

GnG Winter School 2022

# Prediction accuracy and pitfalls

**Huanwei Wang**  
(with thanks to Naomi, Guiyan, and Jian Zeng)

[huanwei.wang@uq.edu.au](mailto:huanwei.wang@uq.edu.au)

- **Discovery/Training/Derivation**

- Estimate the effect sizes ( $\hat{b}$ ) of SNPs on a trait ( $y$ ) – GWAS

- **Tuning/Validation**

- Further estimate some parameters (depends on methods; not all methods require it)

- **Target/Testing/Validation**

- Build a polygenetic risk score (PRS) ( $\hat{y}$ ):

$$\hat{y} = \sum_i \hat{b}_i x_i$$

- $\hat{b}_i$  is the estimated effect size for  $i$ -th SNP
- $x_i$  is the genotype value for  $i$ -th SNP
- Evaluate the prediction performance/accuracy

- $y$  is a quantitative phenotype

$$R^2(y, \hat{y}) = \frac{\text{Cov}(y, \hat{y})^2}{\text{Var}(y)\text{Var}(\hat{y})}$$

- the coefficient of determination
  - or the square of correlation coefficient
  - or the variance of  $y$  explained by  $\hat{y}$
  
  - Reduce:  $y \sim \text{cov}$ ; Full:  $y \sim \text{cov} + \hat{y}$
  - Incremental  $R^2$ :  $R_{full}^2 - R_{reduce}^2$
- Regression of phenotypes ( $y$ ) on PRS ( $\hat{y}$ )
    - Deviation from expectation of the slope
    - Expectation is usually 1
    - If not close to expectation, then biased

- Nagelkerke's  $R^2$
- AUC
- Decile Odds Ratio
- Variance explained on liability scale
- Risk stratification

# 1) Nagelkerke's $R^2$

Logistic regression:

full model:  $y \sim \text{covariates} + \text{score}$

reduced model:  $y \sim \text{covariates}$

- Many pseudo- $R^2$  statistic for logistic regression
- Cox & Snell  $R^2$

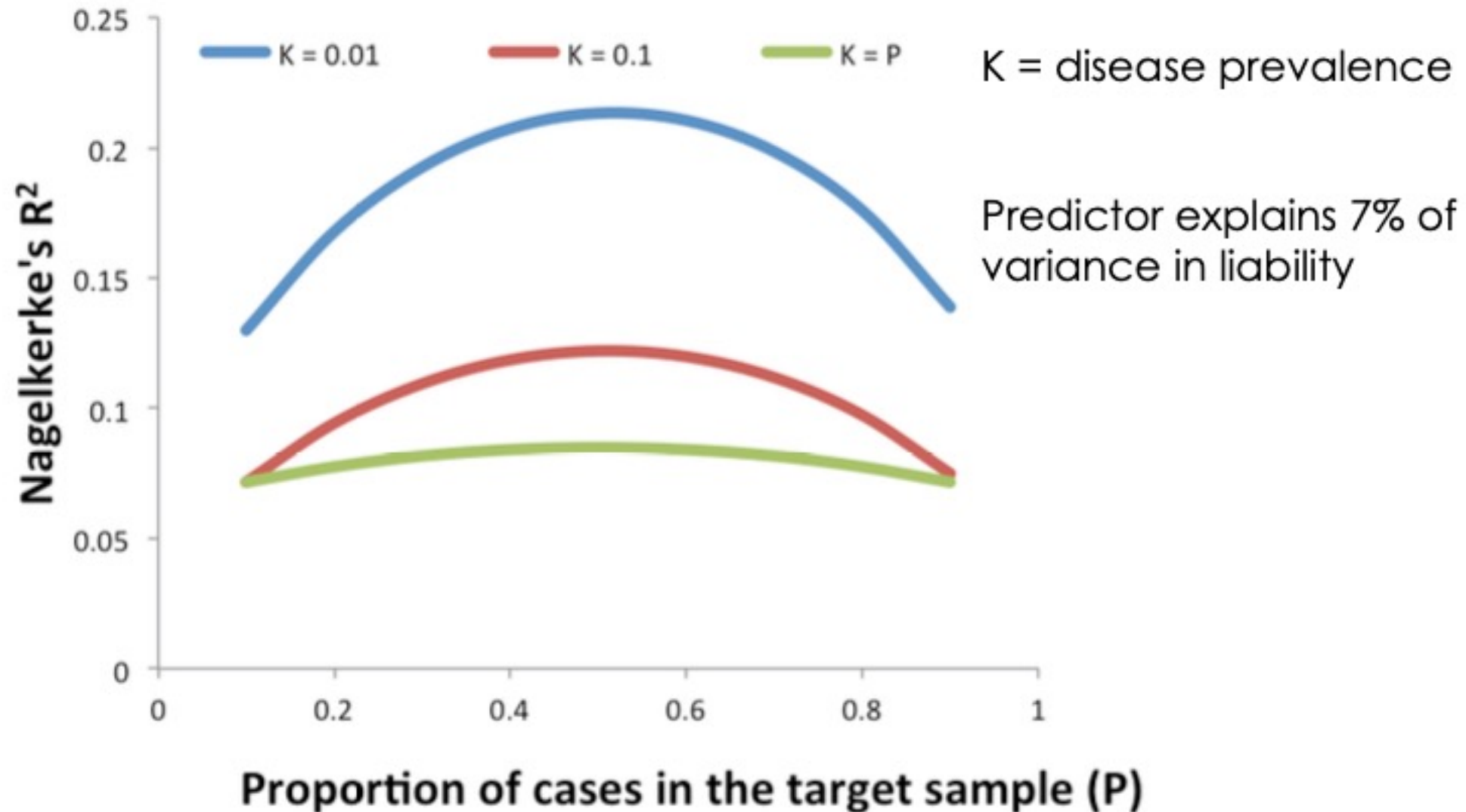
$$1 - \left( \frac{L_{reduced}}{L_{full}} \right)^{\frac{2}{N}} \in [0, 1 - (L_{reduced})^{\frac{2}{N}}]$$

N is the sample size; L is the likelihood

- Nagelkerke's  $R^2$

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

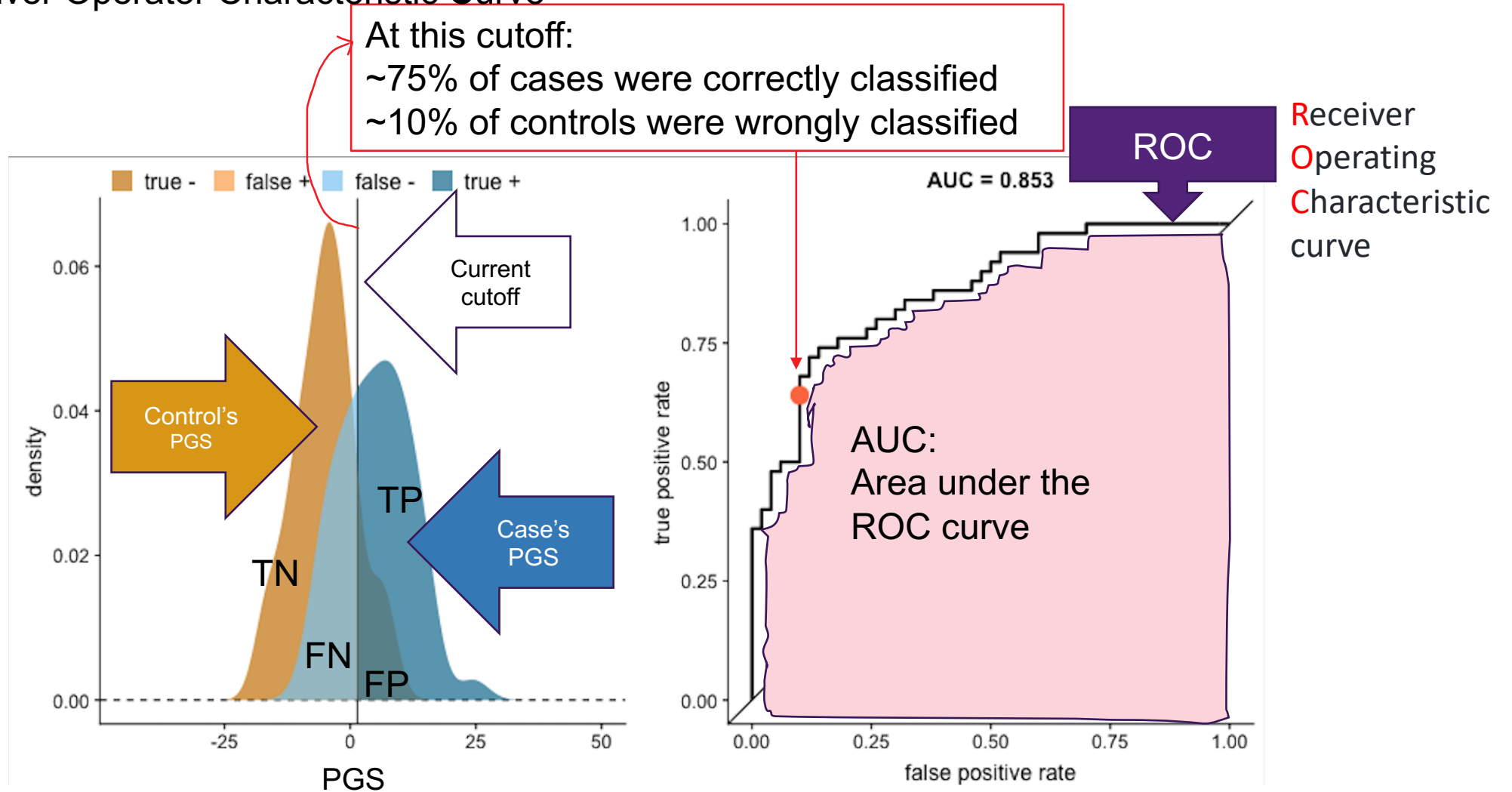
# Nagelkerke's $R^2$ depends on case proportion in the sample



# 2) AUC

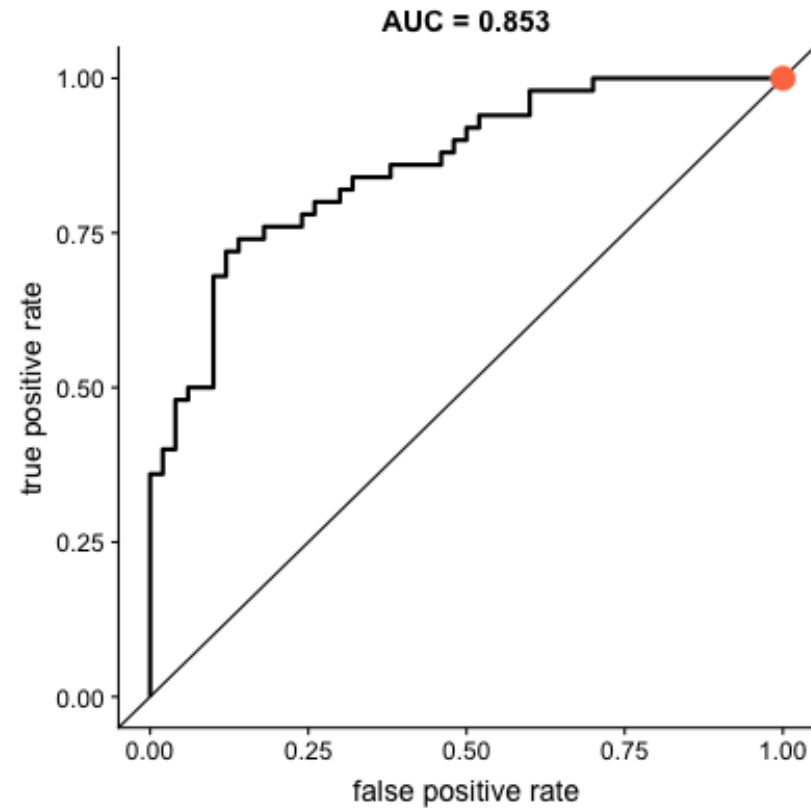
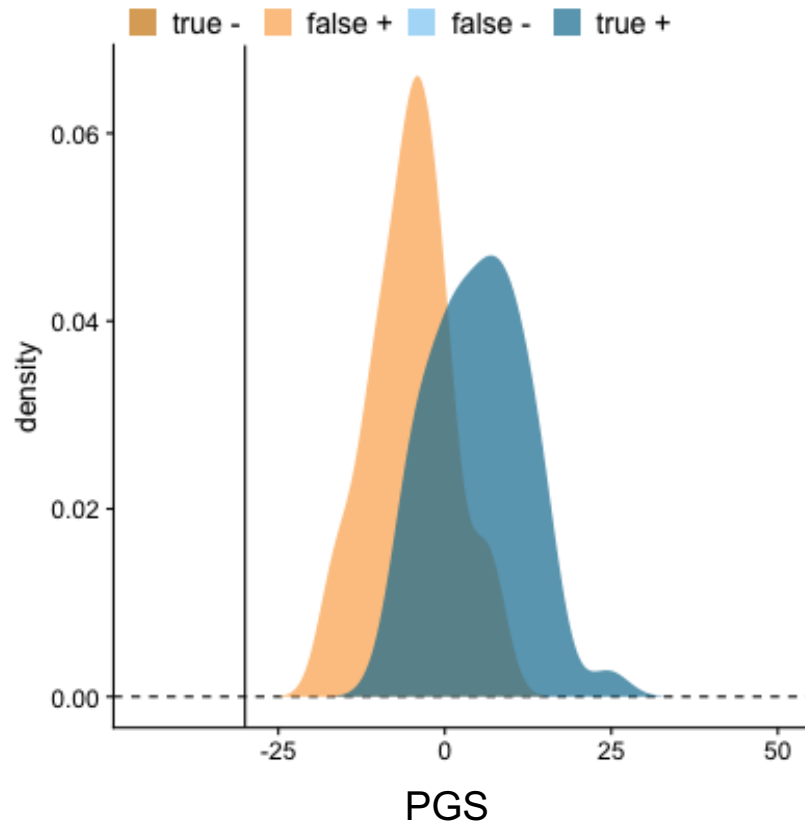
## Area Under Receiver Operator Characteristic Curve

Toy example:



$$\text{True Positive Rate} = \text{TP} / (\text{TP} + \text{FN}) = \text{Sensitivity}$$

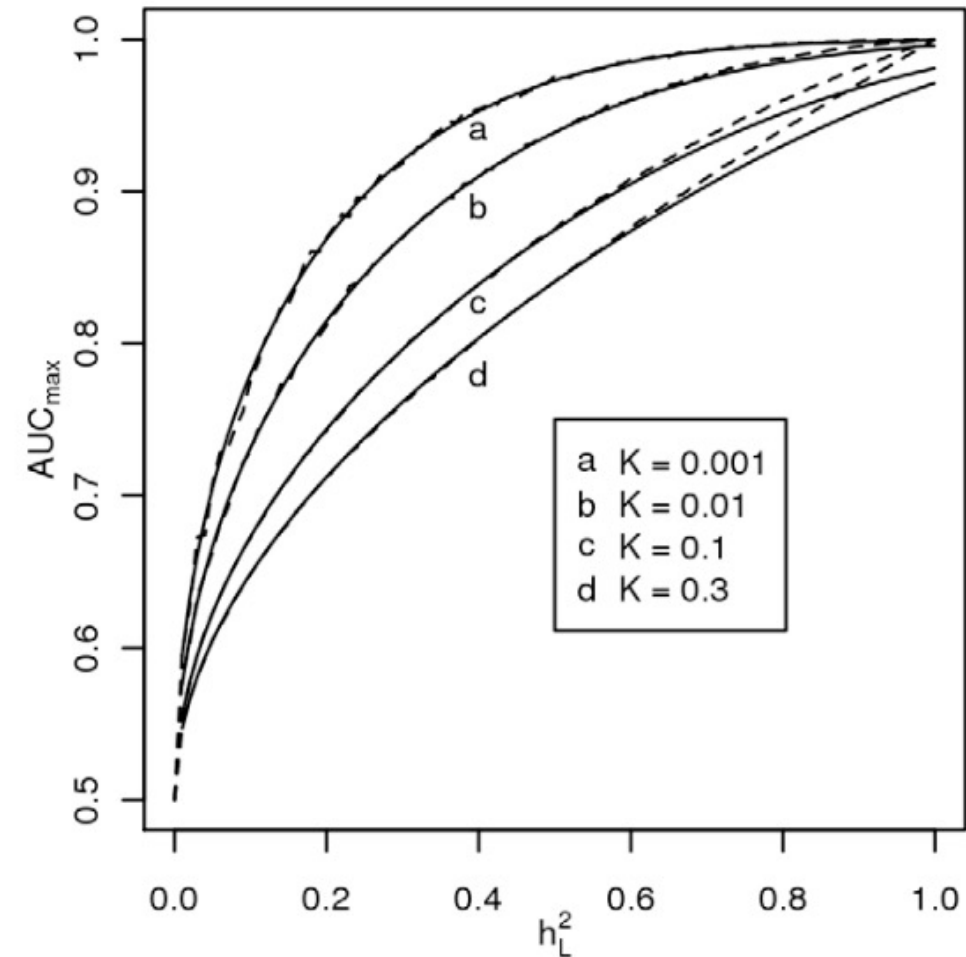
$$\text{False Positive Rate} = \text{FP} / (\text{FP} + \text{TN}) = 1 - \text{Specificity}$$



<https://www.youtube.com/watch?v=y4wTRSGrVuo>

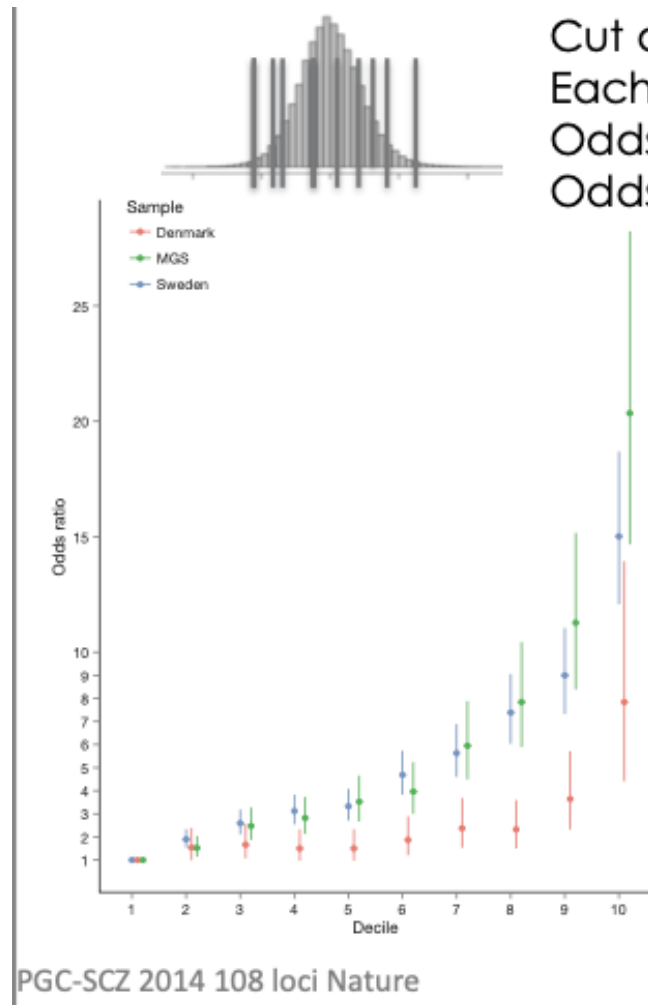


- Range 0.5 to 1;
- 0.5 has no predictive value
- Probability that a randomly selected case has a score higher than a randomly selected control
- Independent to proportion of cases and controls in sample



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

# 3) 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 1<sup>st</sup> 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

13

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

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

$P$  = probability of being case

Toy example:

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

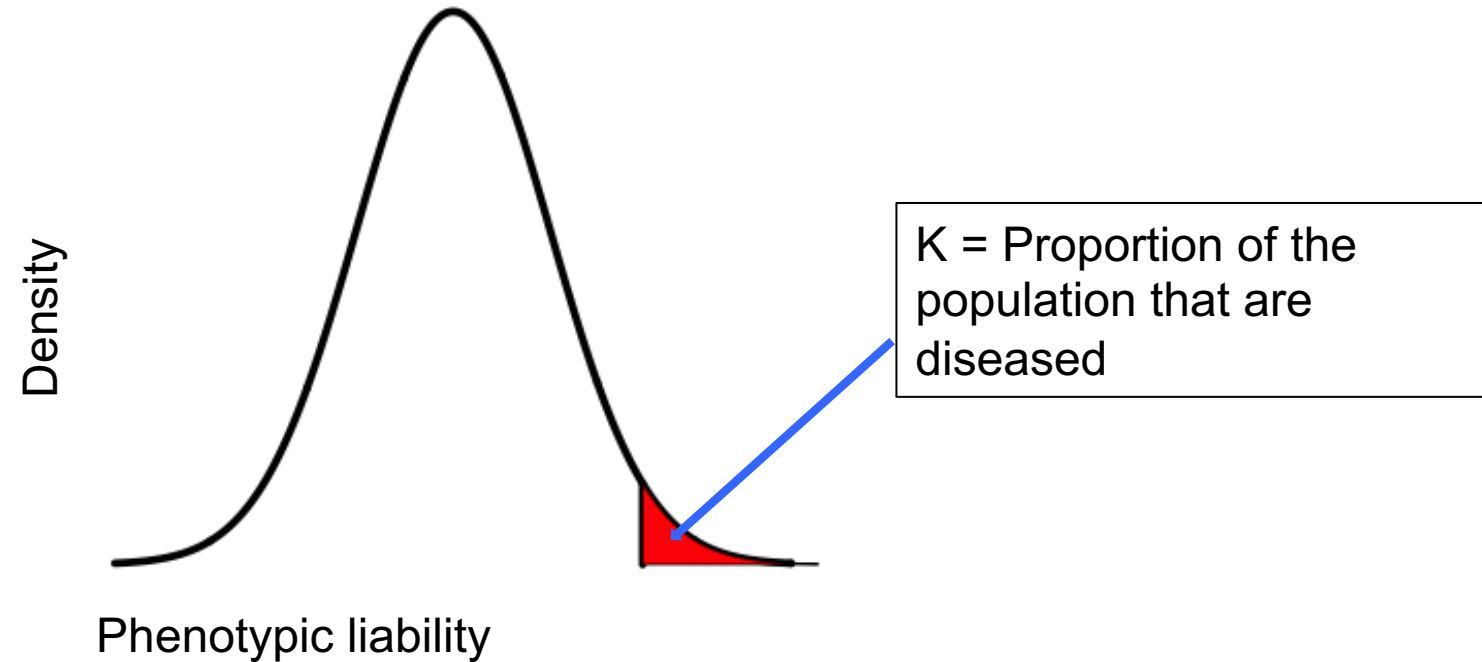
Odds being a case in 1<sup>st</sup> decile  
= 23/103

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

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

## Liability threshold model

- Observed probability 0-1 scale
- Underlying unobserved continuous liability scale
- heritability is independent of disease prevalence



Falconer 1965; Lee 2011

More details in Lecture 9

# 4) R<sup>2</sup> on liability scale

R<sup>2</sup> on the liability scale when using ascertained case-control studies

Linear regression; Y are 0s and 1s

Null: Y = cov + e

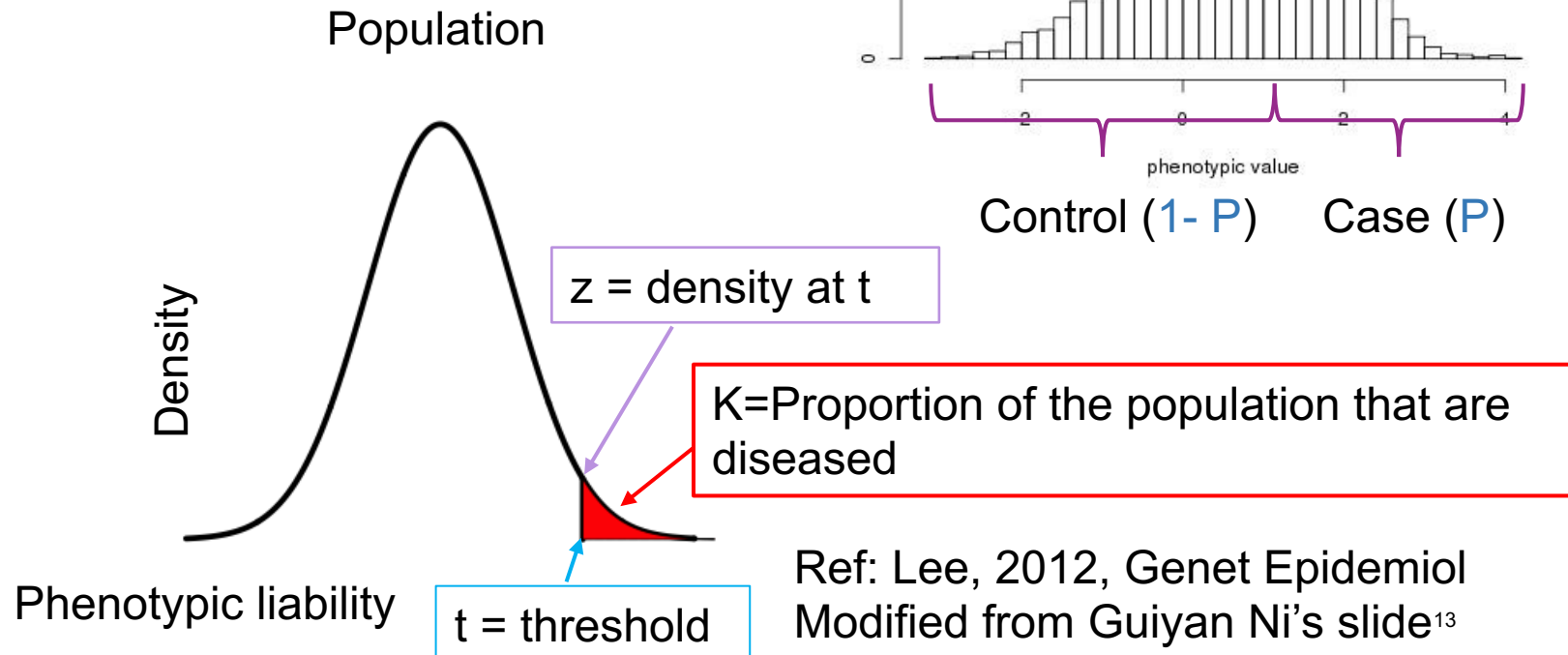
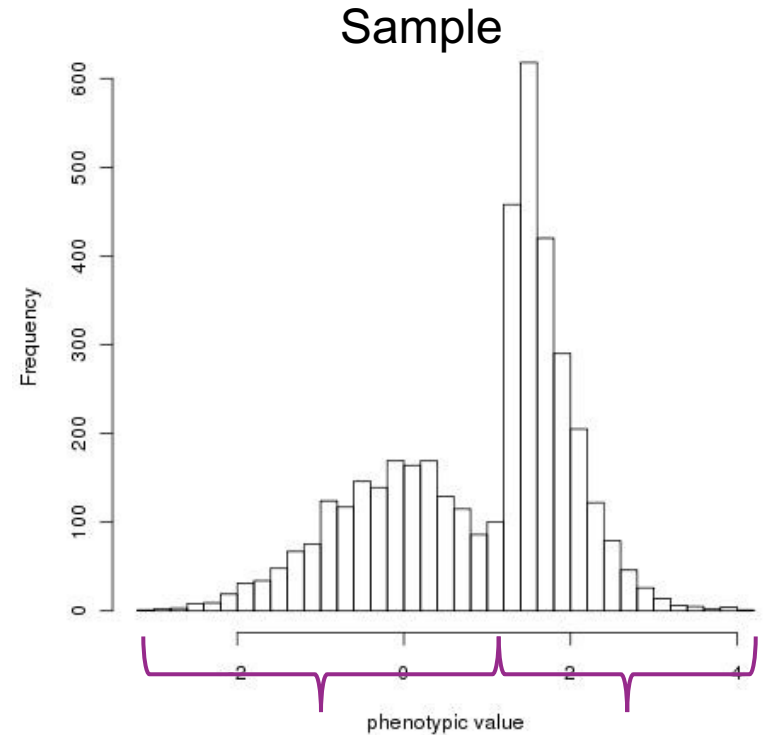
Full: Y = cov + PGS + e

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

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

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

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



Ref: Lee, 2012, Genet Epidemiol  
Modified from Guiyan Ni's slide<sup>13</sup>

# 5) Net reclassification index

Introduced in 2008 (Pencina et al.)

Getting popular, but still under debate

Kathleen et al. 2014

which was corrected after recalibration. Using a risk threshold of 7.5%, addition of the polygenic risk score to pooled cohort equations resulted in a net reclassification improvement of 4.4% (95% CI, 3.5% to 5.3%) for cases and -0.4% (95% CI, -0.5% to -0.4%) for noncases (overall net reclassification improvement, 4.0% [95% CI, 3.1% to 4.9%]).

The NRI, as originally proposed, seeks to quantify whether a new marker provides clinically relevant improvements in prediction. In the definition of “net reclassification indices,” the risk prediction model with established predictors is called the “old” model. The model that adds the new marker is the “new” model. “Events” are cases—persons who have or will have the disease or outcome in the absence of intervention. “Nonevents” are controls. The formula defining the NRI is<sup>4</sup>

$$\text{NRI} = P(\text{up}|\text{event}) - P(\text{down}|\text{event}) + P(\text{down}|\text{nonevent}) - P(\text{up}|\text{nonevent}). \quad (1)$$

“Up” means that the new risk model places a person into a higher risk category than the old model. Similarly, “down” means the new model places a person into a lower risk category. For example,  $\text{NRI}^{0.2}$  means a two-category index with

Example from Elliott et al. 2020

“Old model”: pooled cohort equations

7.5% is the threshold for intervention (e.g. statin for CVD)

“New” model: “Old”+PRS

The prediction accuracy of PRS ( $\hat{y}$ ) for a quantitative trait  $y$

$$R^2(y, \hat{y}) = \frac{\text{Cov}(y, \hat{y})^2}{\text{Var}(y)\text{Var}(\hat{y})}$$

The expected value of this prediction accuracy

$$E(R^2) = \frac{h_M^2}{1 + M/(Nh_M^2)} < h_M^2$$

- N: discovery sample size
- M: the number of SNPs (assume LD-independent)
- $h_M^2$ : the SNP-heritability captured by M SNPs
  
- An upper bound of  $h_M^2$
- Larger N, larger  $R^2$
- The trade-off between M and  $h_M^2$ 
  - More SNPs, larger M, smaller  $R^2$
  - More SNPs, larger  $h_M^2$ , larger  $R^2$

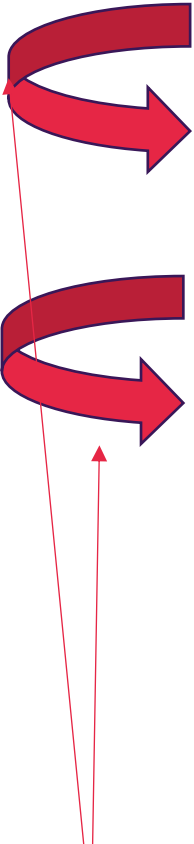
$$R^2(y, \hat{y}) = \frac{\text{Cov}(y, \hat{y})^2}{\text{Var}(y)\text{Var}(\hat{y})}$$

$$\begin{aligned} E(\text{Cov}(y, \hat{y})) &= E\left(\text{Cov}\left(\sum_i^M b_i x_i + e, \sum_i^M \hat{b}_i x_i\right)\right) = \sum_i^M E(\text{Cov}(b_i x_i, \hat{b}_i x_i)) = \sum_i^M b_i E(\hat{b}_i) \text{Var}(x_i) \\ &= \sum_i^M b_i^2 \text{Var}(x_i) = h_M^2 \text{Var}(y) \end{aligned}$$

$$\begin{aligned} E(\text{Var}(\hat{y})) &= E\left(\text{Var}\left(\sum_i^M \hat{b}_i x_i\right)\right) = \sum_i^M E(\hat{b}_i^2) \text{Var}(x_i) = \sum_i^M (b_i^2 + \text{Var}(\hat{b}_i)) \text{Var}(x_i) = \sum_i^M b_i^2 \text{Var}(x_i) + \sum_i^M \text{Var}(\hat{b}_i) \text{Var}(x_i) \\ &\approx h_M^2 \text{Var}(y) + M * \text{Var}(y)/N \end{aligned}$$

$$E(R^2(y, \hat{y})) = \frac{h_M^2 * h_M^2}{h_M^2 + M/N} = \frac{h_M^2}{1 + M/(Nh_M^2)}$$



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

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

x: M markers for N samples

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

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

$$E(R^2) = M/N \quad \text{By chance}$$

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

$$E(R^2) \gg m/N \quad \text{+ winner's curse}$$

## Out-of-sample prediction

# The *Drosophila melanogaster* Genetic Reference Panel

Trudy F. C. Mackay<sup>1\*</sup>, Stephen Richards<sup>2\*</sup>, Eric A. Stone<sup>1\*</sup>, Antonio Barbadilla<sup>3\*</sup>, Julien F. Ayroles<sup>1†</sup>, Dianhui Zhu<sup>2</sup>, Sònia Casillas<sup>3†</sup>, Yi Han<sup>2</sup>, Michael M. Magwire<sup>1</sup>, Julie M. Cridland<sup>4</sup>, Mark F. Richardson<sup>5</sup>, Robert R. H. Anholt<sup>6</sup>, Maite Barrón<sup>3</sup>, Crystal Bess<sup>2</sup>, Kerstin Petra Blankenburg<sup>2</sup>, Mary Anna Carbone<sup>1</sup>, David Castellano<sup>3</sup>, Lesley Chaboub<sup>2</sup>, Laura Duncan<sup>1</sup>, Zeke Harris<sup>1</sup>, Mehwish Javaid<sup>2</sup>, Joy Christina Jayaseelan<sup>2</sup>, Shalini N. Jhangiani<sup>2</sup>, Katherine W. Jordan<sup>1</sup>, Fremiet Lara<sup>2</sup>, Faye Lawrence<sup>1</sup>, Sandra L. Lee<sup>2</sup>, Pablo Librado<sup>7</sup>, Raquel S. Linheiro<sup>5</sup>, Richard F. Lyman<sup>1</sup>, Aaron J. Mackey<sup>8</sup>, Mala Munidasa<sup>2</sup>, Donna Marie Muzny<sup>2</sup>, Lynne Nazareth<sup>2</sup>, Irene Newsham<sup>2</sup>, Lora Perales<sup>2</sup>, Ling-Ling Pu<sup>2</sup>, Carson Qu<sup>2</sup>, Miquel Ràmia<sup>3</sup>, Jeffrey G. Reid<sup>2</sup>, Stephanie M. Rollmann<sup>1†</sup>, Julio Rozas<sup>7</sup>, Nehad Saada<sup>2</sup>, Lavanya Turlapati<sup>1</sup>, Kim C. Worley<sup>2</sup>, Yuan-Qing Wu<sup>2</sup>, Akihiko Yamamoto<sup>1</sup>, Yiming Zhu<sup>2</sup>, Casey M. Bergman<sup>5</sup>, Kevin R. Thornton<sup>4</sup>, David Mittelman<sup>9</sup> & Richard A. Gibbs<sup>2</sup>

## 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

“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.)

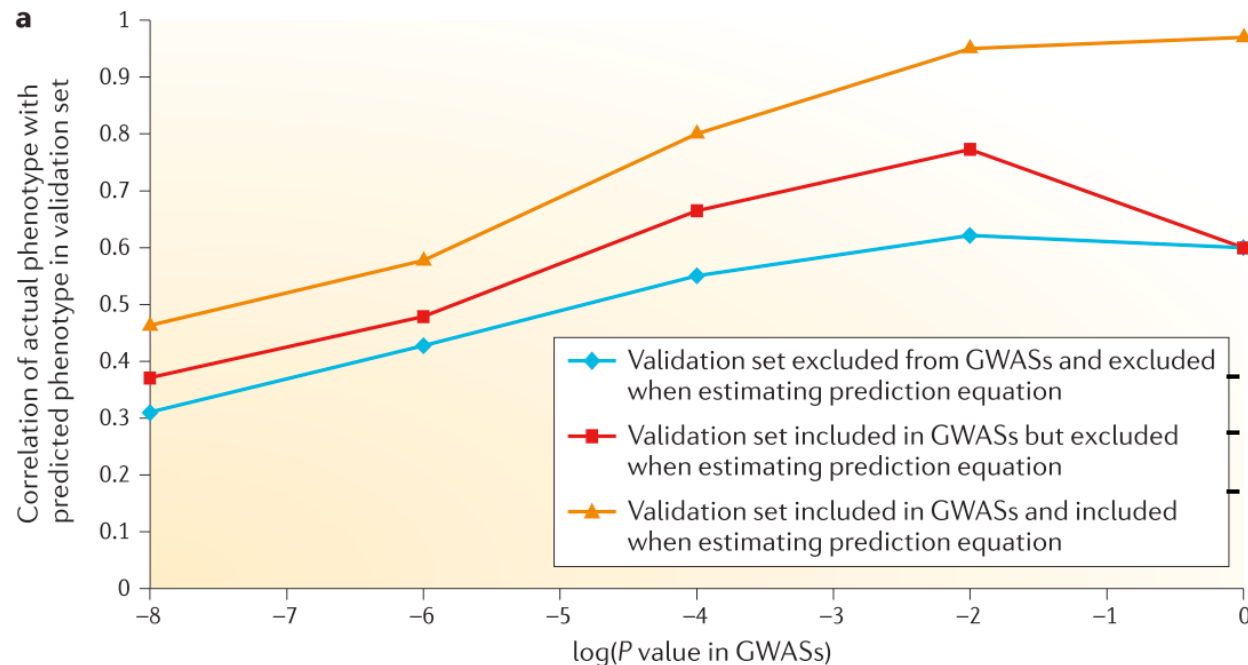
- Overlapping target and discovery sample
- Greater similarity between target and discovery sample (such as relatedness)
  - Cross-validation: not a pitfall, but to be aware

$$\begin{aligned}\text{cov}(\hat{y}_i, y_i) &= \text{cov}\left\{\sum_{j=1}^m (x_{ij} \hat{b}_j), \sum_{j=1}^m x_{ij} b_j + e_i\right\} \\ &= \sum_{j=1}^m \text{var}(x_{ij}) \hat{b}_j b_j + \sum_{j=1}^m x_{ij} \text{cov}(\hat{b}_j, e_i)\end{aligned}$$

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

# Pitfall 3: 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

Estimate SNP effects in total sample

Direct report R<sup>2</sup> in the discovery sample

- measurement of prediction performance
  - $R^2$  for quantitative traits
  - for binary traits
    - Pseudo- $R^2$  (Nagelkerke's  $R^2$ )
    - AUC
    - Decile Odds Ratio
    - variance explained on liability scale
    - risk stratification (Net reclassification index)
- factors affecting prediction accuracy
  - SNP-heritability ( $h_M^2$ ),
  - number of SNPs (M)
  - discovery sample size (N)
- pitfalls
  - No target sample (only discovery sample)
  - Overlapping discovery & target sample
  - non-independence

Thank you for your attention