

Evaluation of PRS for diseases

Statistics and Visualisation

Jian Zeng

j.zeng@uq.edu.au

Prediction accuracy

Squared correlation between phenotype and PGS in the validation sample

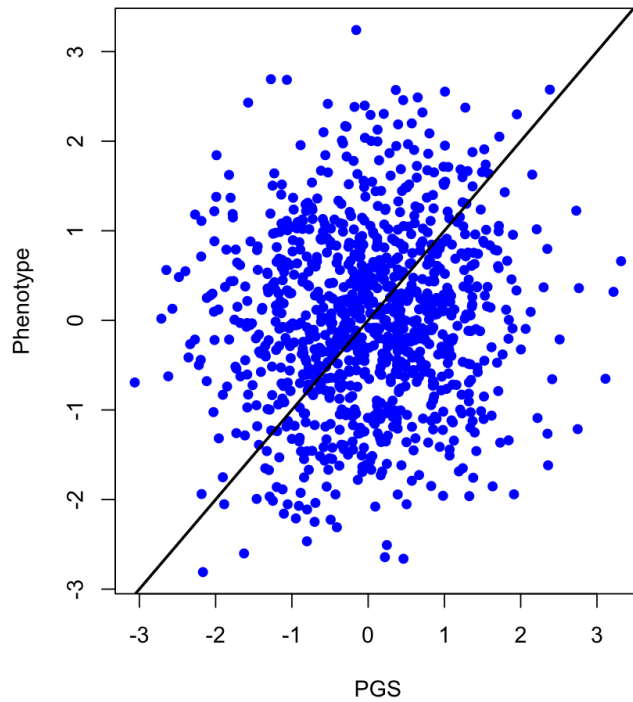
- The proportion of phenotypic variance explained by PGS (prediction R^2)
- The SNP-based heritability is its upper bound

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

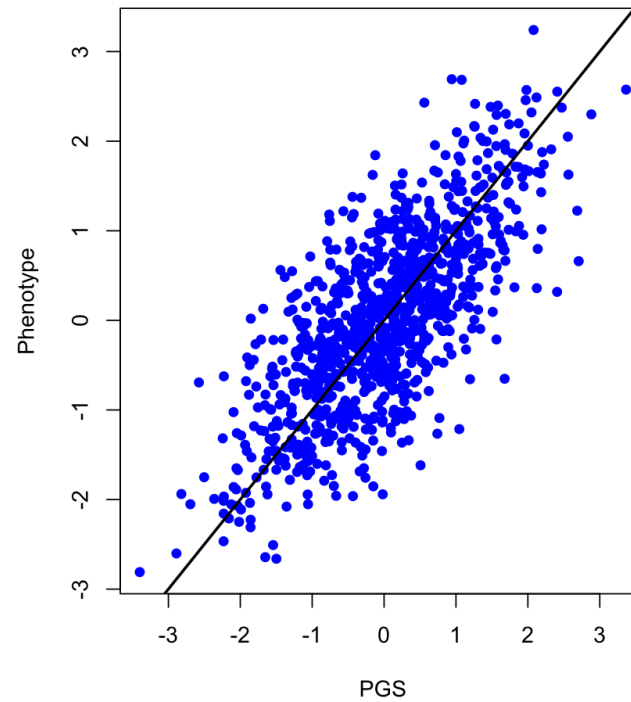
- Null model: $y = \text{covariates} + e$
- Full model: $y = \text{covariates} + \text{PGS} + e$
- Incremental R^2 : $R_{Full}^2 - R_{Null}^2$

Prediction accuracy

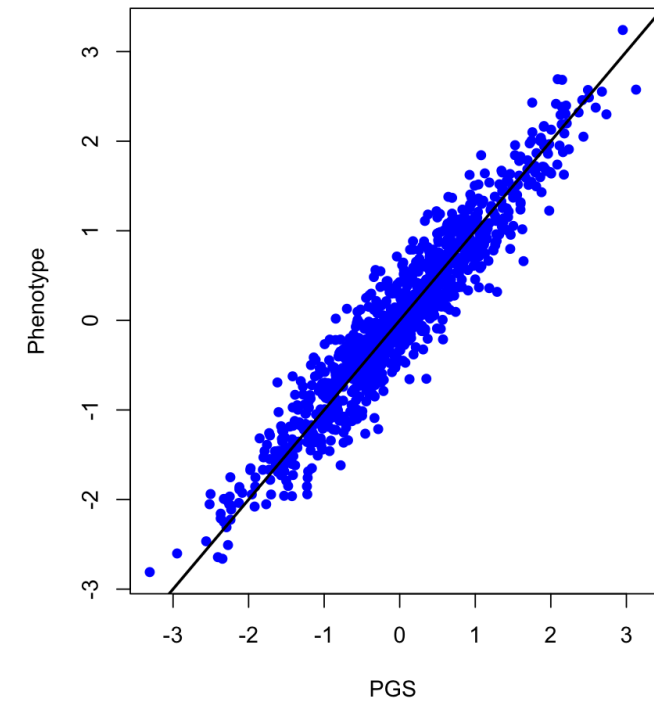
$R^2 = 0$



$R^2 = 0.5$



$R^2 = 0.9$



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

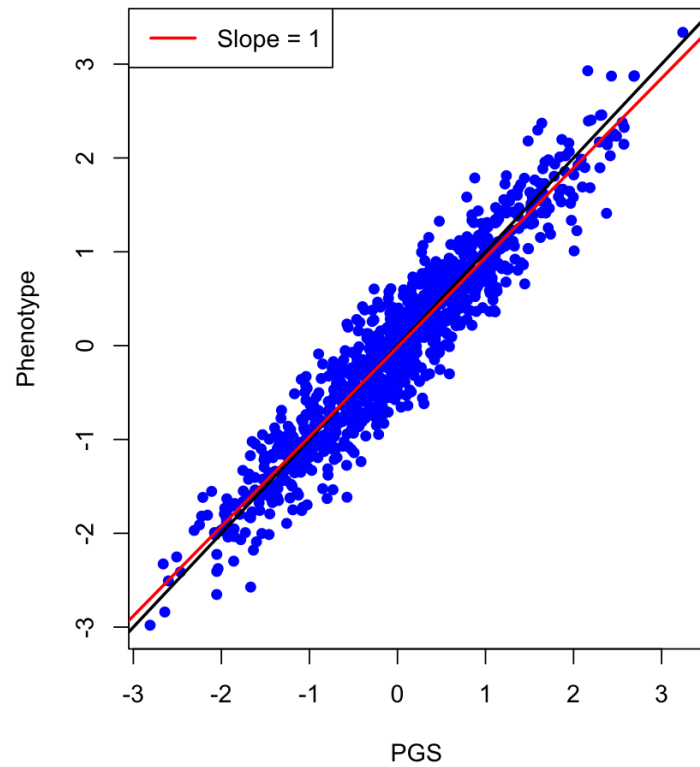
- 1 unit increase in PGS leads to >1 unit increase in phenotype
- The PGS are downward biased

If the slope < 1 , then

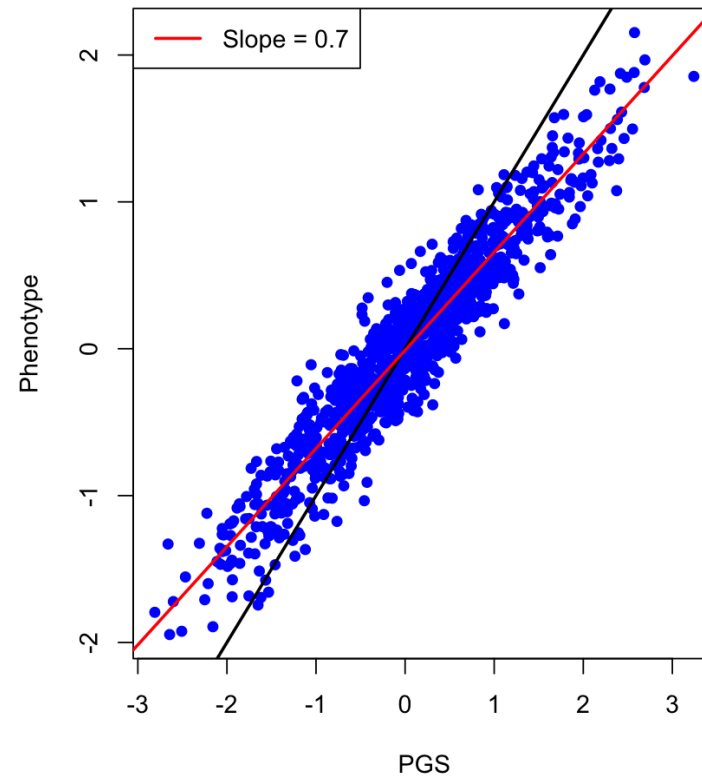
- The PGS are upward biased

Prediction bias

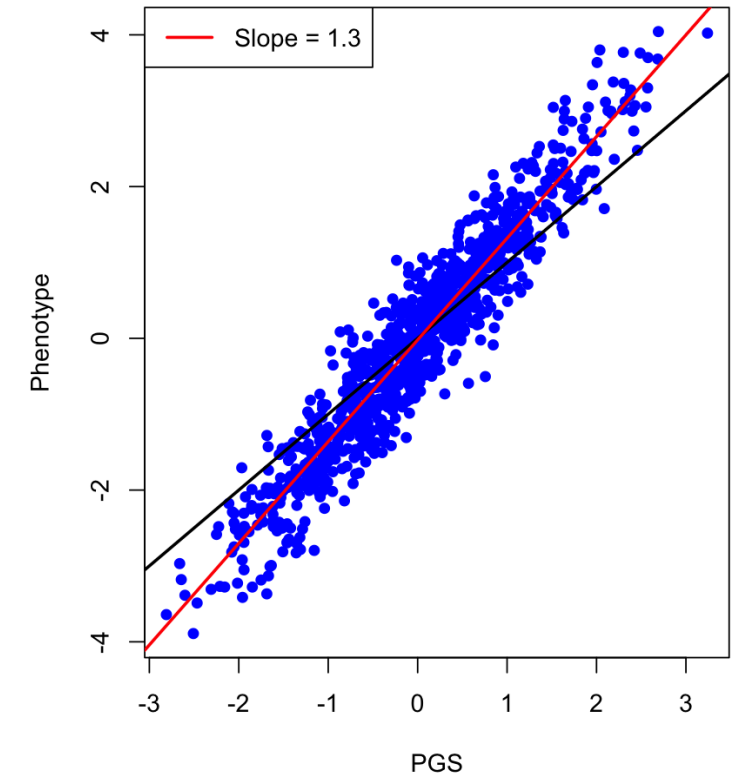
Unbiased 😊



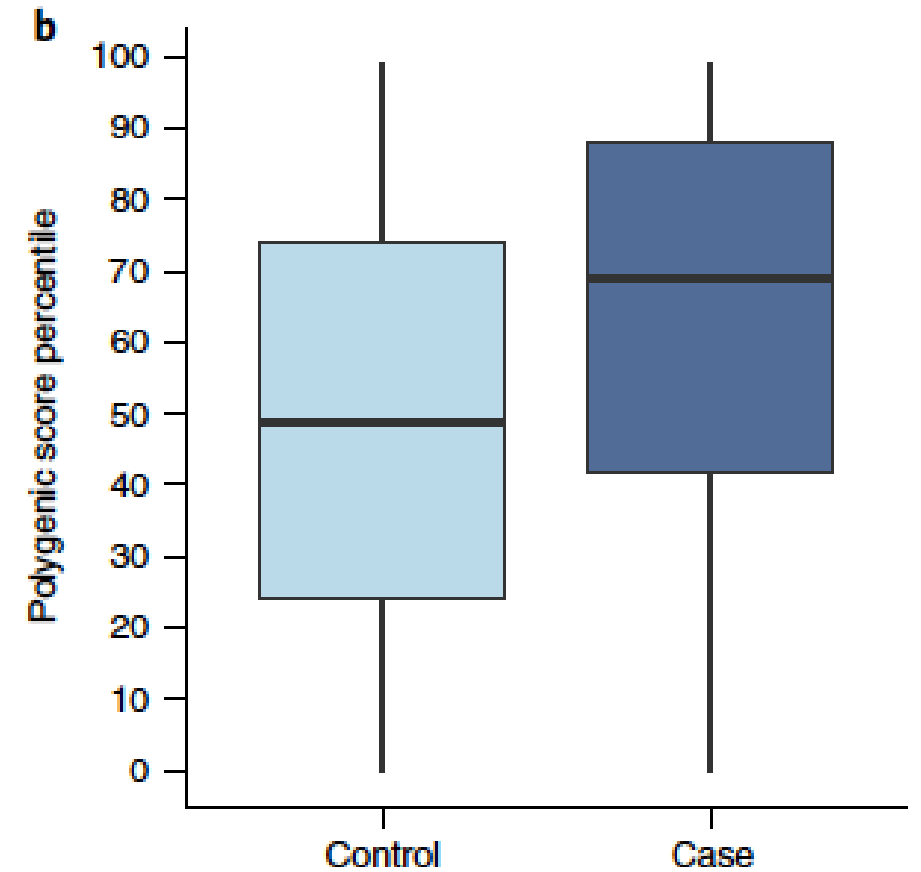
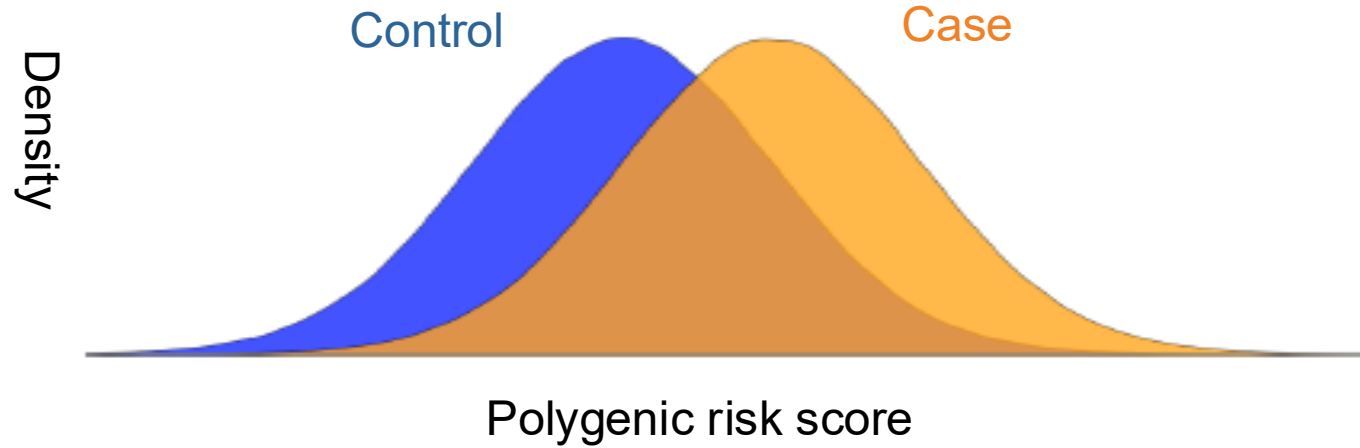
Upward biased 😞



Downward biased 😞



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
- Decile odds ratio (OR)
- Risk stratification

Logistic regression:

- Null model: $y = \text{logistic}(\text{covariates} + e)$
- Full model: $y = \text{logistic}(\text{covariates} + \text{PGS} + e)$

Many pseudo R^2 statistics available for logistic regression

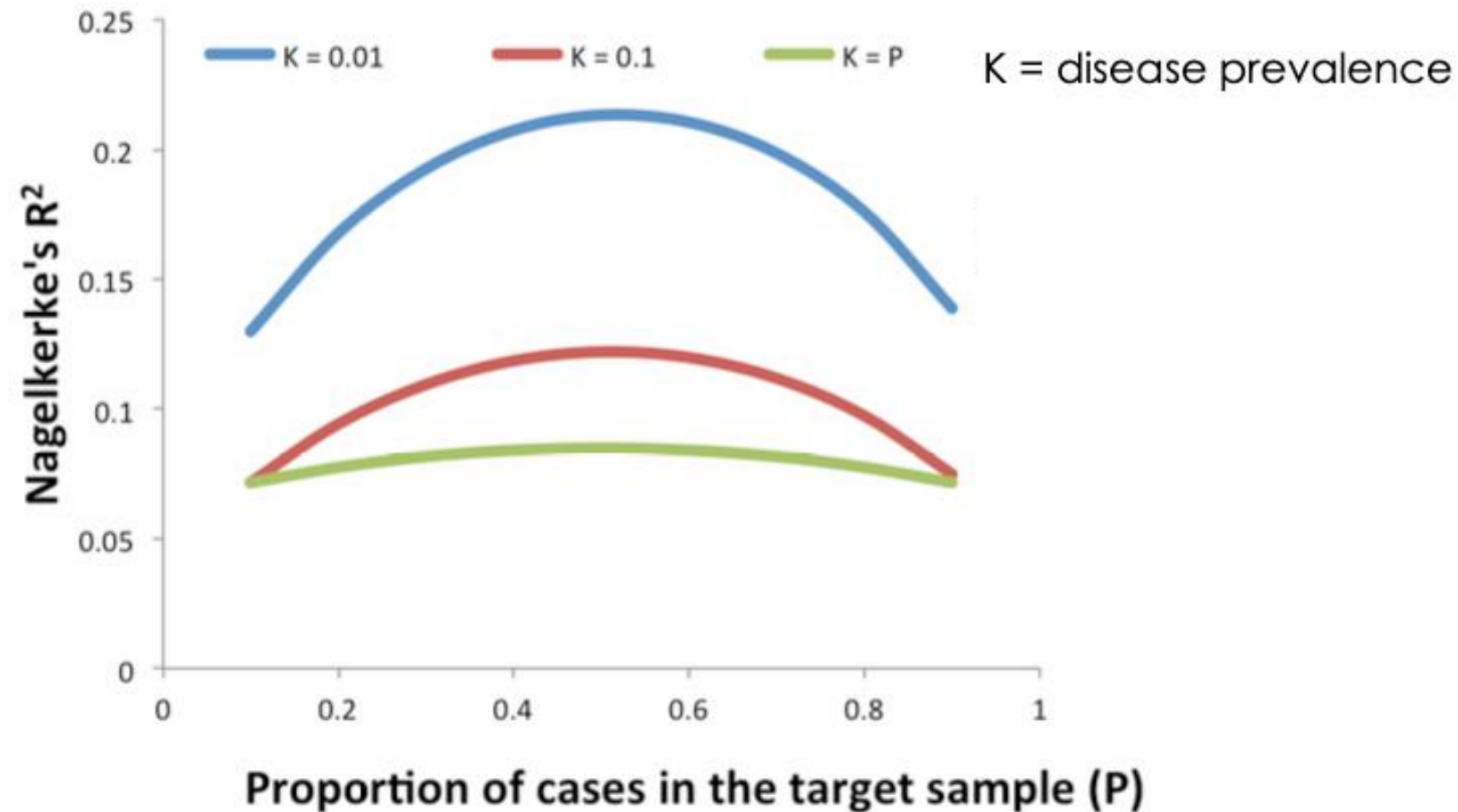
e.g., Nagelkerke's R^2

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

For a review of pseudo R^2 statistics, check [this link](#)

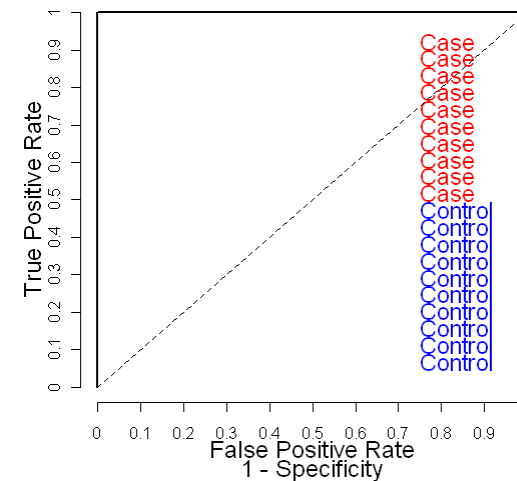
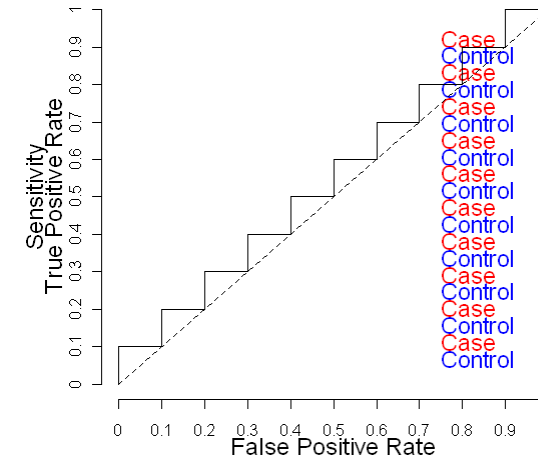
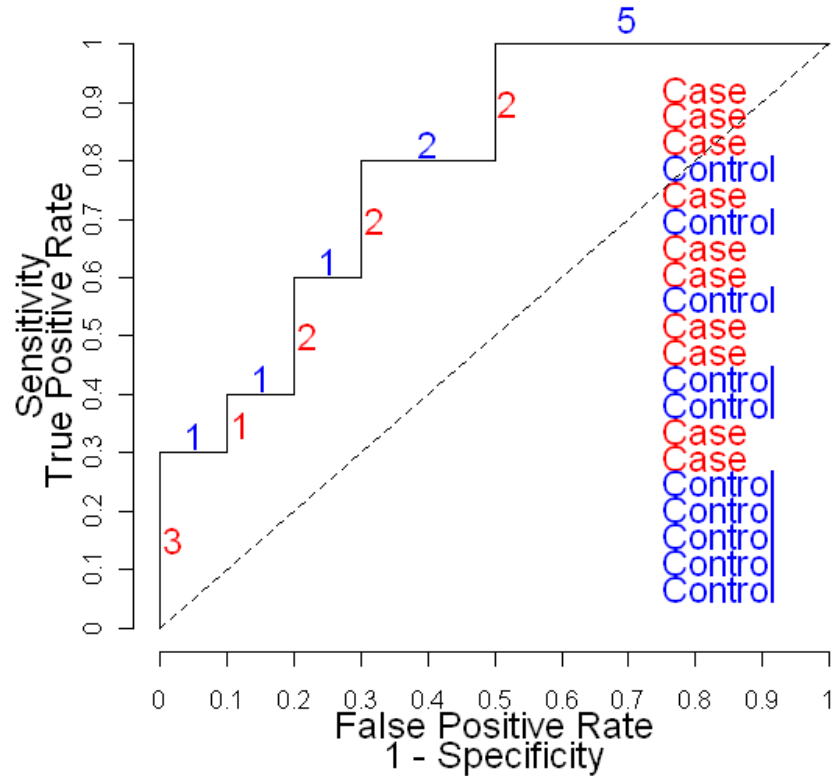
Property of pseudo R^2

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



- 😊 - 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

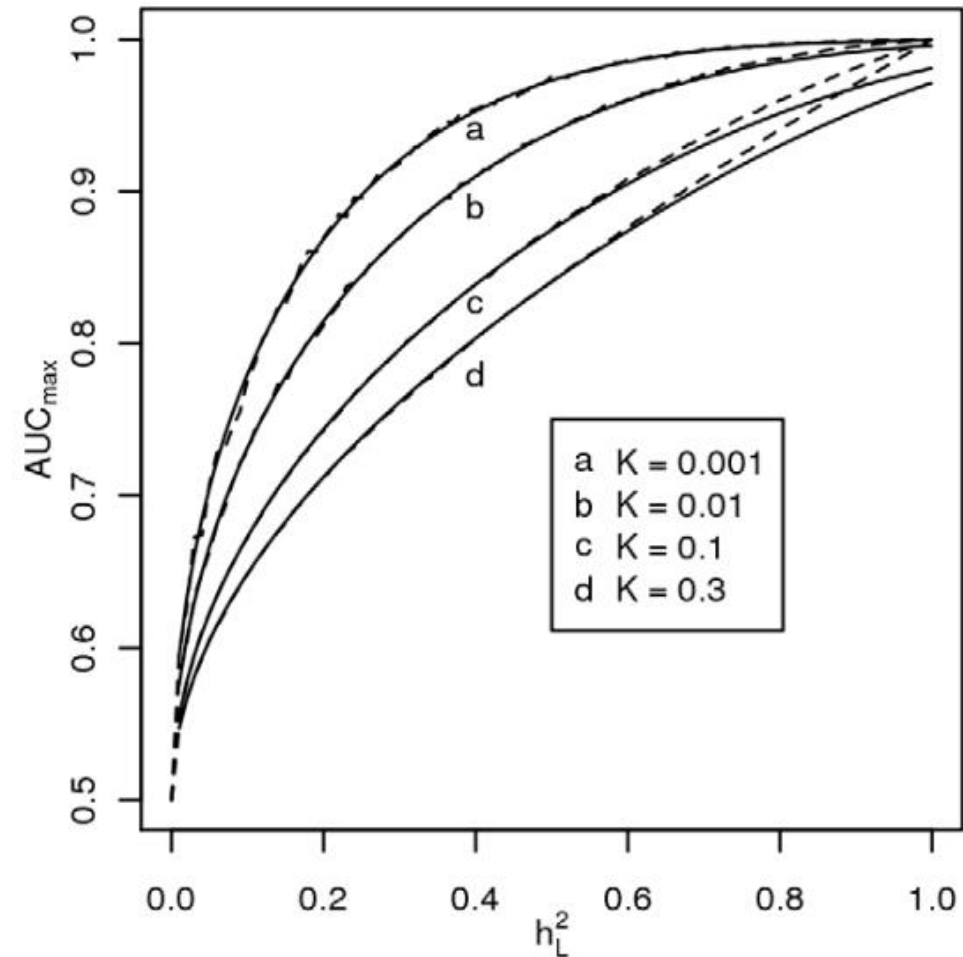


Figure 2. Relationship between maximum AUC (AUC_{max}) from a genomic profile and heritability on the liability scale h_L^2 . For

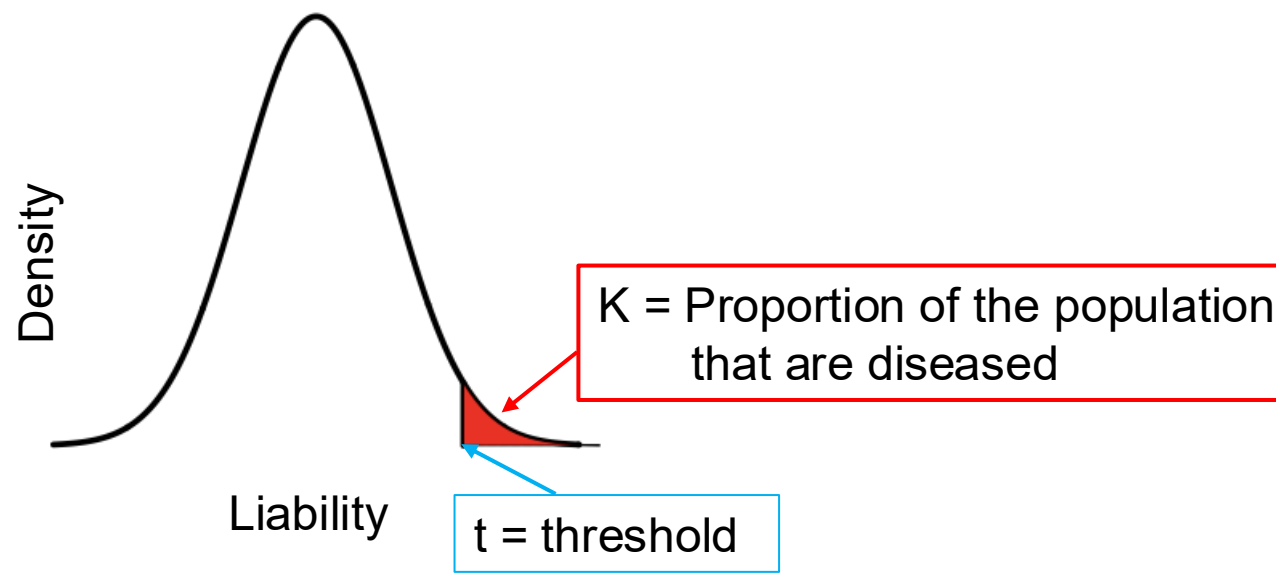
The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling

Naomi R. Wray^{1*}, Jian Yang¹, Michael E. Goddard^{2,3}, Peter M. Visscher¹

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

Liability threshold model

Map variance explained on observed probability 0-1 scale (R_o^2)
To underlying unobserved continuous liability scale (R_l^2).



Prediction R^2 on liability scale

Linear regression; Y are 0s and 1s

Null: $Y = \text{covariates} + e$

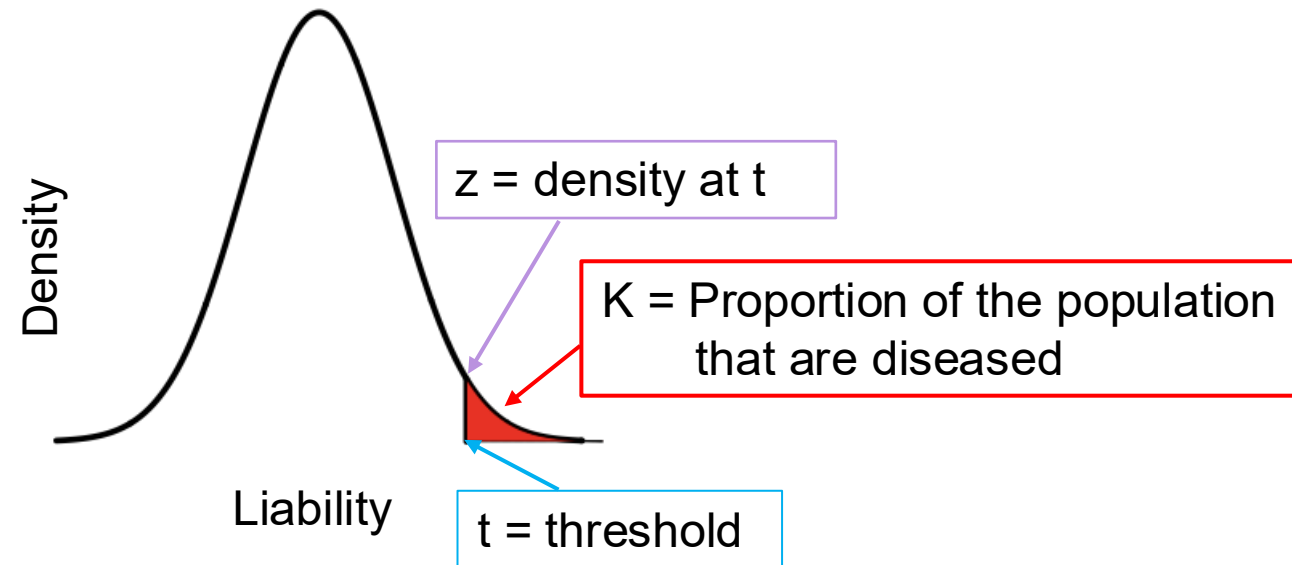
Full: $Y = \text{covariates} + \text{PGS} + e$

R^2 on the observed scale

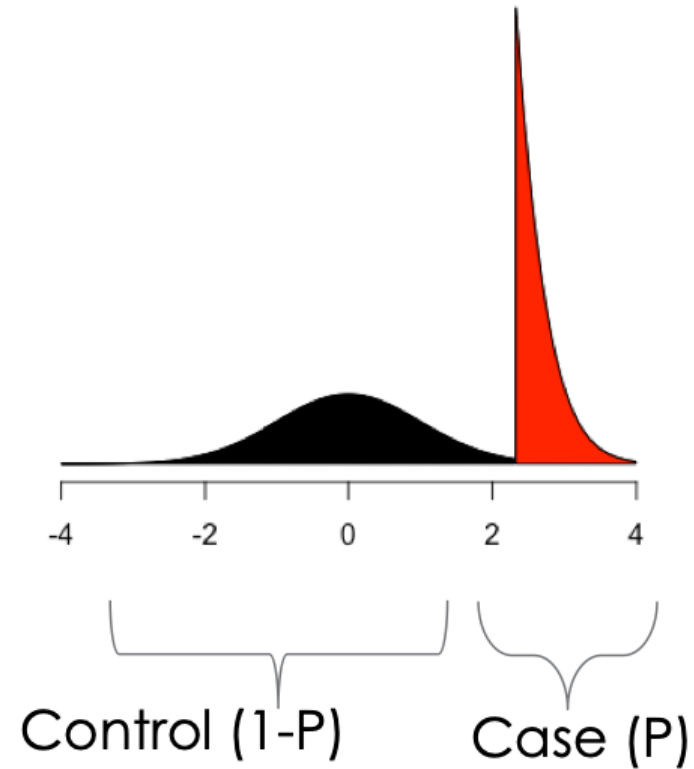
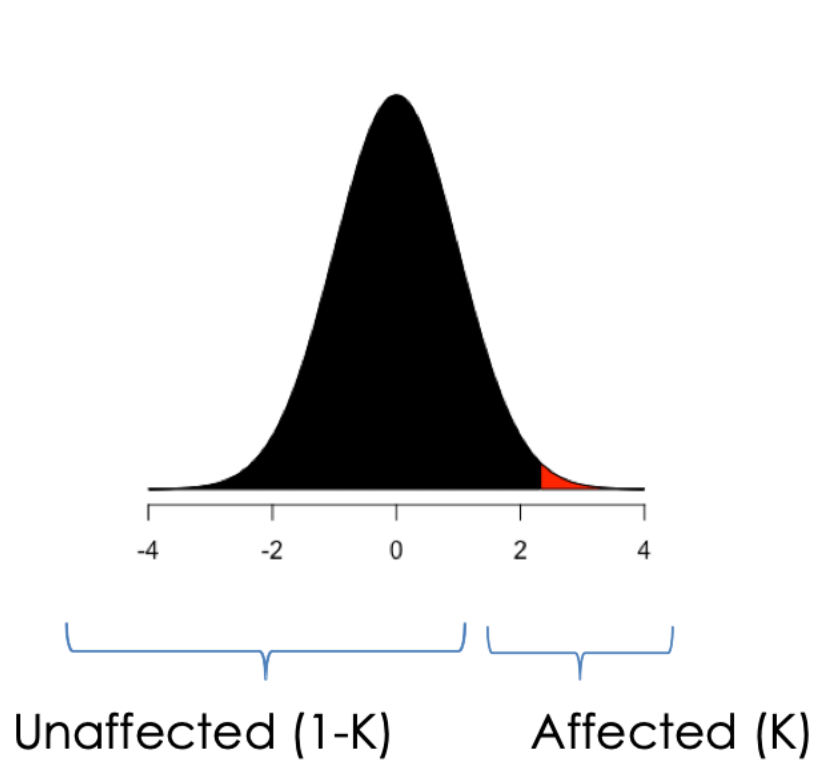
$$R_o^2 = 1 - \left(\frac{\text{Likelihood}_{null}}{\text{Likelihood}_{full}} \right)^{2/N}$$

R^2 on the liability scale

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



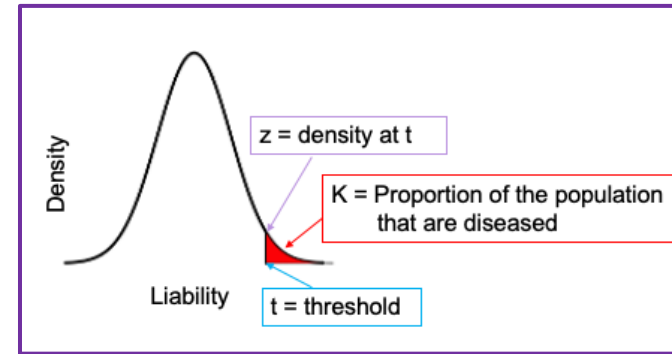
Ascertainment in case-control studies



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



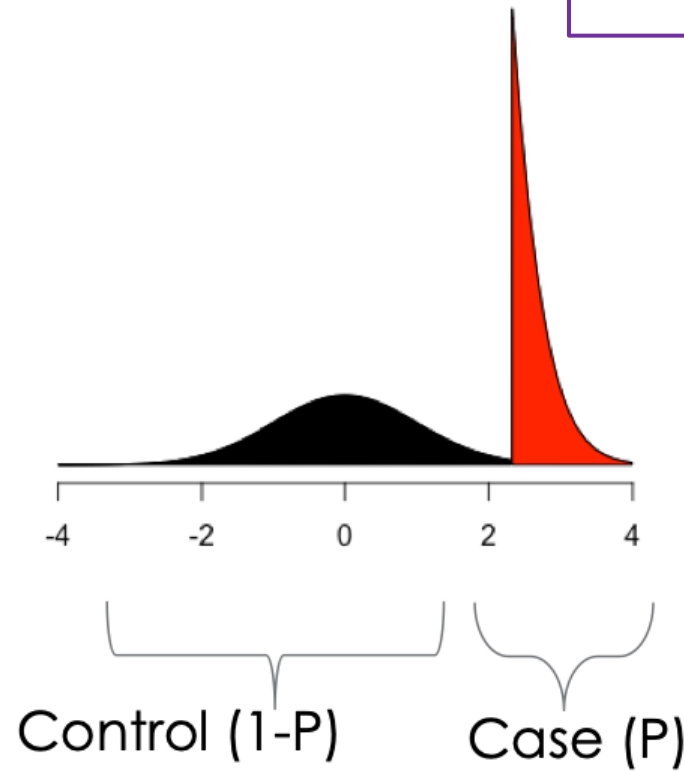
Ascertainment in case-control studies



$$R_{l_{cc}}^2 = \frac{R_{o_{cc}}^2 * C}{1 + R_{o_{cc}}^2 * \theta * C}$$

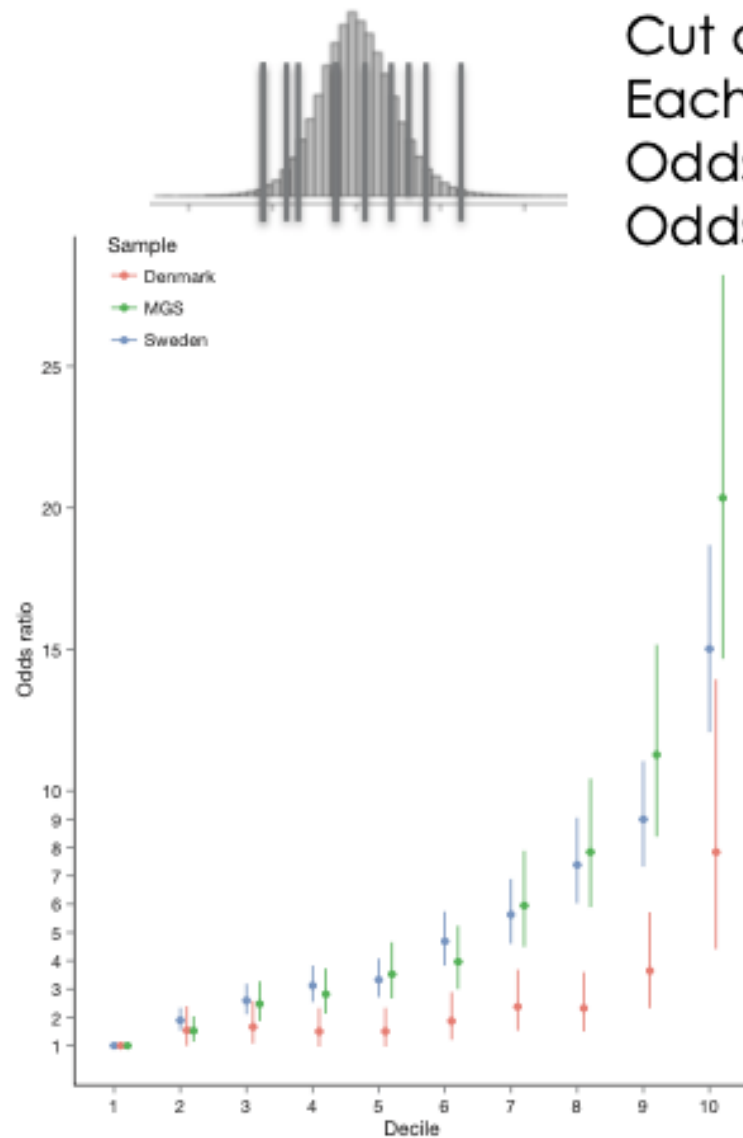
$$C = \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

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



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





Cut distribution into deciles

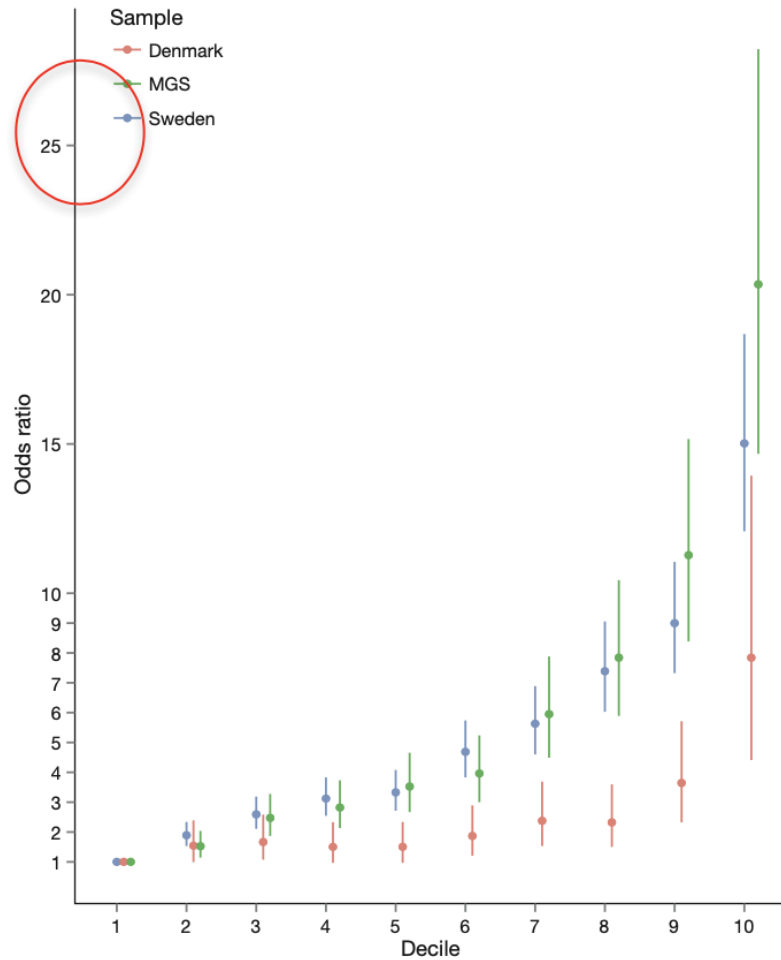
Each decile will include both cases and controls

Odds of being a case in each decile

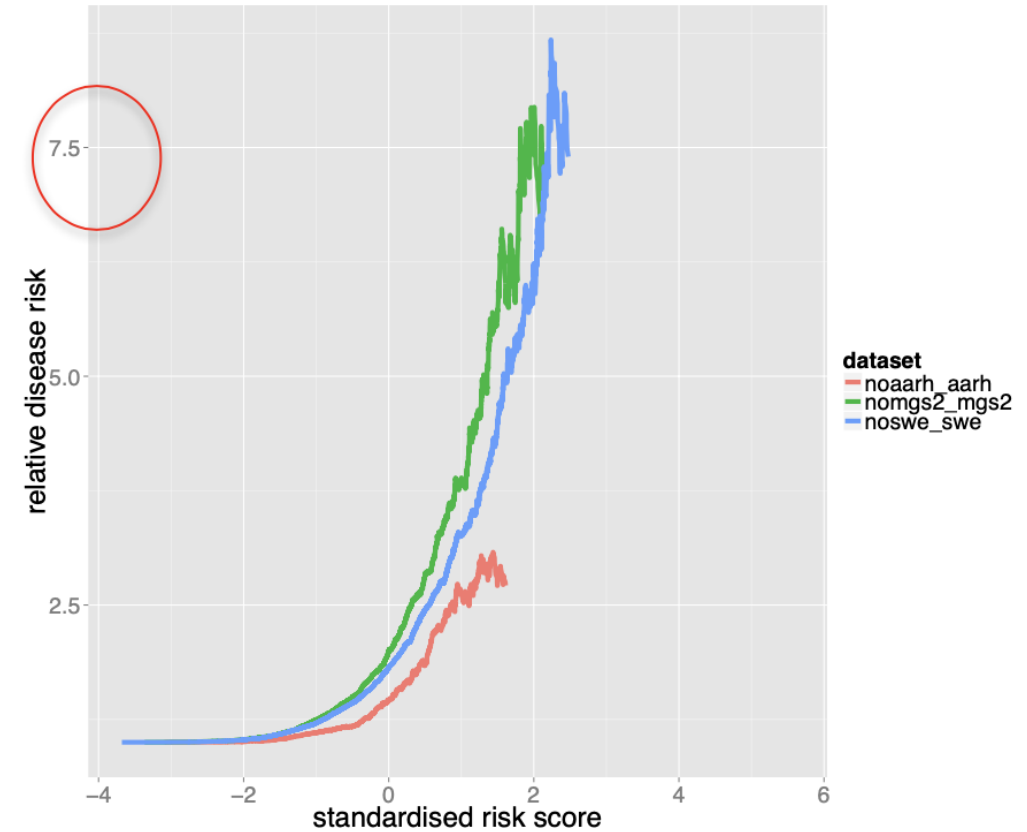
Odds ratio for each decile compared to the 1st decile

- 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

In case control samples



Same data scaled to population risk



Toy example:

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

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

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

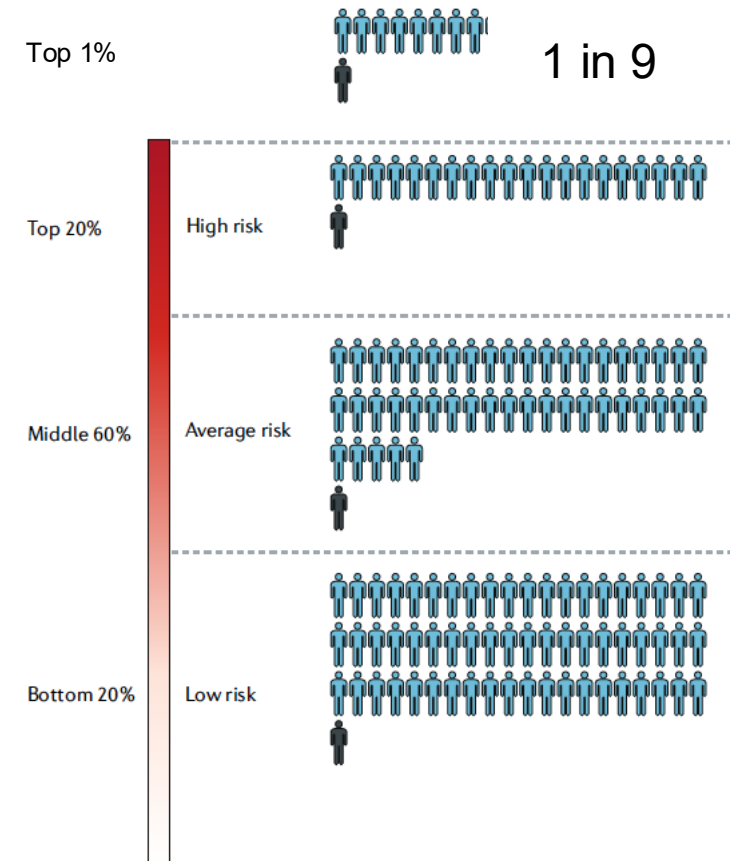
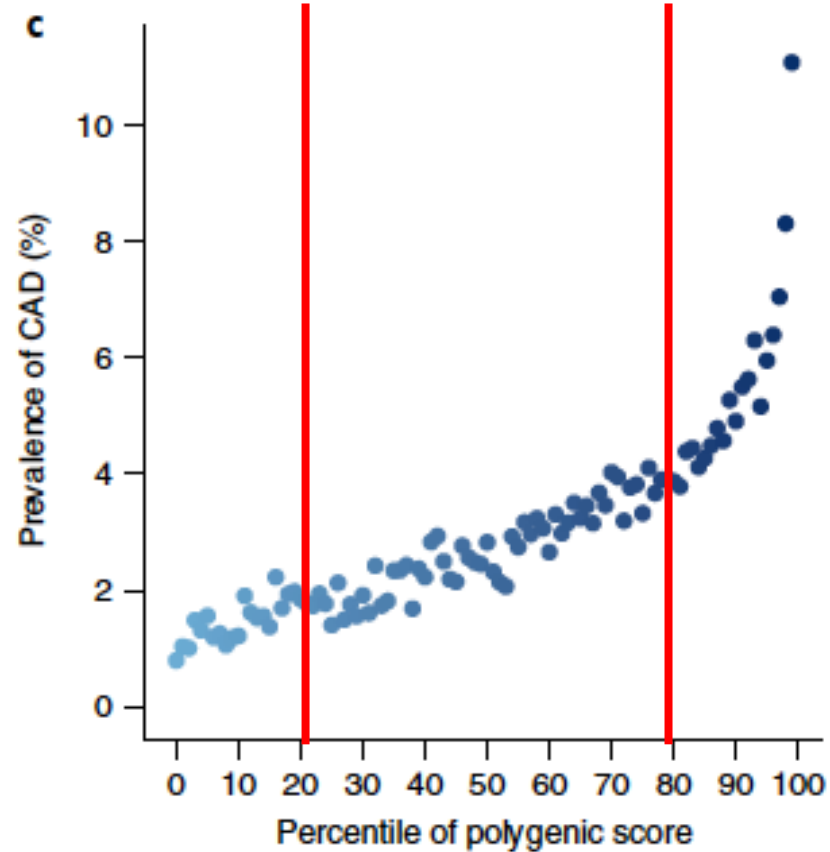
Odds ratio between 10th and 1st decile
= (83/40) / (23/103) = 9.3

$$\text{Odds ratio} = \frac{\text{Odds}_1}{\text{Odds}_0} = \frac{P_1/1-P_1}{P_0/1-P_0}$$

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

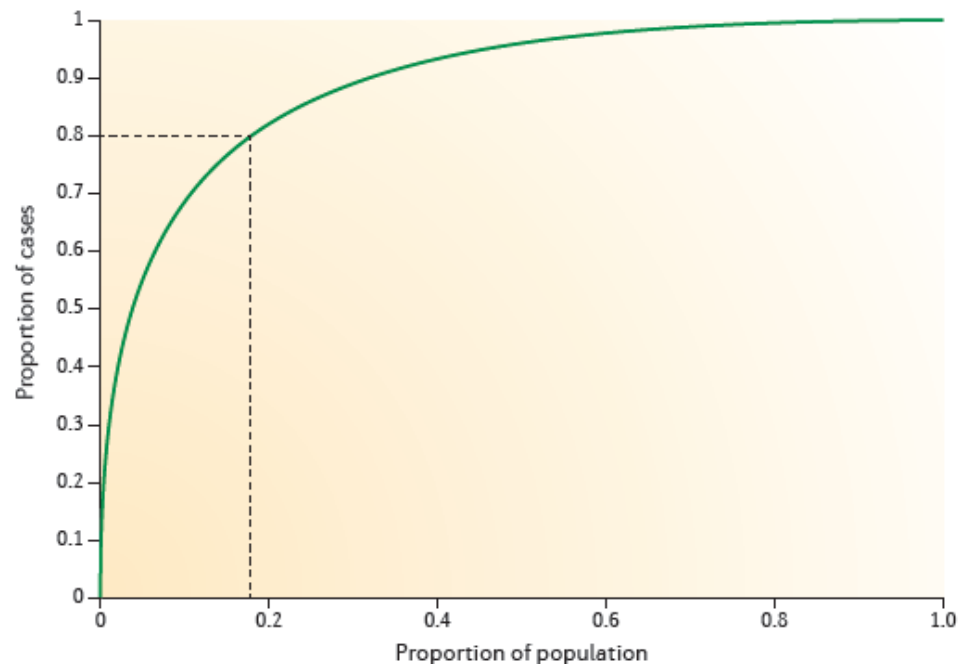
P = probability of being case

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



Population risk of 1%

80% of cases in
top 18% of genetic risk

For every 1,000 people treated with intervention could “save” 10
Treat only 18% = 180 and “save” 8

Number of people treated to save 1 reduced from 100 to 22.5

Polychronakos & Li NRG (2011) Understanding Type I Diabetes through genetics. Nat Rev Genetics

- 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

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**)

Questions?

5 min break



Practical 5: Evaluation of PRS prediction

https://cnsgenomics.com/data/teaching/GNGWS26/module5/Practical5_Evaluation.html

To log into your server, type command below in **Terminal** for Mac/Linux users or in **Command Prompt** or **PowerShell** for Windows users.

```
ssh username@hostname
```

And then key in the provided password.