope > 1, then

- 1 unit increase in PGS leads to >1 unit increase in phenotype
- The PGS are downward biased

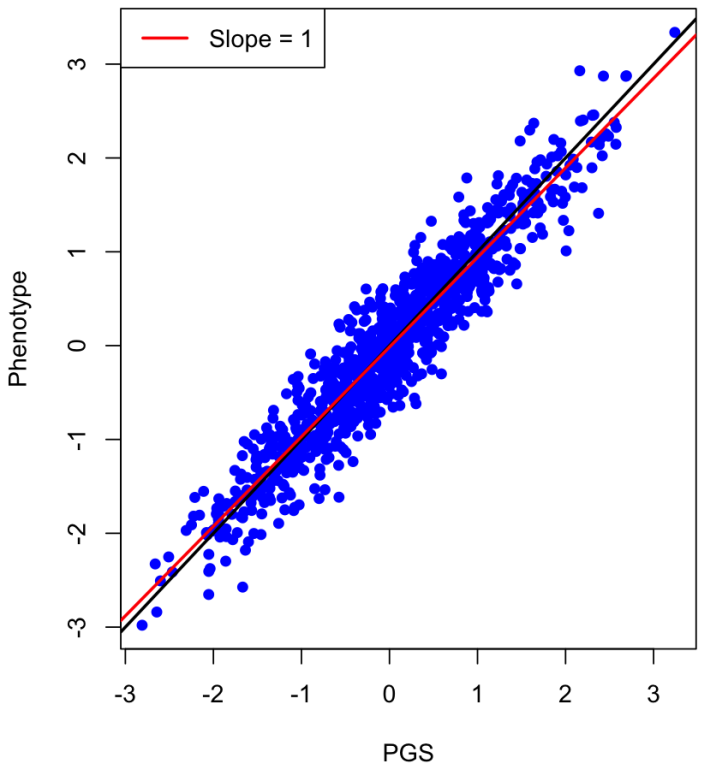
If the slope < 1, then

- The PGS are upward biased

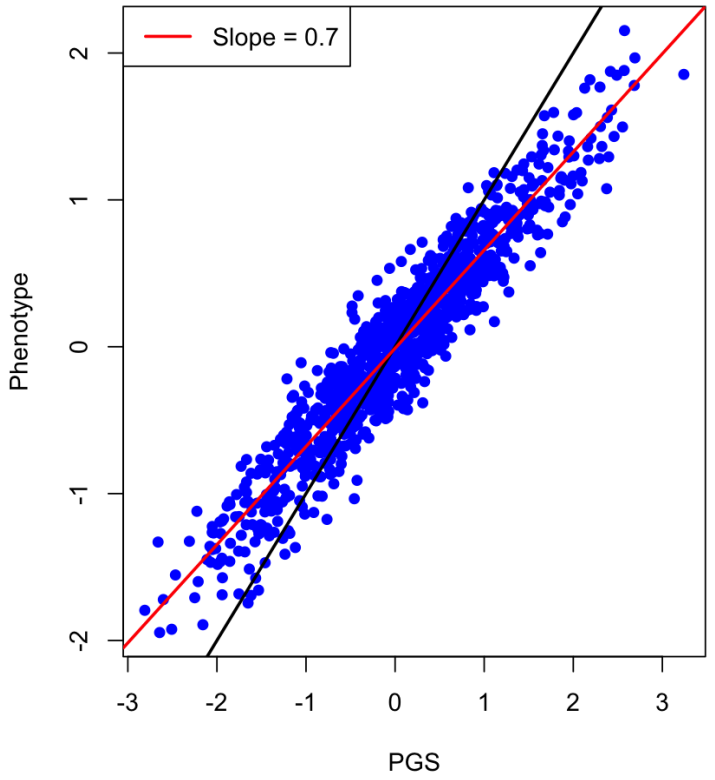
PGS evaluation in quantitative traits

Prediction bias

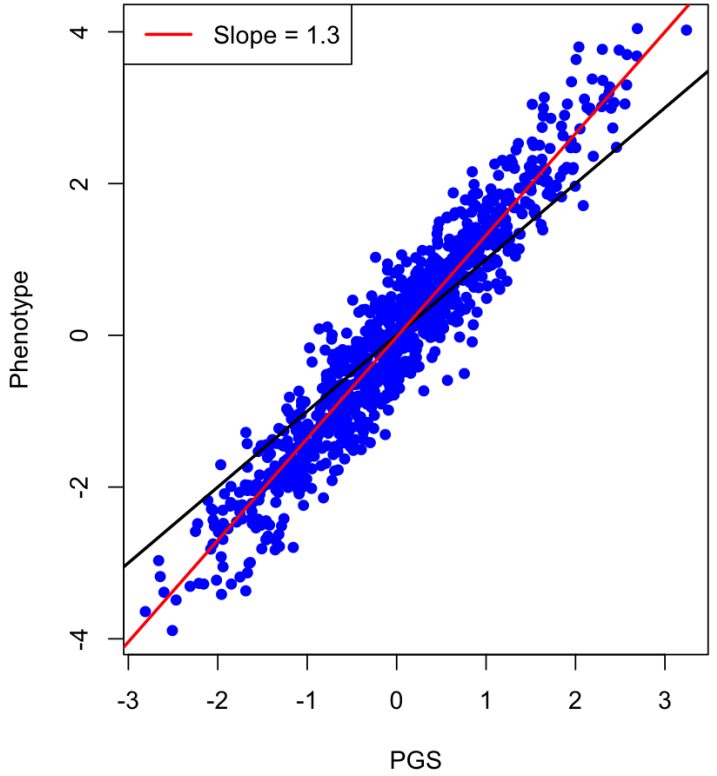
Unbiased



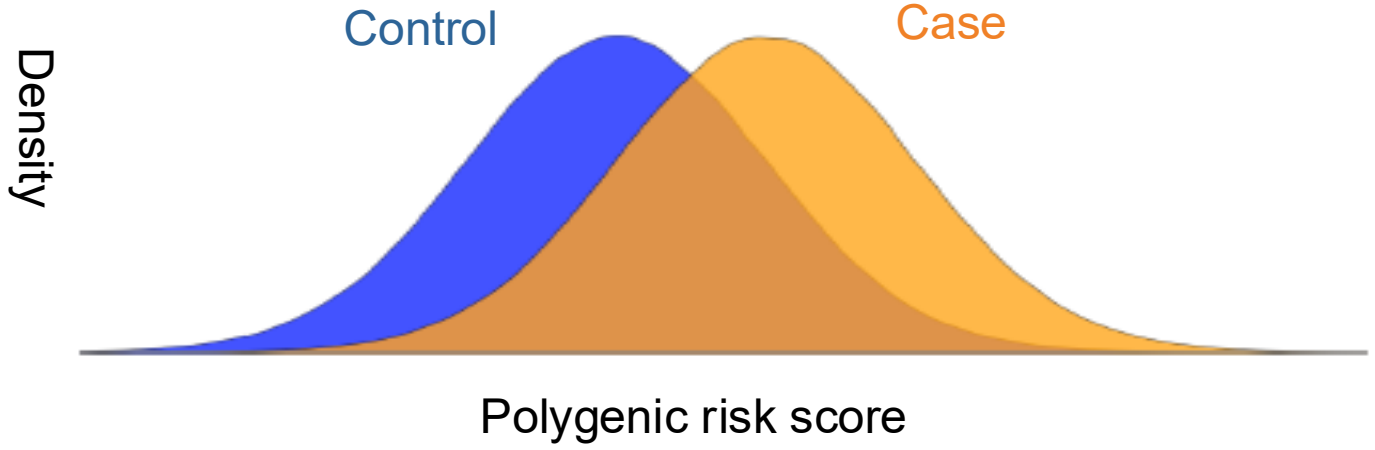
Upward biased

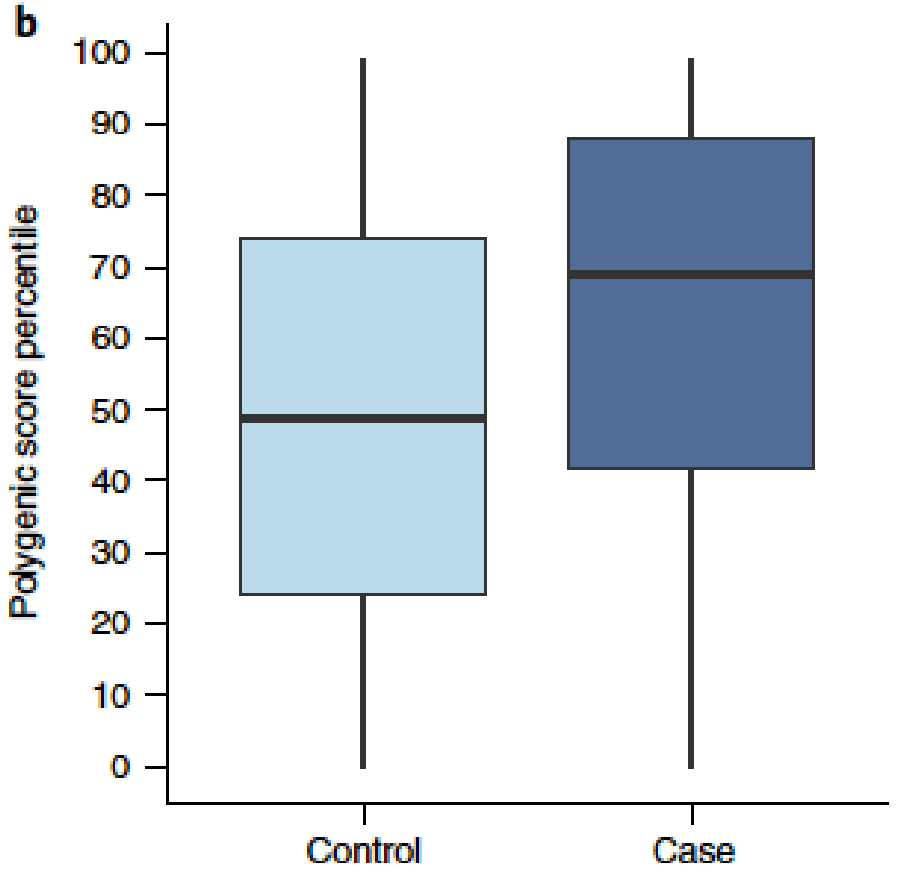


Downward biased



PGS evaluation in diseases (binary traits)





PGS evaluation in diseases (binary traits)

Statistics to measure prediction accuracy

- Pseudo R^2 from logistic regression
- AUC (area under the ROC curve)
- Variance explained on liability scale
- Risk stratification
- Decile odds ratio (OR)

Pseudo R^2

Logistic regression:

- Null model: y = logistic(covariates + e)
- Full model: y = logistic(covariates + PGS + e)

Many pseudo R^2 statistics available for logistic regression

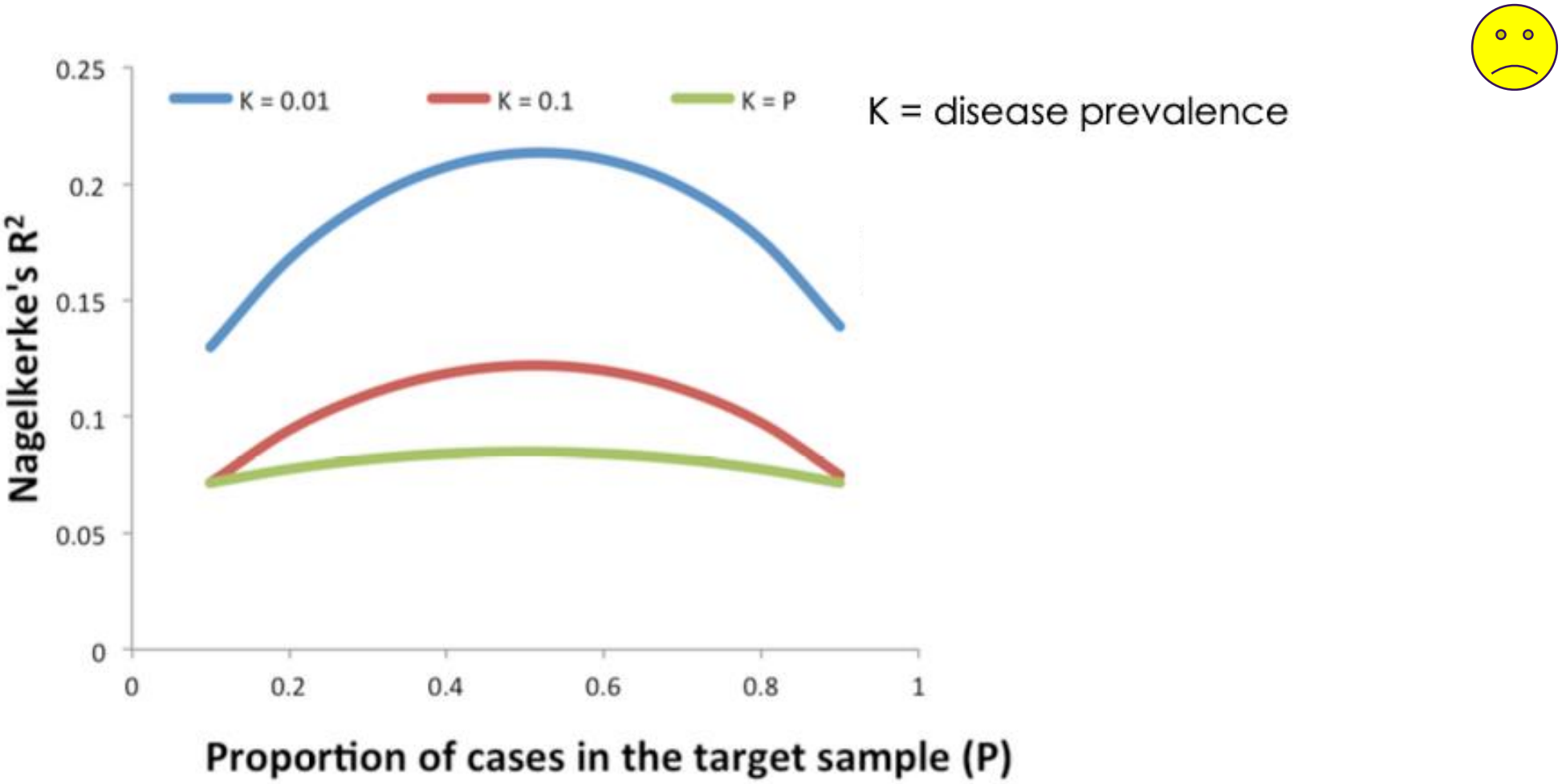
e.g., Nagelkerke's R^2

$$\frac{1-\left(\frac{L_{Null}}{L_{Full}}\right)^{\frac{2}{N}}}{1-(L_{Null})^{\frac{2}{N}}} \in [0,1]$$

For a review of pseudo R^2 statistics, check this link

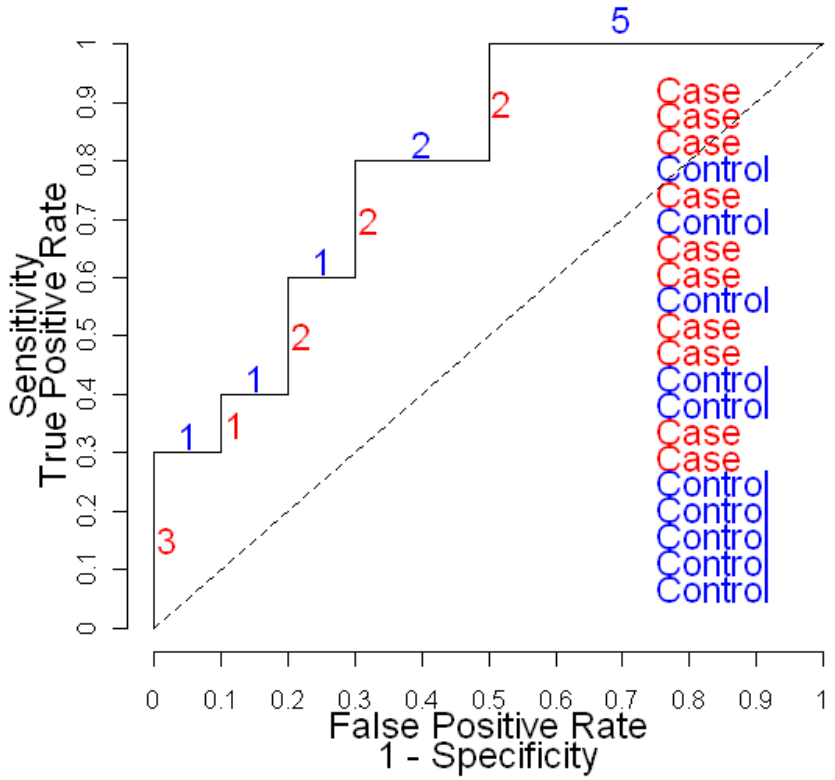
Property of pseudo R^2

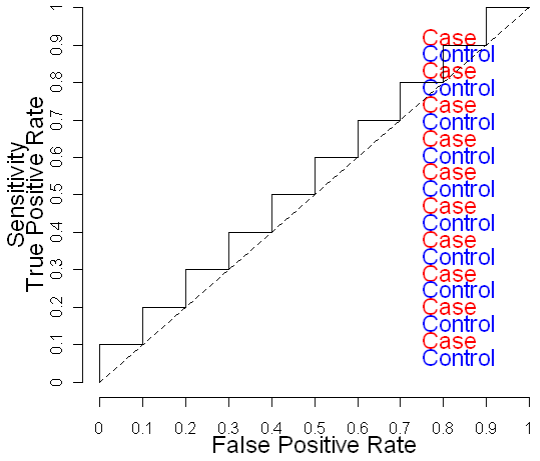
Problem: Nagelkerke's R^2 depends on case proportion in the sample

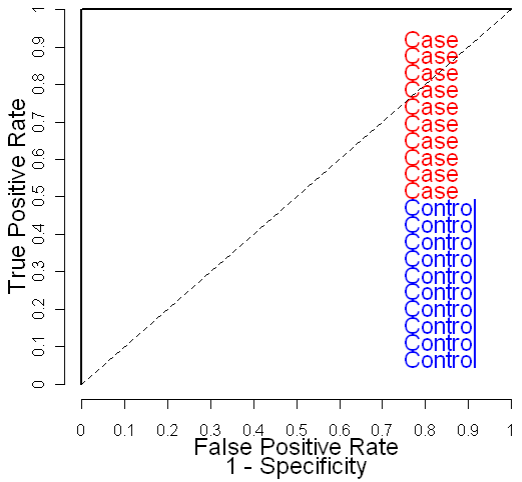


AUC (area under the ROC curve)

AUC = Probability that a randomly selected case has a higher test score than a randomly selected control







Property of AUC

- Nice property - independent to proportion of cases and controls in sample
  - Can be used to compare results between case-control studies
- Max AUC depends on heritability and disease prevalence
  - Use caution when comparing populations with different prevalence

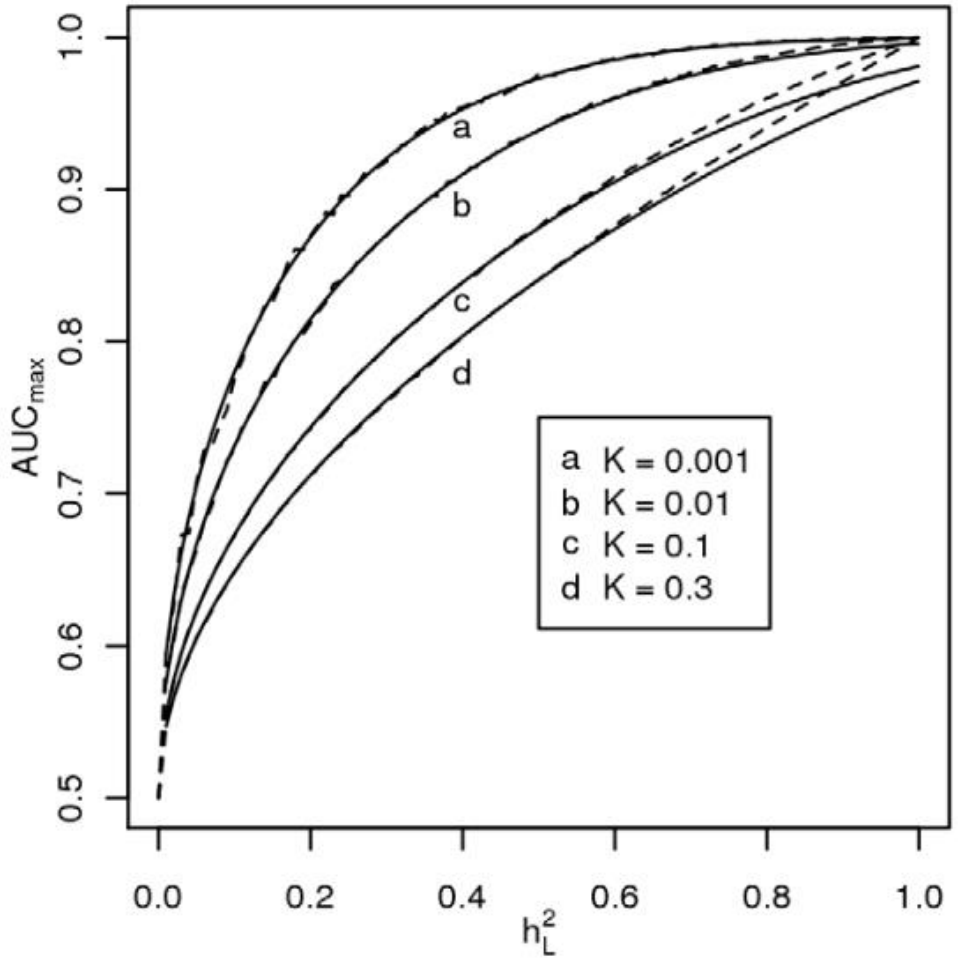


Figure 2. Relationship between maximum AUC (AUC_{max}) from a genomic profile and heritability on the liability scale h_L^2. For

OPEN ACCESS Freely available online PLoS GENETICS

The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling

Naomi R. Wray[1]*, Jian Yang[1], Michael E. Goddard[2,3], Peter M. Visscher[1]

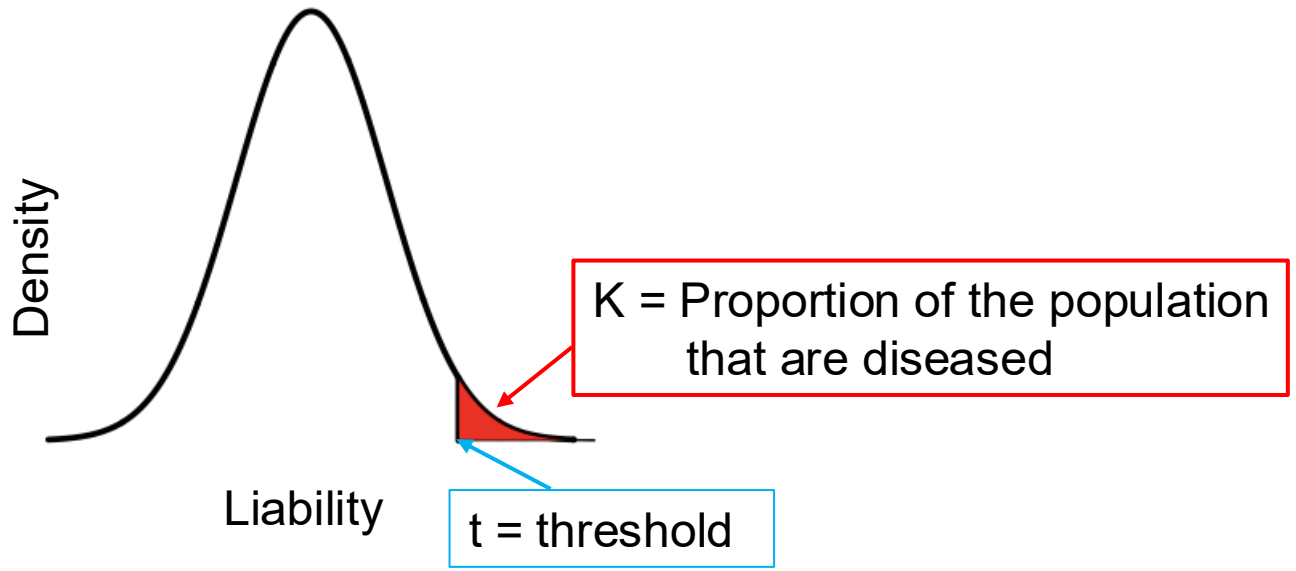
1 Genetic Epidemiology and Queensland Statistical Genetics, Queensland Institute of Medical Research, Brisbane, Australia, **2** Department of Food and Agricultural Systems, University of Melbourne, Melbourne, Australia, **3** Victoria Department of Primary Industries, Melbourne, Australia

Prediction R^2 on liability scale ★

Liability threshold model

Map variance explained on observed probability 0-1 scale (R_o^2)

To underlying unobserved continuous liability scale (R_l^2).



Falconer 1965; Lee et al 2011 AJHG; Lee, 2012, Genet Epidemiol

Prediction R^2 on liability scale ★

Linear regression; Y are 0s and 1s

Null: Y= covariates + e

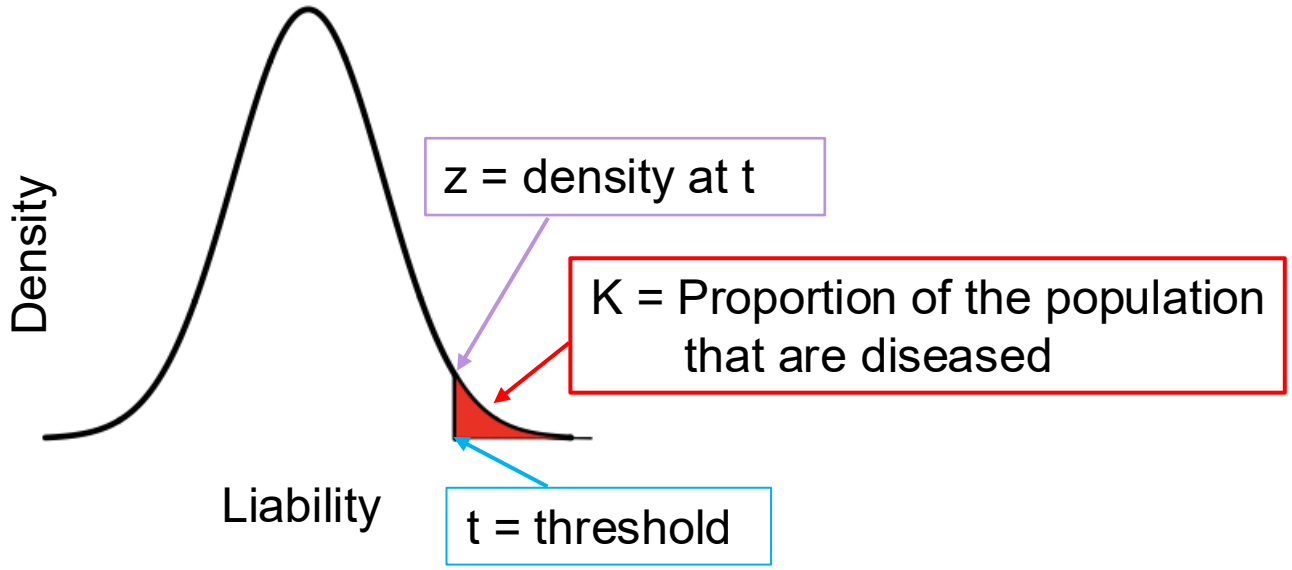
Full: Y= covariates + PGS + e

R^2 on the observed scale

$$R_o^2 = 1 - \left(\frac{Likelihood_{null}}{Likelihood_{full}}\right)^{2/N}$$

R^2 on the liability scale

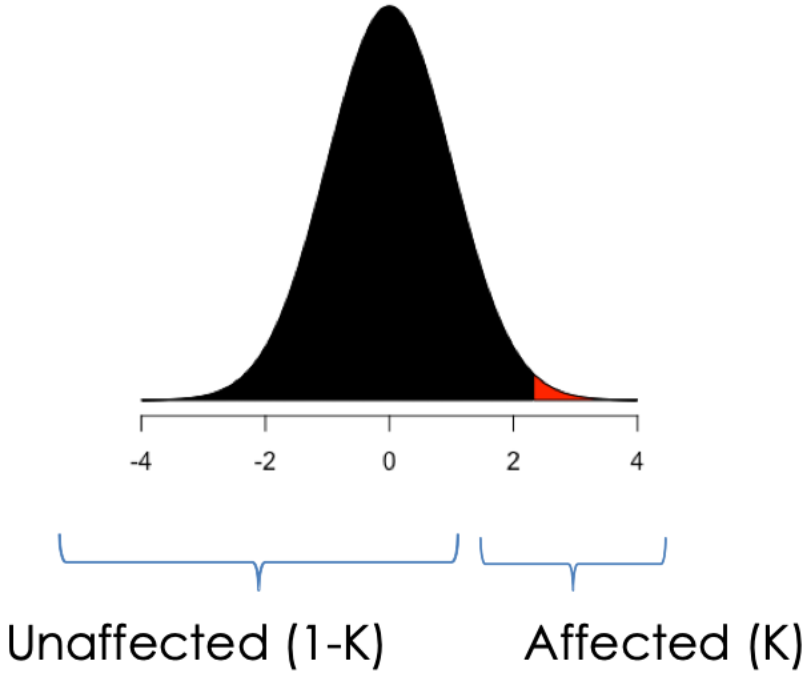
$$R_l^2 = R_o^2 \frac{K(1-K)}{z^2}$$

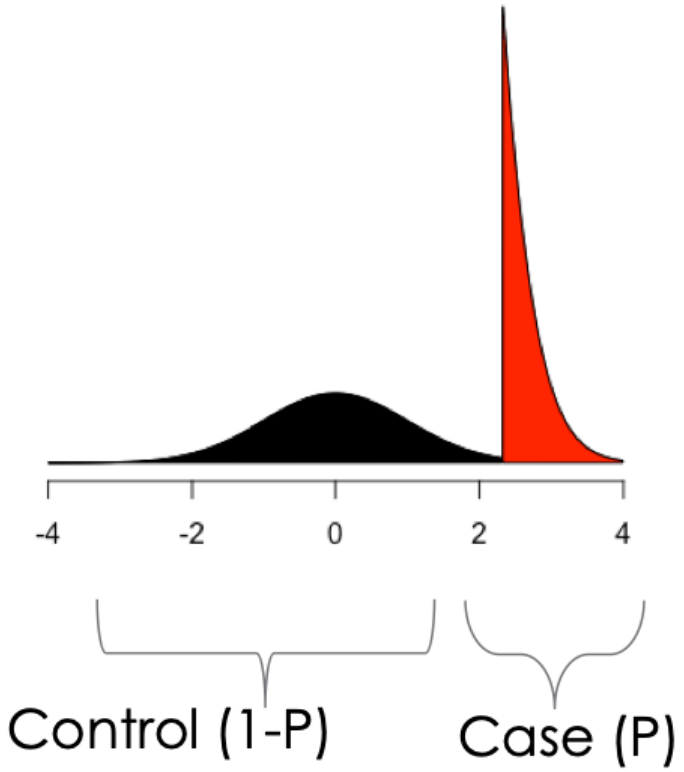


Lee, 2012, Genet Epidemiol

Prediction R^2 on liability scale ★

Ascertainment correction for case-control studies



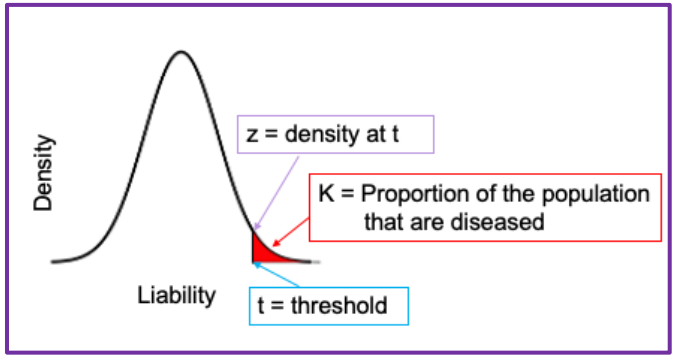


$$R_l^2 = R_o^2 \frac{K(1-K)}{z^2}$$

?

Prediction R^2 on liability scale ★

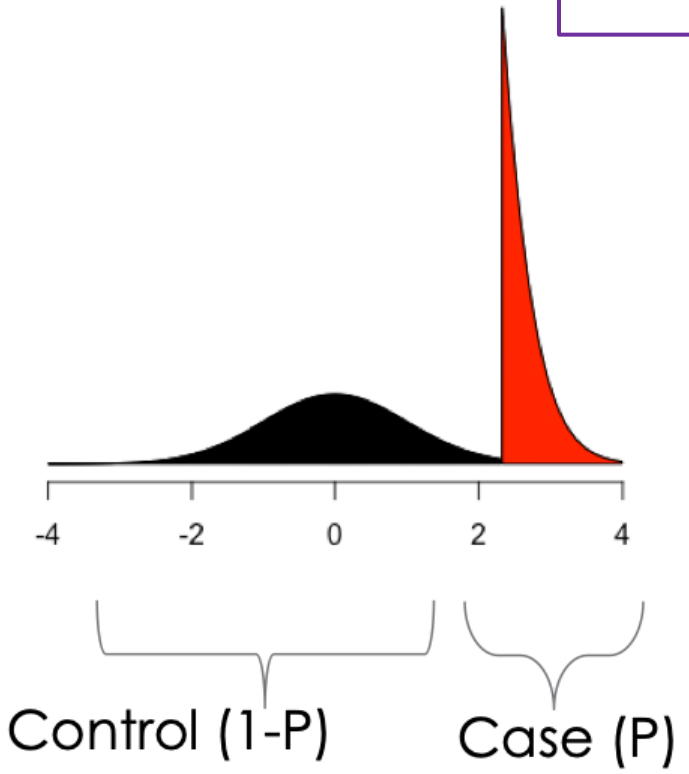
Ascertainment correction



$$R^2_{l\_cc} = \frac{R^2_{o\_cc} * C}{1 + R^2_{o\_cc} * \theta * C}$$

$$C = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}$$

$$\theta = \frac{z}{K}\left(\frac{P-K}{1-K}\right)\left(\frac{z}{K}\frac{P-K}{1-K} - \mathrm{t}\right)$$

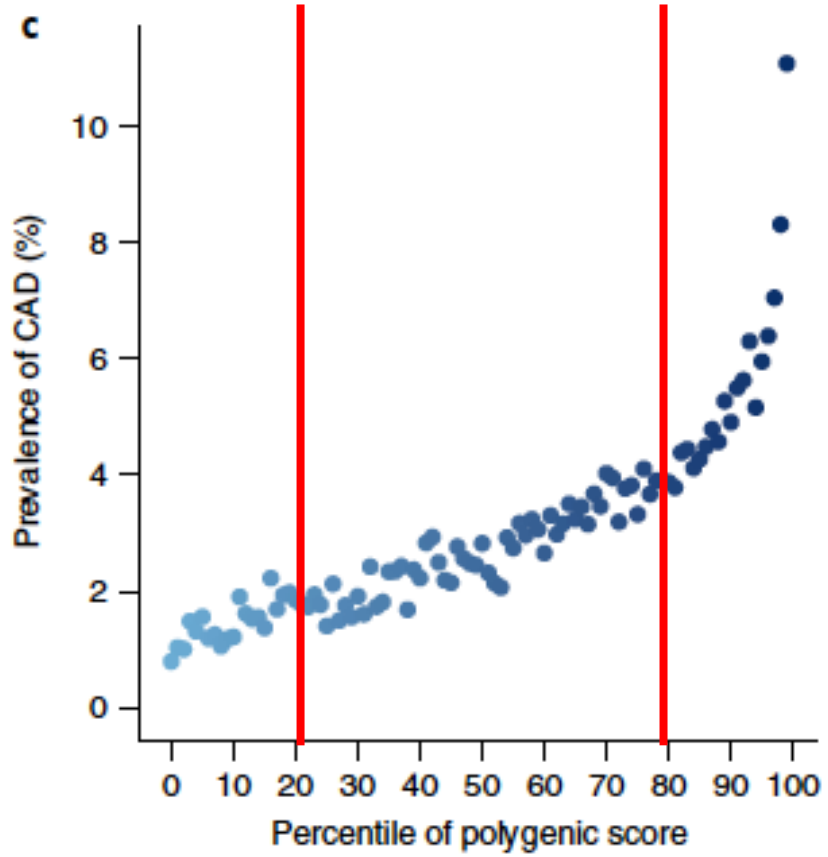


Lee, 2012, Genet Epidemiol

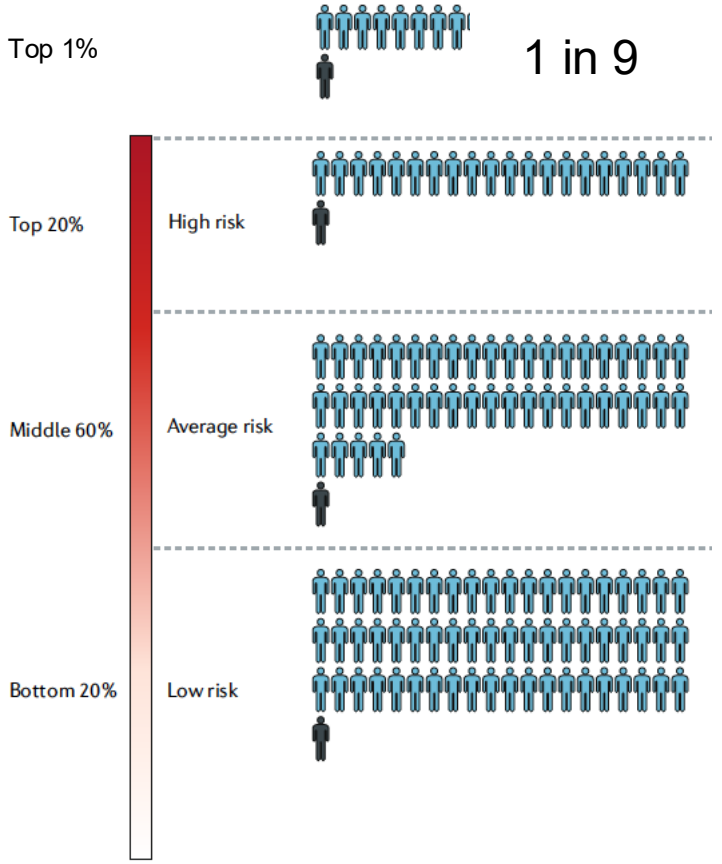
Property of R^2 on liability scale

- heritability is independent of disease prevalence
- $R^2_{l\_cc}$ is on the same sale as heritability estimated from family studies or genotypes
- Provide a direct measure of how well the predictor performs relative to capturing all genetic variation

Risk stratification

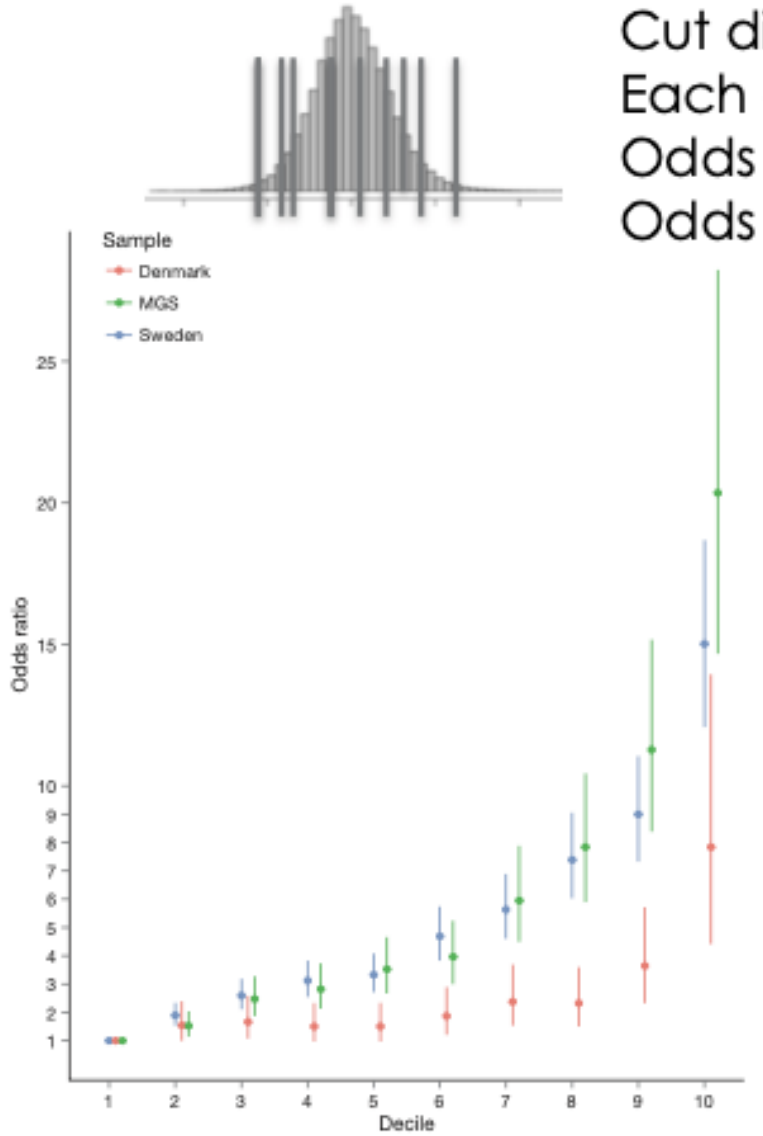


Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics



Torkamani et al, Nat Rev Genetics, 2018

Decile odds ratio



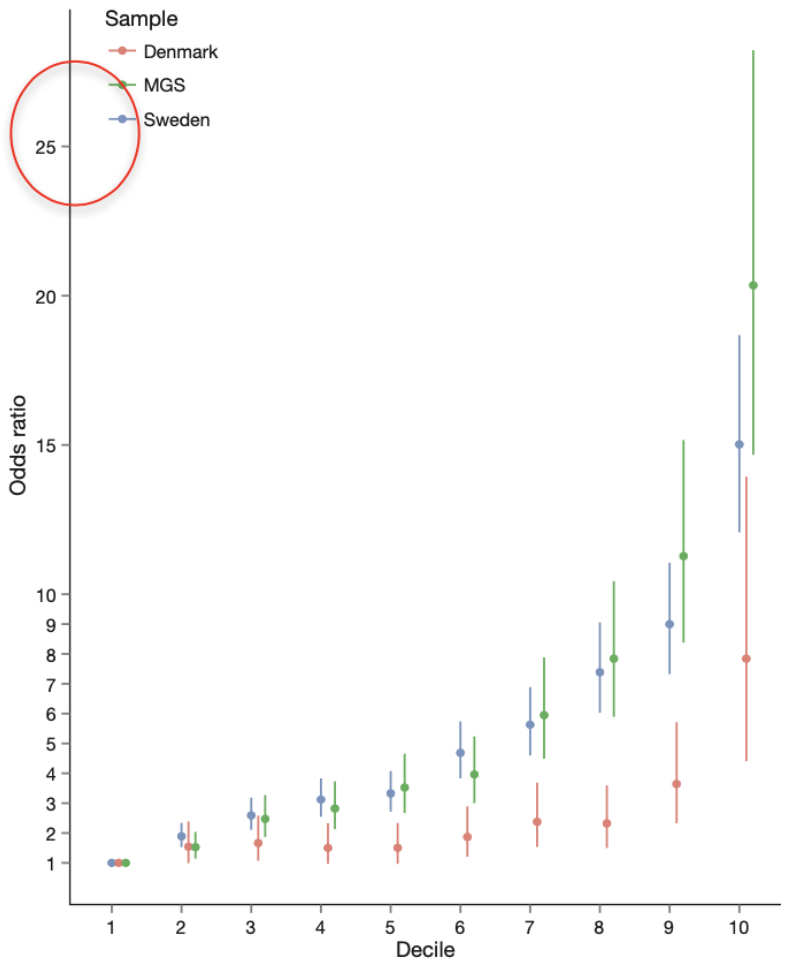
Cut distribution into deciles
Each decile will include both cases and controls
Odds of being a case in each decile
Odds ratio for each decile compared to the 1st decile

- Good visualisation
- Shows that there could be utility in using high vs low profile risk scores
- But remember case-control samples are 50% cases
- Would look less impressive if a population sample

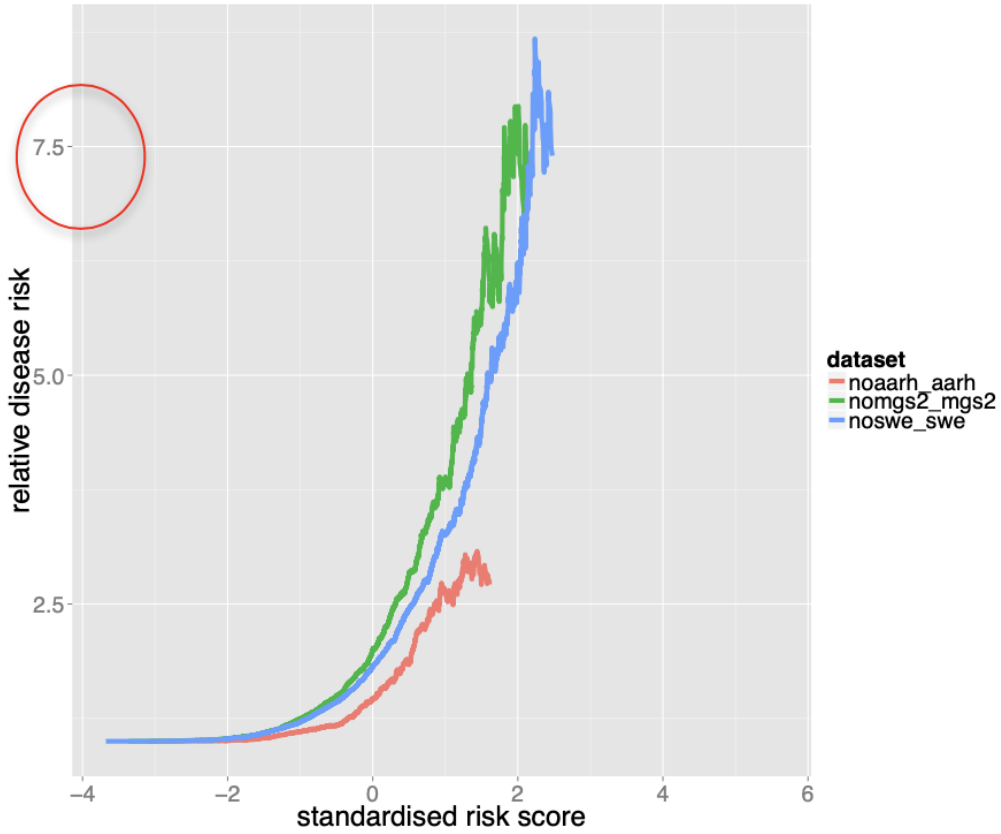
PGC-SCZ 2014 108 loci Nature

Decile odds ratio

In case control samples



Same data scaled to population risk



Decile odds ratio

Toy example:

	1st decile (Bottom 10%)	10th decile (Top 10%)
Case	23	83
Control	103	40

Odds being a case in 10th decile
= 83/40

Odds being a case in 1st decile
= 23/103

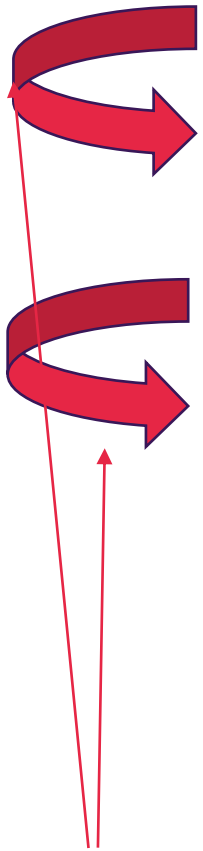
Odds ratio between 10th and 1st decile
= (23/103) / (83/40) =9.3

$$Odds\ ratio = \frac{Odds_1}{Odds_0} = \frac{P_1/_{1-P_1}}{P_0/_{1-P_0}}$$

$$Odds = \frac{P}{1-p}$$

P = probability of being case

Pitfalls of polygenic prediction



- **<u>Discovery</u>/Training/Derivation**
  - Estimate the effect sizes ($\hat{b}$) of SNPs on a trait (y) – GWAS
- **<u>Tunning</u>/Validation**
  - Further estimate some parameters (depends on methods; not all methods require it)
- **<u>Target</u>/Testing/Validation**
  - Build a polygenetic risk score (PRS) ($\hat{y}$):
  - Evaluate the prediction performance/accuracy

Should be independent; no overlap; out-of-sample prediction

GENOME-WIDE ASSOCIATION STUDIES — OPINION

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

Pitfalls of polygenic prediction

Pitfall 1: No target sample – report R^2 in discovery sample

x: M markers for N samples

y from N(0,1) independently (null hypothesis)

1) Multiple linear regression of y on x (when M<N)

$E(R^2) = M/N$ variation "explained" by chance

2) Select m "best" markers out of M in total, and conduct multiple linear regression in the same dataset

$E(R^2) \gg m/N$ winner's curse

Pitfalls of polygenic prediction

Pitfall 2: target sample overlapped with discovery sample

- Overlapping target and discovery sample
- Greater similarity between target and discovery sample (such as relatedness)
  - Cross-validation: not a pitfall, but to be aware

$$\mathrm{cov}(\hat{y}_i, y_i) = \mathrm{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij} b_j + e_i\}$$

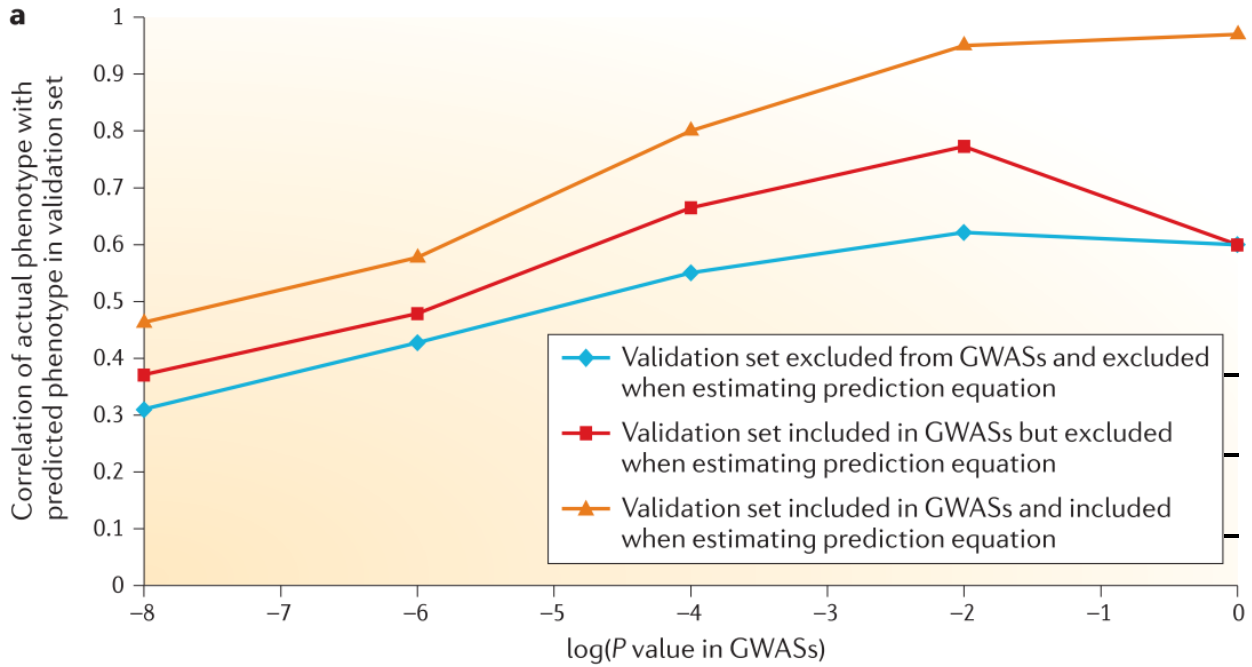
$$= \sum_{j=1}^{m} \mathrm{var}(x_{ij})\hat{b}_j b_j + \sum_{j=1}^{m} x_{ij}\,\mathrm{cov}(\hat{b}_j, e_i)$$

If b estimated from the same data in which prediction is made, then the second term is non-zero

Pitfalls of polygenic prediction

Pitfall 3: Less obvious non-independence

- Estimate SNP effects and/or select SNPs from total sample (discovery + target sample)
- Re-estimate effects in the target sample after selecting in the discovery sample



- Out-of-sample prediction
- Select SNPs in total sample
- Direct report R2 in the discovery sample

Summary

- Assessing prediction accuracy and bias is straightforward for quantitative traits
- For disease traits
  - Pseudo-R^2 is subject to sample prevalence (P)
  - AUC is widely used but max AUC depends on population prevalence (K)
  - R^2 on the liability scale is independent of both P and K (nice property!)
  - Decile odds ratio is great for visualization but not portable across samples
- Pitfalls in the prediction analysis – non-independence of discovery & target samples

Recommended reading

1. Wray NR, *et al*. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet*. 2010 Feb 26;6(2):e1000864. (**Interpretation of ROC curve**)

2. FALCONER DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*. 1965. 29: 51-76. (**Theory of liability model**)

3. Lee SH, *et al*. A Better Coefficient of Determination for Genetic Profile Analysis. *Genet. Epidemiol*. 2012. 36: 214-224. (**Methodology for R^2 on liability scale**)

4. Wray NR, *et al*. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet*. 14, 507–15 (2013). (**Pitfalls of interpretation**)