THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Evaluation of Prediction Performance

Measurement, Visualisation, Theory & Pitfalls

Jian Zeng

j.zeng@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Slides credit: Naomi Wray, Huanwei Wang

Outline

- How to measure PGS prediction in quantitative traits?
- How to measure PGS prediction in diseases?
- What parameters determine the accuracy of PGS prediction?
- What are the pitfalls in the prediction analysis?

PGS evaluation in quantitative traits

Prediction accuracy

Squared correlation between phenotype and PGS in the validation sample

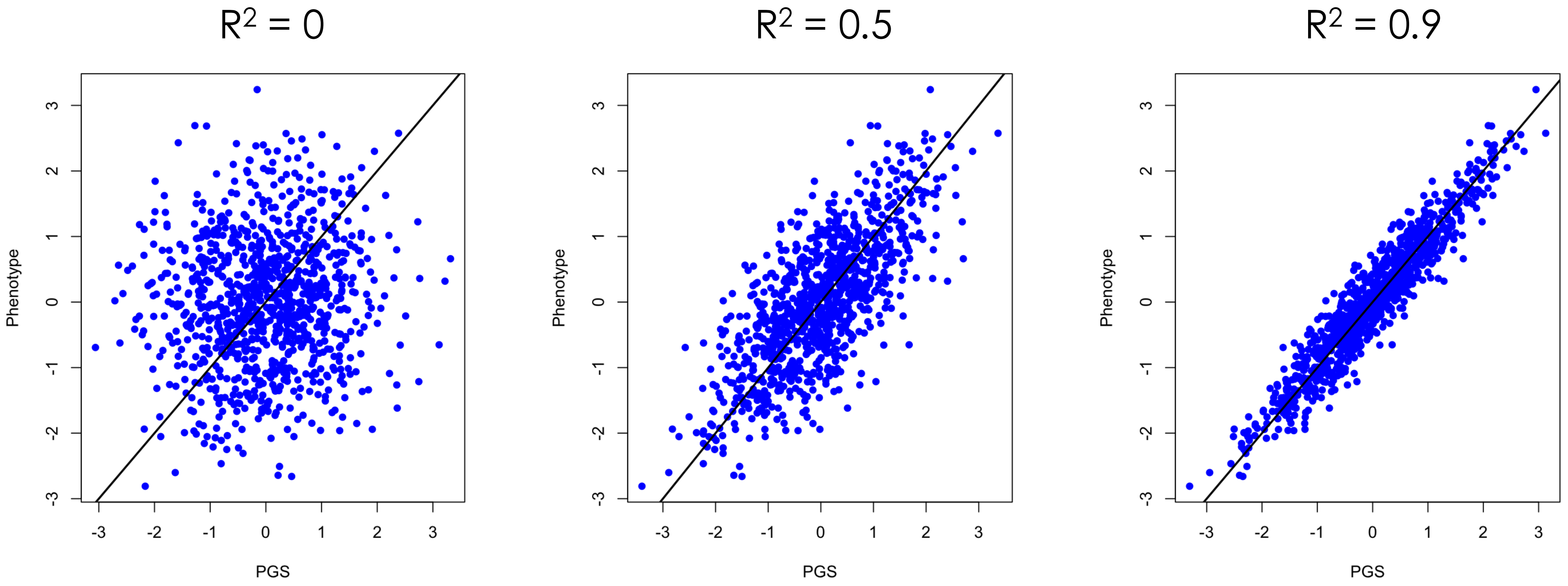
- The proportion of phenotypic variance explained by PGS (prediction R^2)
- The SNP-based heritability is its upper bound

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

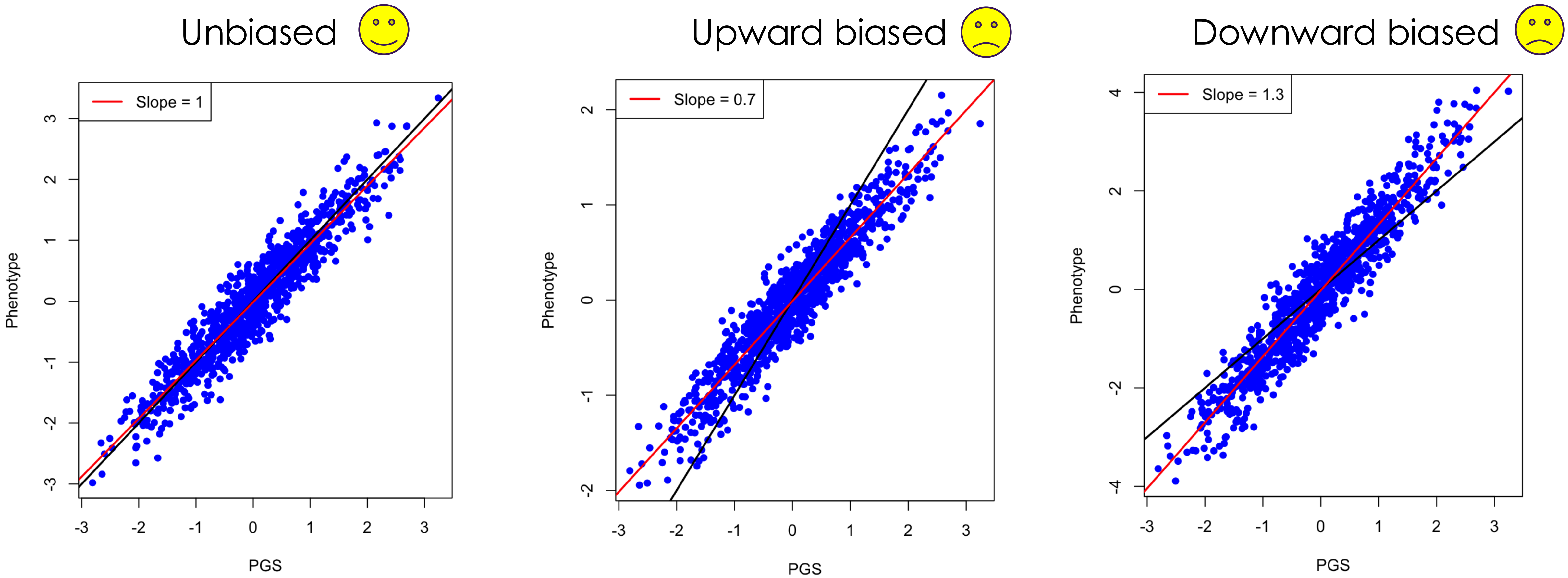
- 1 unit increase in PGS leads to >1 unit increase in phenotype
- The PGS are downward biased

If the slope < 1, then

- The PGS are upward biased

PGS evaluation in quantitative traits

Prediction bias

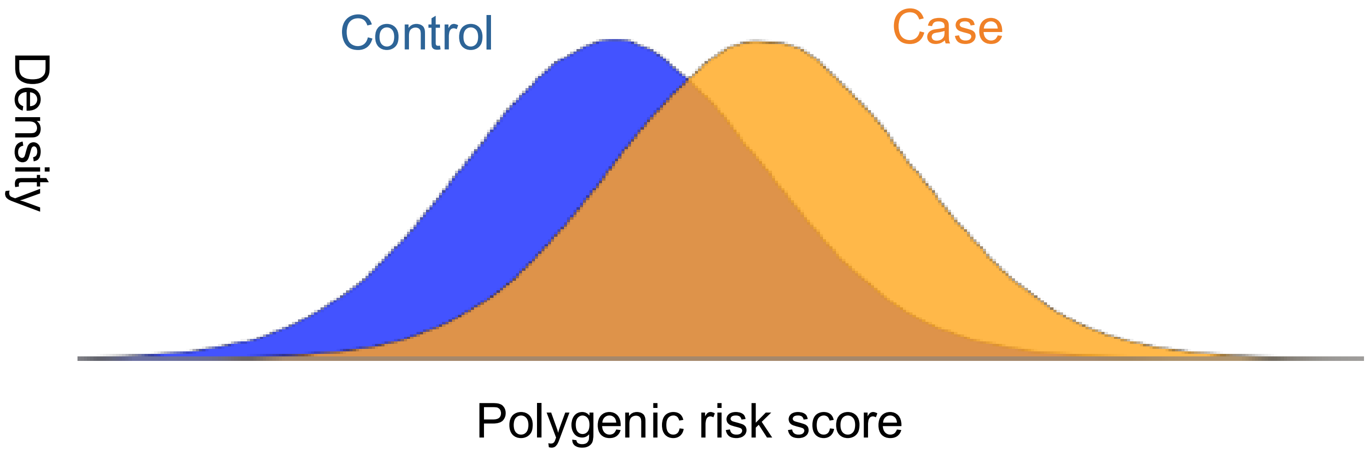


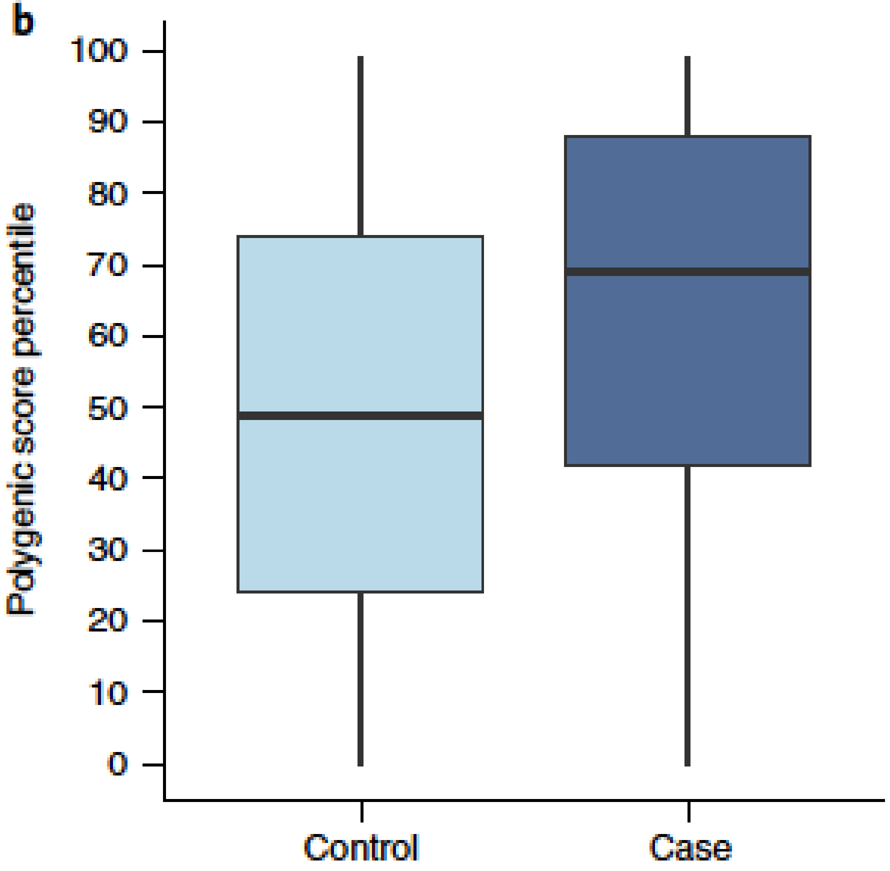
Unbiased

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
- Decile odds ratio (OR)
- Risk stratification

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

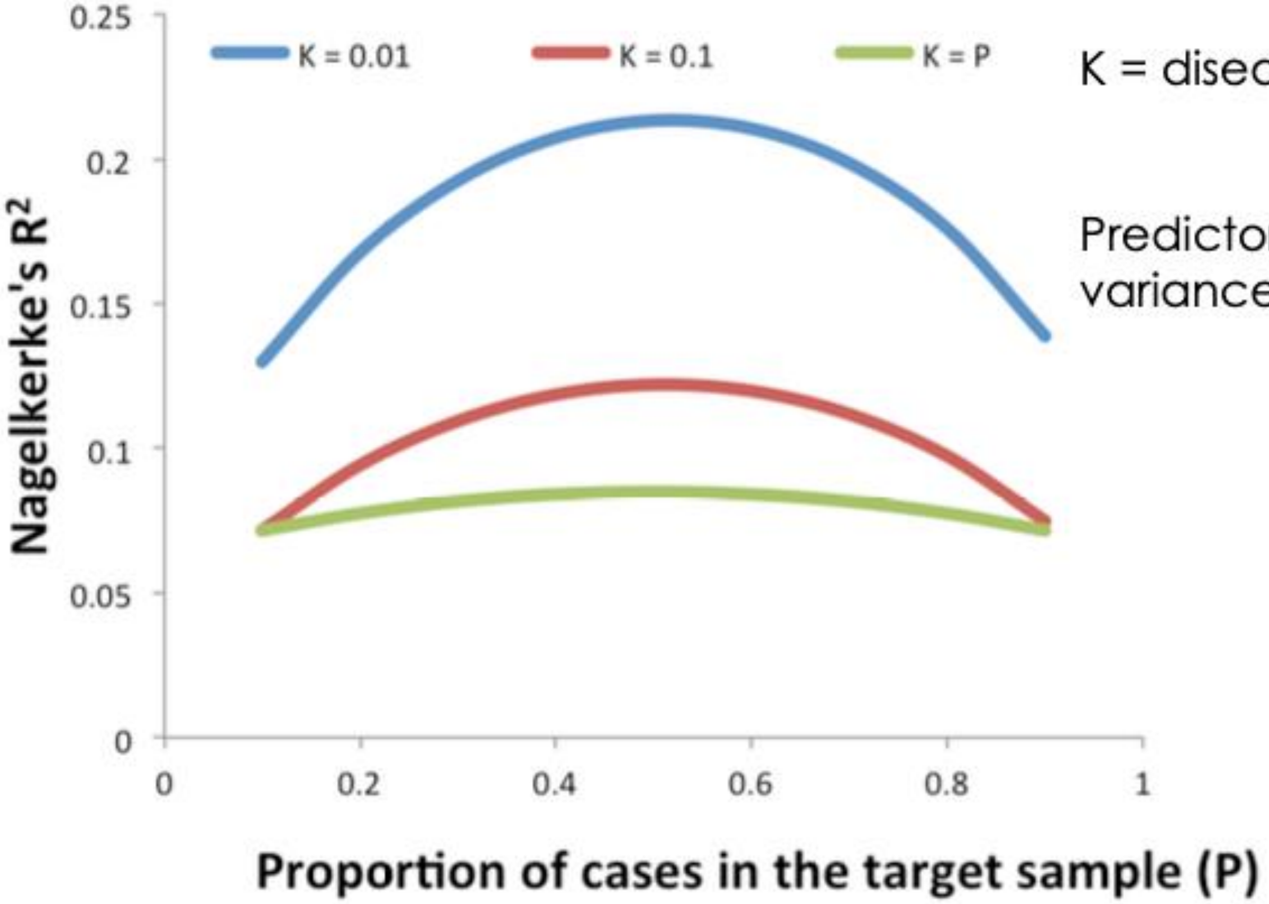
e.g., Nagelkerke’s R^2

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

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

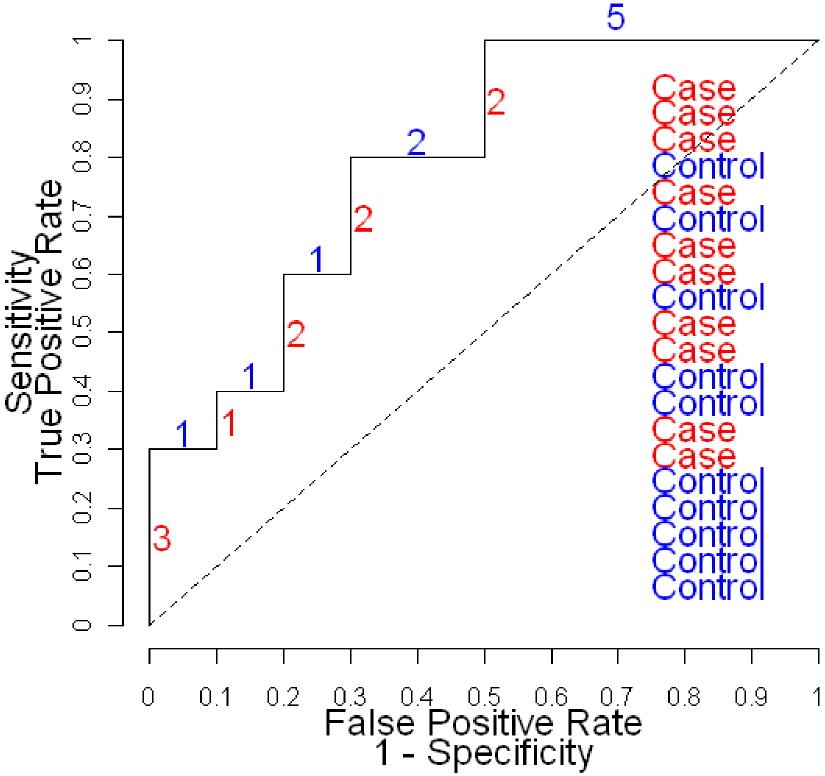


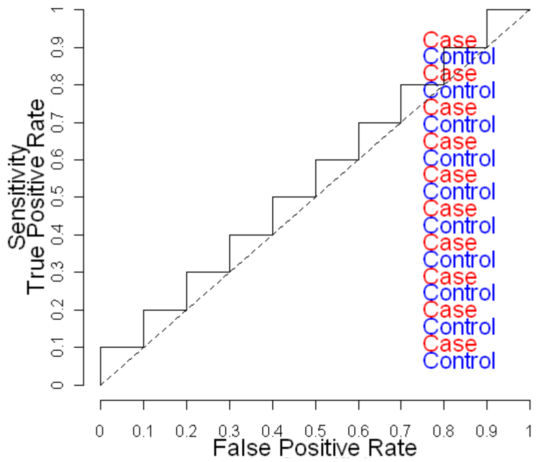
K = disease prevalence

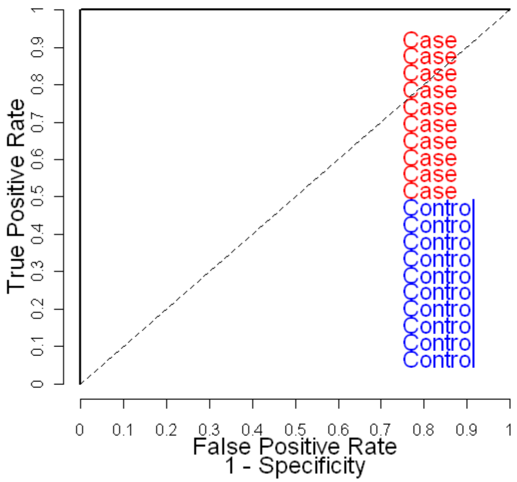
Predictor explains 7% of variance in liability

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 difference prevalence

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

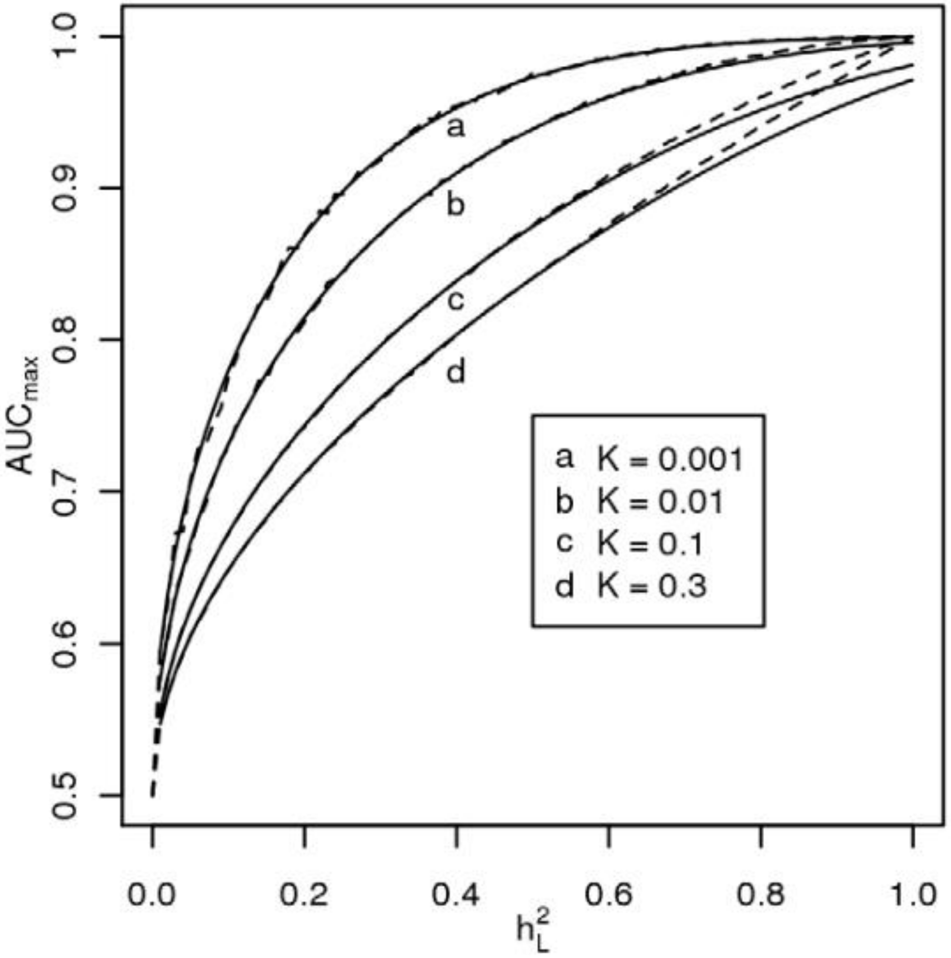


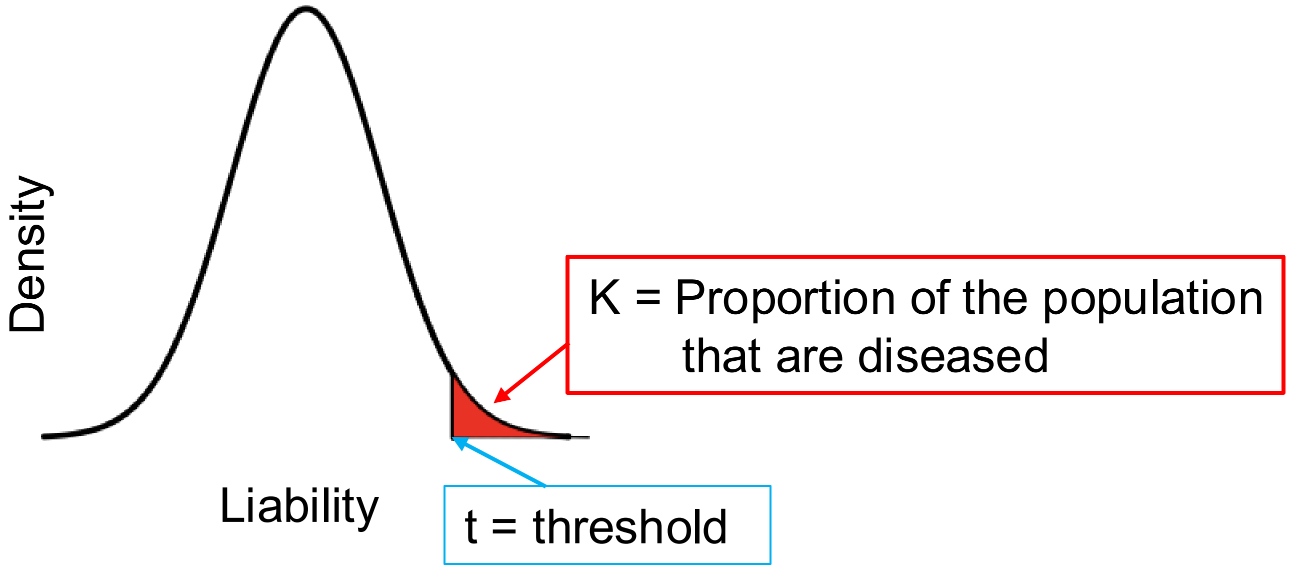
Figure 2. Relationship between maximum AUC (AUC_{max}) from a genomic profile and heritability on the liability scale h^2_L. For

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 R² on liability scale ★

Linear regression; Y are 0s and 1s

Null: Y= covariates + e

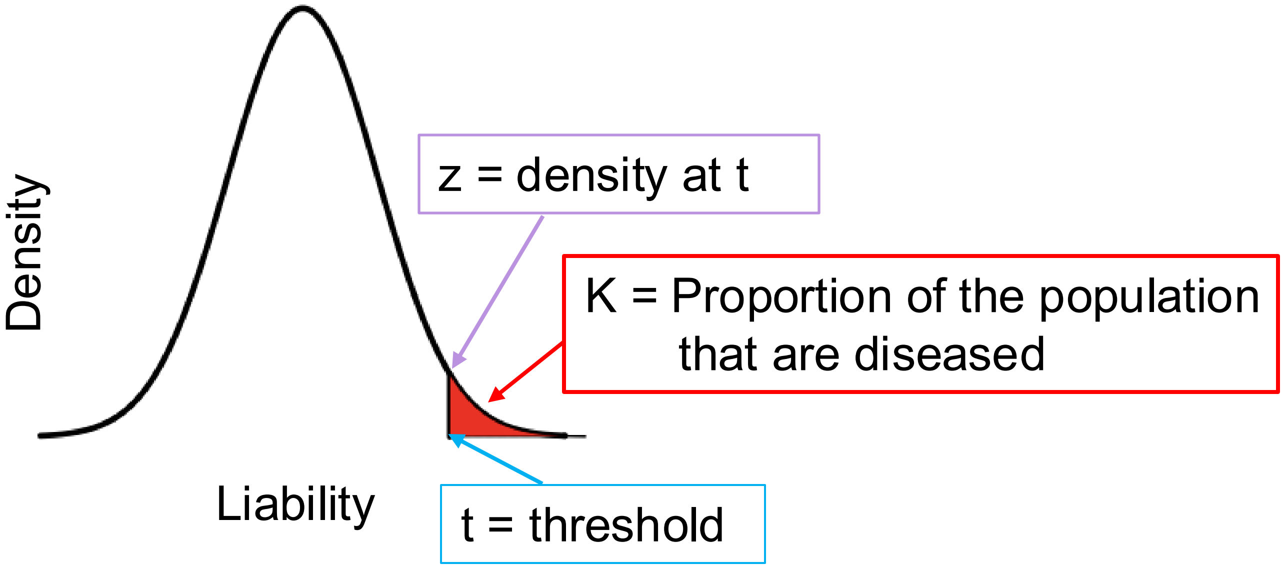
Full: Y= covariates + PGS + e

R² on the observed scale

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

R² on the liability scale

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



z = density at t

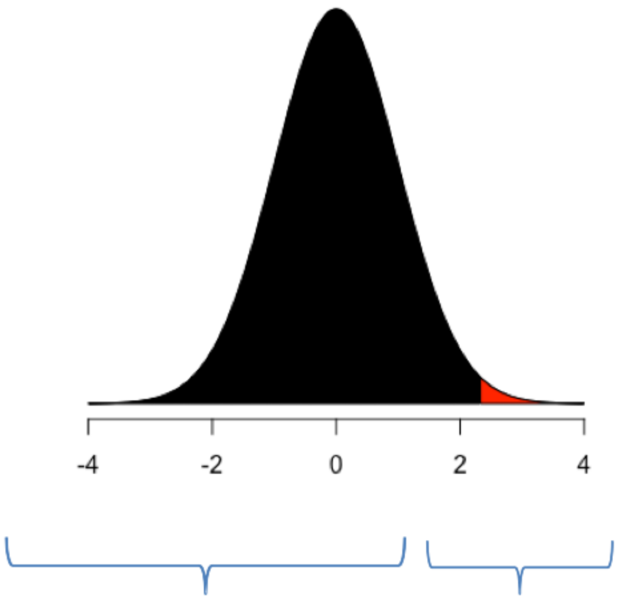
K = Proportion of the population that are diseased

t = threshold

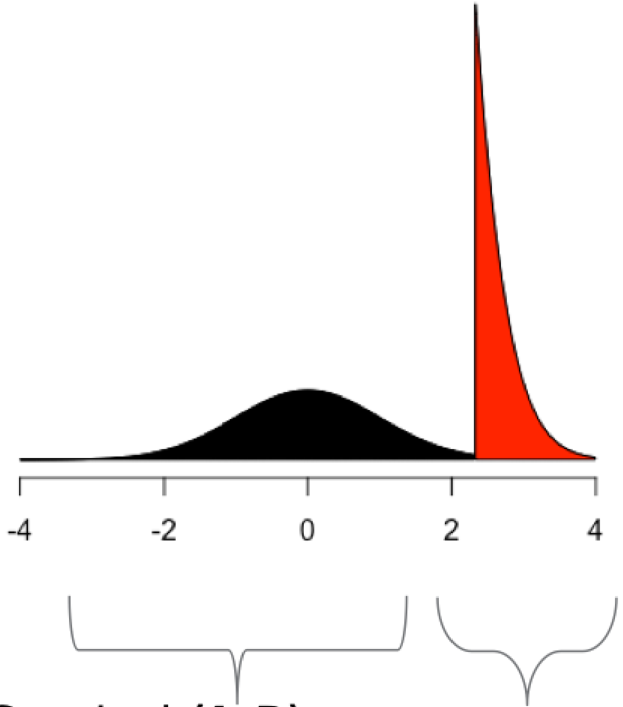
Lee, 2012, Genet Epidemiol

Prediction R^2 on liability scale

Ascertainment in case-control studies



Unaffected (1-K) Affected (K)



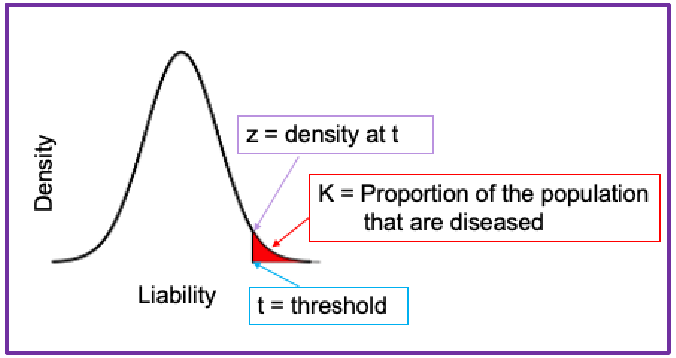
Control (1-P) Case (P)

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

?

Prediction R^2 on liability scale

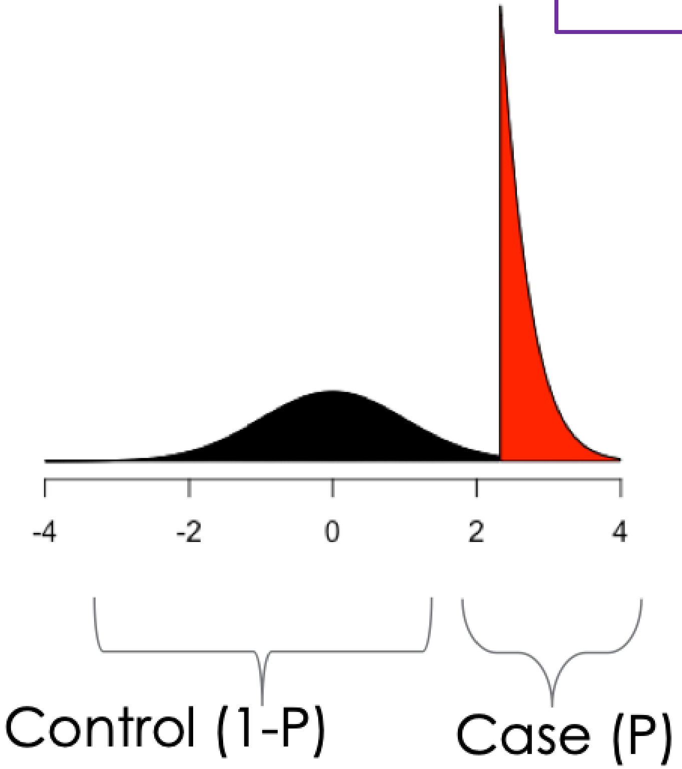
Ascertainment in case-control studies



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

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

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

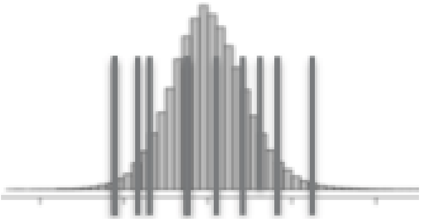


Lee, 2012, Genet Epidemiol

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

Decile odds ratio

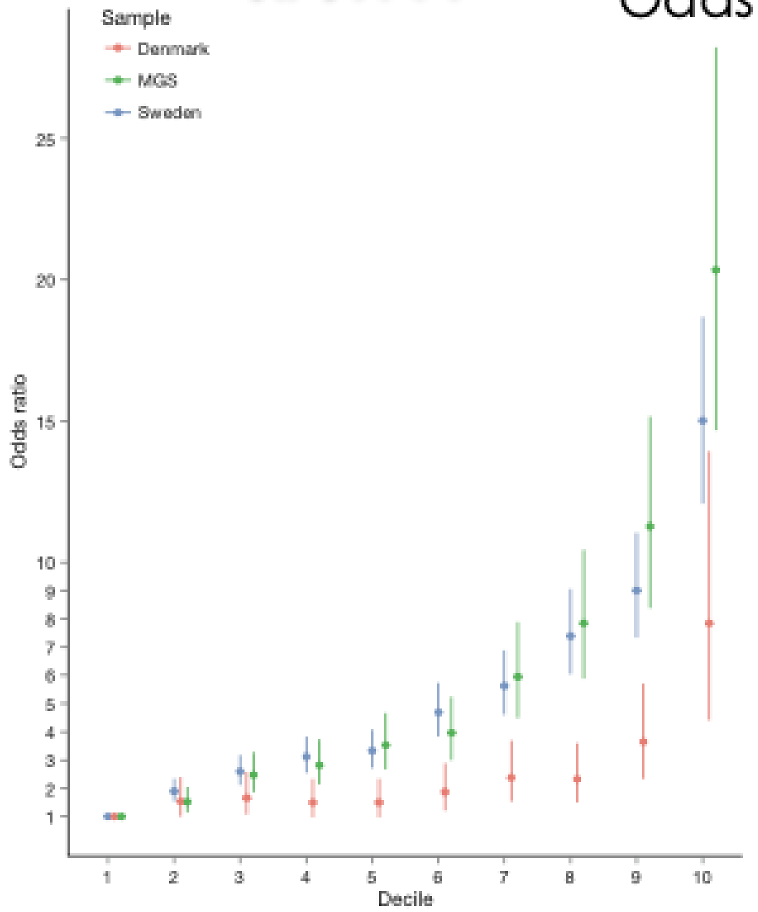


Cut distribution into deciles

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

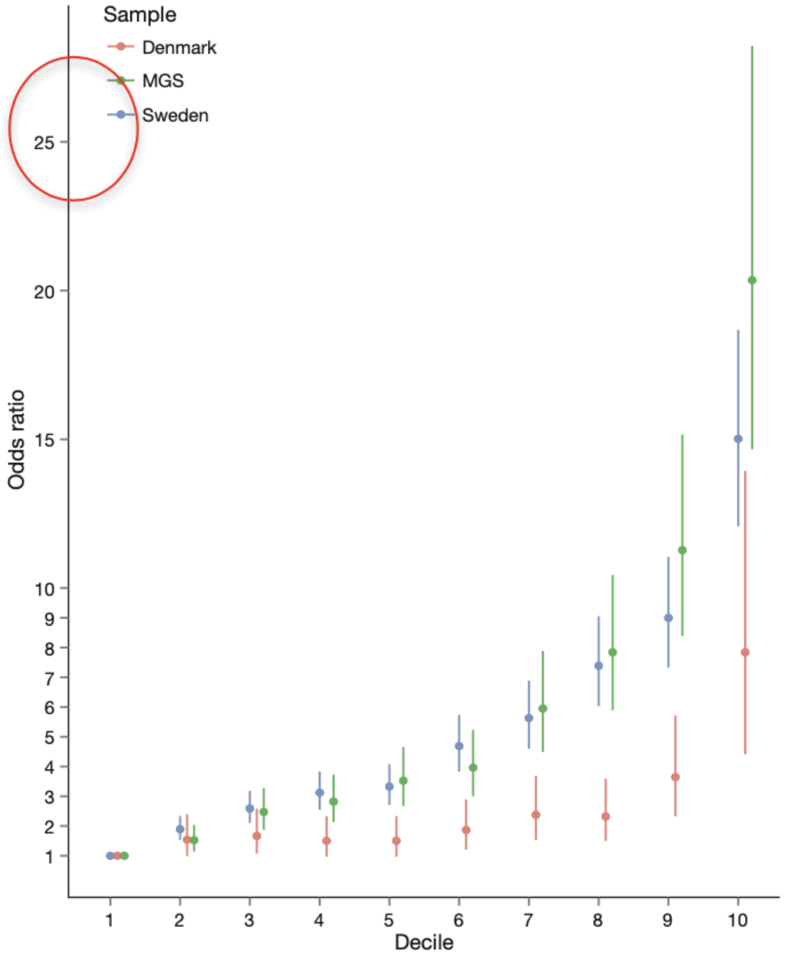


- Good visualisation
- Shows that there could be utility in using high vs low profile risk scores
- But remember case-control samples are 50% cases
- Would look less impressive if a population sample

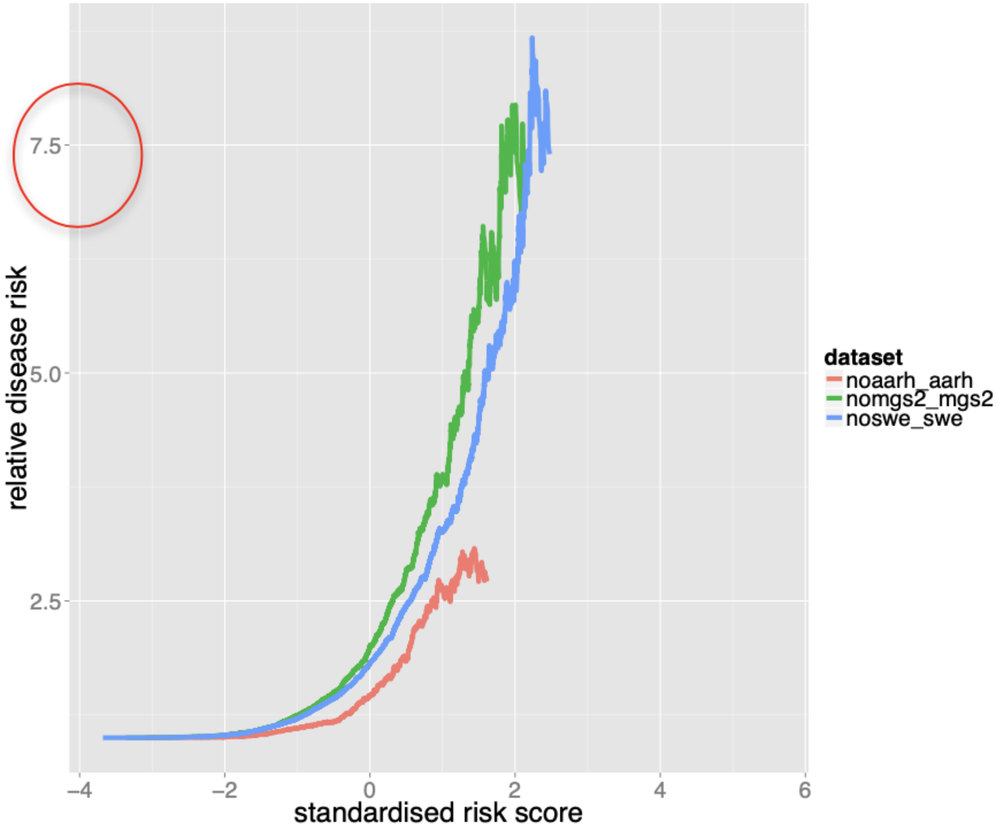
PGC-SCZ 2014 108 loci Nature

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 1st decile

= 23/103

Odds being a case in 10th decile

= 83/40

Odds ratio between 10th and 1st decile

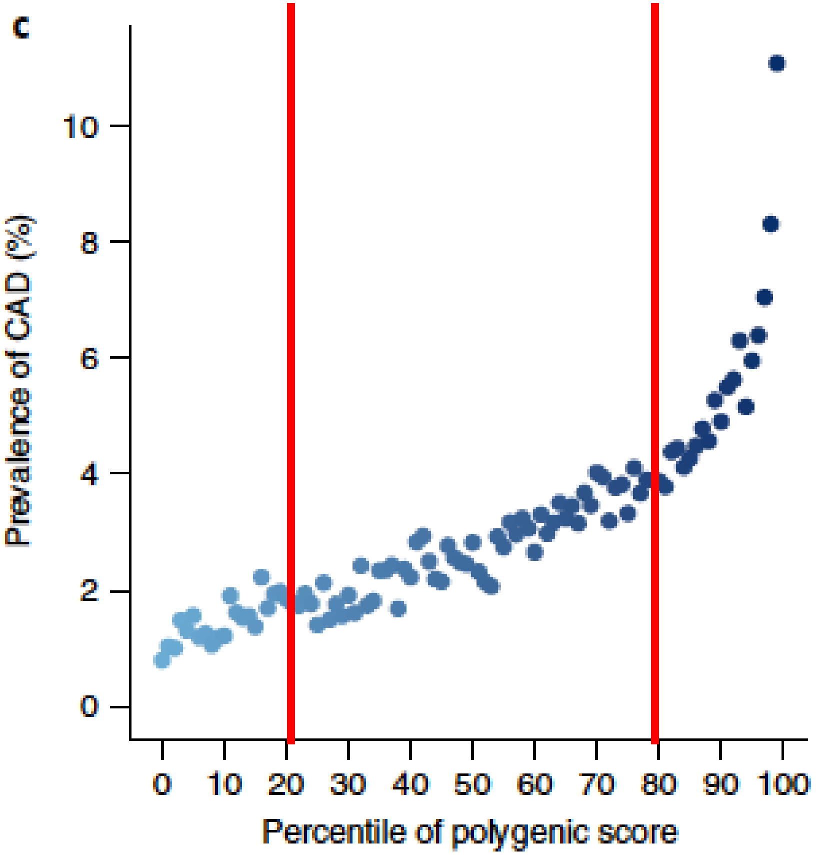
= (23/103) / (83/40) =9.3

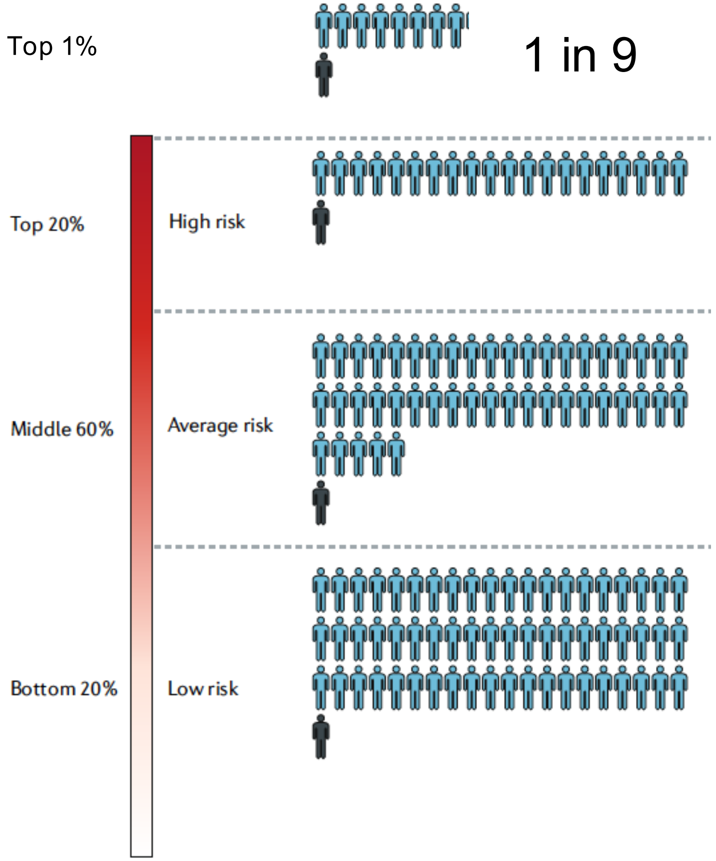
$$Odds\,ratio = \frac{Odds_1}{Odds_0} = \frac{P_1/_{1-P_1}}{P_0/_{1-P_0}}$$

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

P = probability of being case

Risk stratification

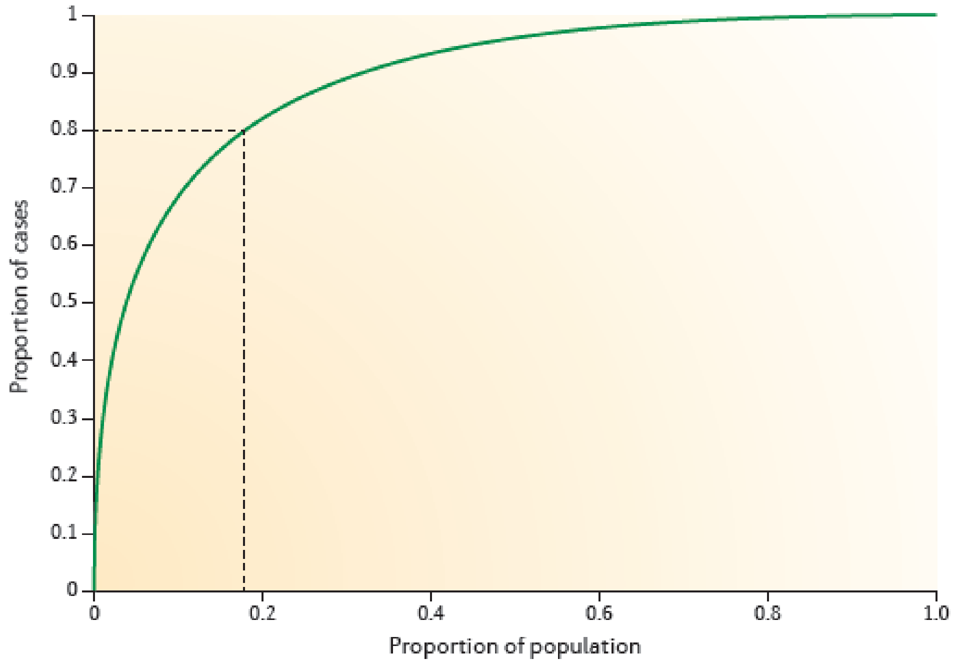




Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Torkamani et al, Nat Rev Genetics, 2018

Stratification & health economics



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

Parameters determining the prediction accuracy

The expected value of prediction accuracy:

Variance explained by the predictor

$$R^2 = \frac{h_m^2}{1 + C}$$

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

$$C \approx \frac{m}{N h_m^2}$$

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

- *N*: discovery sample size
- *m*: the number of SNPs (assume LD-independent)
- h_m^2: the SNP-heritability captured by *m* SNPs

Wray et al (2019) Complex trait prediction from genome data. Genetics

What is the maximum prediction accuracy we can get?

Variance explained by the predictor

$$R^2 = \frac{h_m^2}{1 + C}$$

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

We want C to be as small as possible:

- C decreases as Discovery sample N increases
- C decreases as the number of SNPs in the SNP set m decreases

$$C \approx \frac{m}{N h_m^2}$$

Catch-22

As m gets smaller, h_m^2 also gets smaller

How to optimise m and h_m^2 to get max R^2 ?

Wray et al (2019) Complex trait prediction from genome data. Genetics

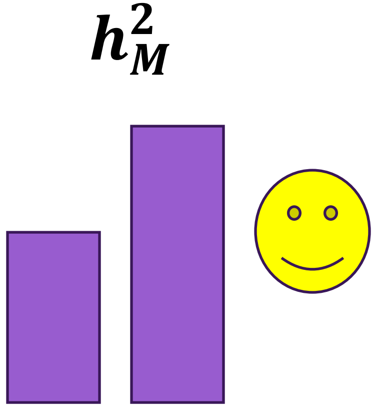
How about whole genome sequencing?

Maximum depends on maximising h_m^2

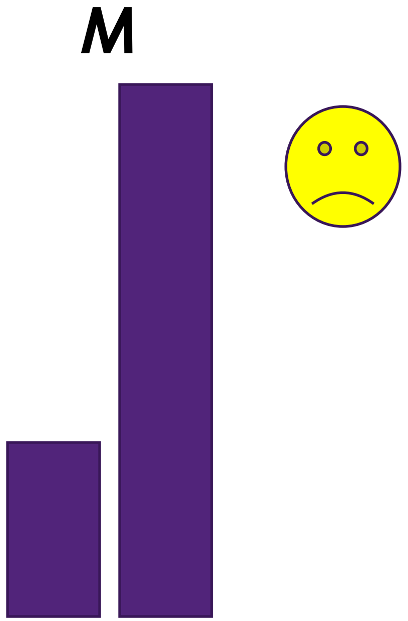
We use GWAS data so the maximum h_m^2 is the SNP-based heritability

Theoretical maximum depends on the heritability of the trait

$$R^2 \approx \frac{h_m^2}{1 + \frac{m}{N h_m^2}}$$



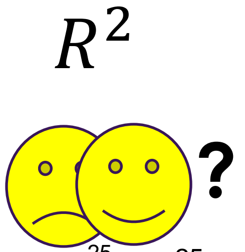
With whole genome sequencing the variance captured by all measured SNPs will increase



But the number of SNPs that we have estimate effect sizes for increases much more

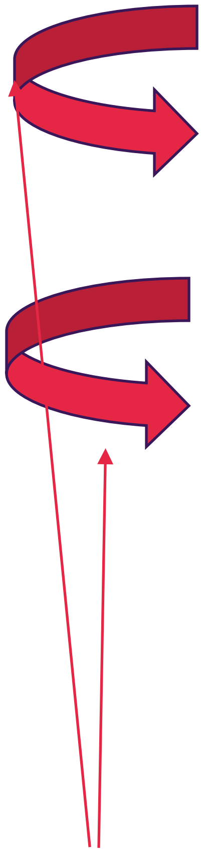
Need MASSIVE discovery sample sizes for WGS associations

Also… rare variants are less likely to be shared across populations



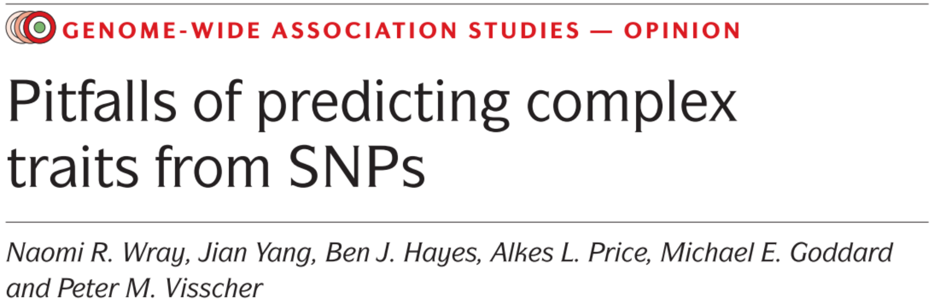
Wray et al (2019) Complex trait prediction from genome data. Genetics

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

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)

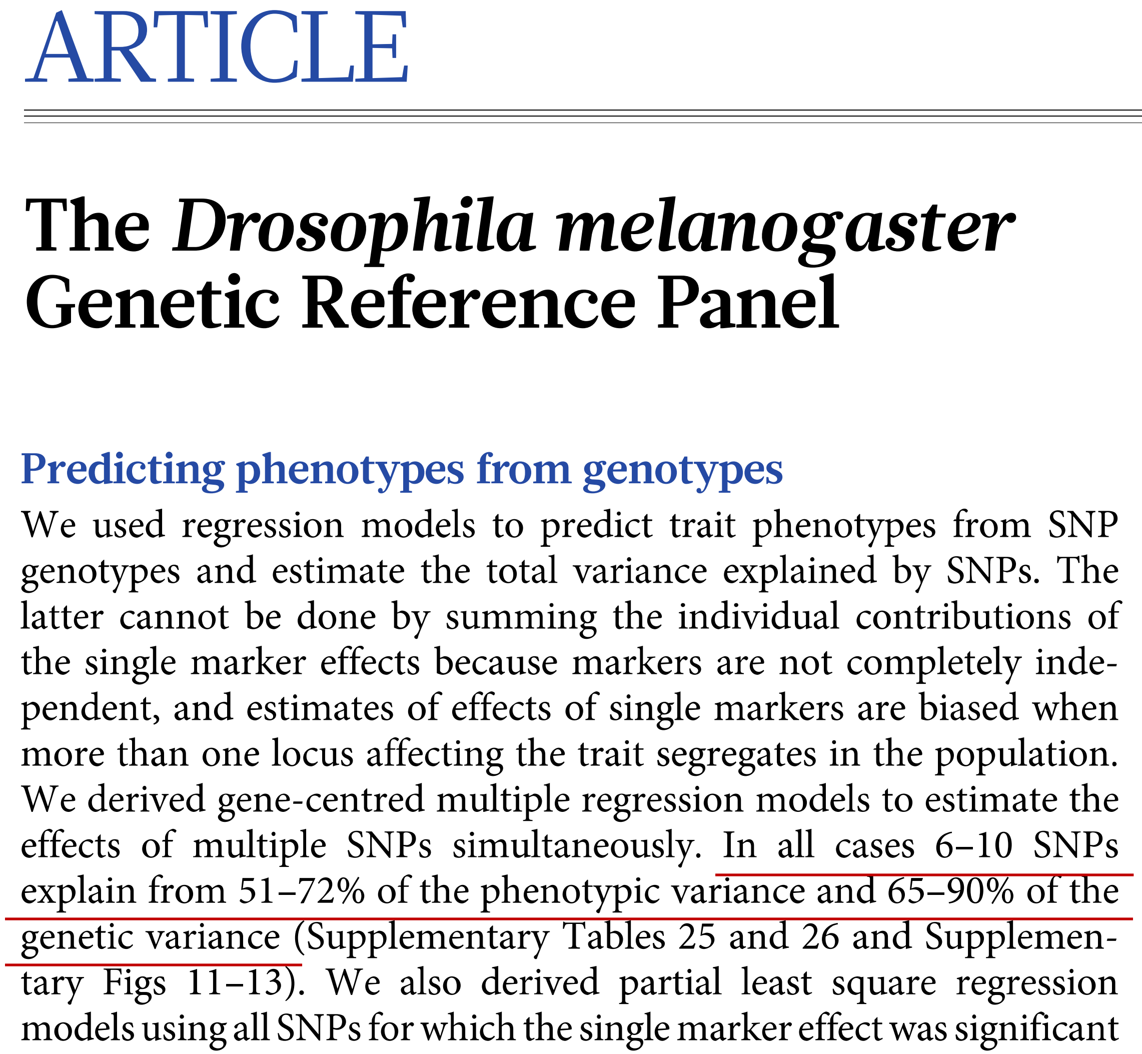
1) Multiple linear regression of y on x (when M<N)

$E(R^2) = M/N$ 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

Winner’s curse



ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel

Predicting phenotypes from genotypes

We used regression models to predict trait phenotypes from SNP genotypes and estimate the total variance explained by SNPs. The latter cannot be done by summing the individual contributions of the single marker effects because markers are not completely independent, and estimates of effects of single markers are biased when more than one locus affecting the trait segregates in the population. We derived gene-centred multiple regression models to estimate the effects of multiple SNPs simultaneously. In all cases 6–10 SNPs explain from 51–72% of the phenotypic variance and 65–90% of the genetic variance (Supplementary Tables 25 and 26 and Supplementary Figs 11–13). We also derived partial least square regression models using all SNPs for which the single marker effect was significant

~10 best markers selected from 2.5 million markers

> “A cross-validated Bayesian prediction analysis using all genetic markers on the same data found that only 6% of phenotypic variation could be explained by the predictor.”
> (Wray et al., 2013. Nat. Rev. Genet.)

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

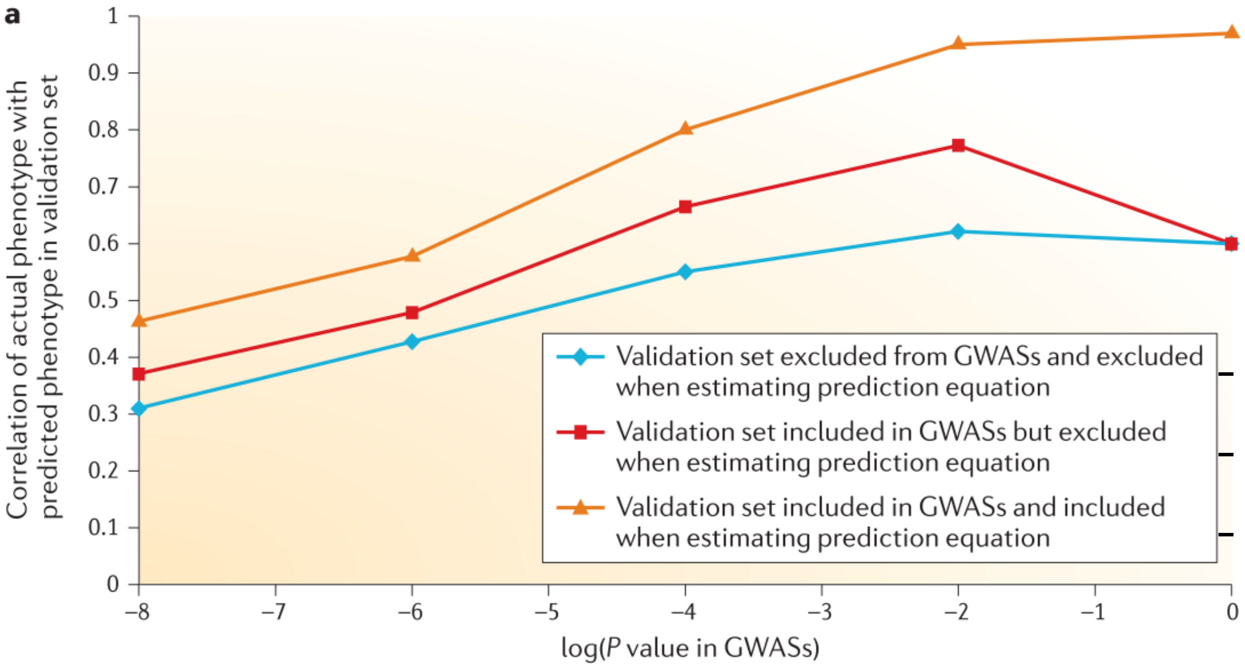
$$\operatorname{cov}(\hat{y}_i, y_i) = \operatorname{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij}b_j + e_i\}$$

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

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

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

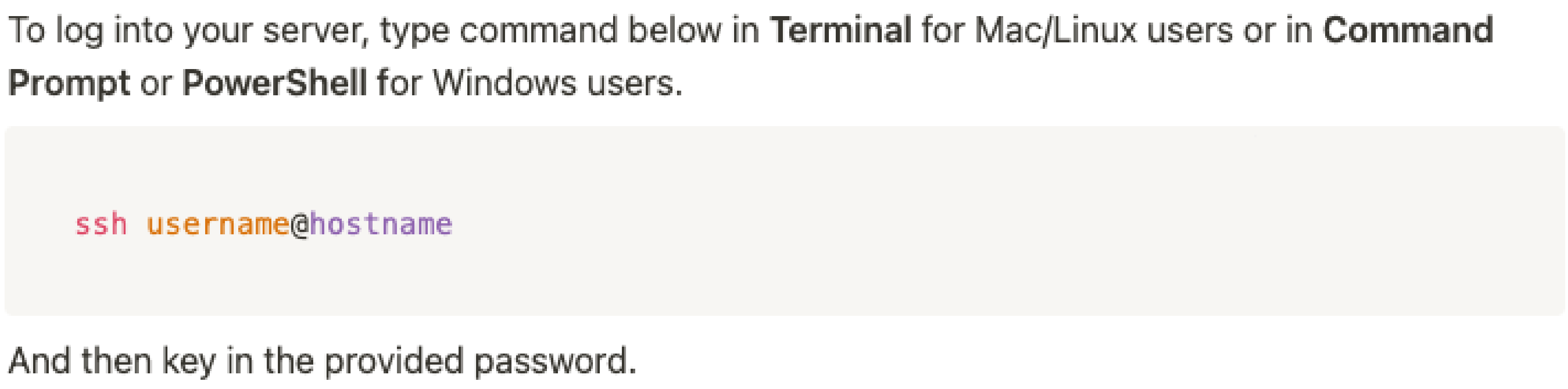
Summary

- Evaluation of prediction performance
  - Prediction accuracy and bias for quantitative traits
  - Different statistics for disease traits with pros and cons
- Parameters determining the prediction accuracy
  - SNP-based heritability (h_m^2)
  - Number of SNPs (m)
  - Discovery sample size (N)
- Pitfalls in the prediction analysis
  - No target sample (only discovery sample)
  - Overlapping discovery & target sample
  - Less obvious non-independence

Questions?

Practical 2: Evaluation of PRS prediction

https://cnsgenomics.com/data/teaching/GNGWS25/module5/Practical2_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.