

GnG Winter School 2023

Prediction accuracy and pitfalls

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

huanwei.wang@uq.edu.au

Genetic prediction

- Discovery/Training/Derivation

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

- Tunning/Testing/Validation

- Further estimate some parameters
- Optional: C+PT: yes; SBayesR: no

- 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

Table 1. Summary of Methods Used to Generate Polygenic Scores

Method	Distribution of SNP Effects (β)	Tuning Sample	Predefined Parameters	Parameters Estimated in Tuning Sample
PC+T	None	Yes	–	p -value threshold
SBLUP	$\beta \sim N\left(0, \frac{h_g^2}{m}\right)$ h_g^2 : SNP-based heritability, m : number of SNPs; $\lambda = m(1 - h_g^2)/h_g^2$	No	λ LD radius in kb	–
Ldpred2-Inf	Same as SBLUP	No	h_g^2 LD radius in cM or kb	–
Ldpred-funct	$\beta_j \sim N(0, c\sigma_j^2)$ $\sum_{j=1}^M 1_{\sigma_j^2 > 0} c\sigma_j^2 = h_g^2$, c is a normalizing constant, σ_j^2 is the expected per SNP heritability under the baseline-LD annotation model estimated by stratified LDSC from the discovery GWAS within Ldpred-funct software	No	h_g^2 LD radius in number of SNPs	–
Ldpred2	$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi m}\right), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1 - \pi \end{cases}$ When sparsity is “true,” the β_j for SNPs in the $(1 - \pi)$ partition are all set to zero	Yes	h_g^2 π software default values, LD radius in cM or kb	π , sparsity
Lassosum	$f(\beta) = \mathbf{y}^T \mathbf{y} + (1 - s) \beta^T \mathbf{X}^T \mathbf{X} \beta - 2 \beta^T \mathbf{X}^T \mathbf{y} + s \beta^T \beta + 2 \lambda \ \beta\ _1$ \mathbf{X} : $n \times m$ matrix of genotypes of LD reference sample, where n is sample size	Yes	LD blocks	λ, s
PRS-CS	$\beta_j \sim N\left(0, \frac{\sigma_j^2}{n} \psi_j\right)$ $\psi_j \sim G(a, \hat{\psi}_j)$ $\hat{\psi}_j \sim G(b, \hat{\phi})$, $\hat{\phi}$ is a global scaling parameter	Yes	$a = 1, b = 0.5$ n LD blocks	ϕ
PRS-CS-auto	Same as PRS-CS, but estimates ϕ from the discovery GWAS	No	$a = 1, b = 0.5$ n LD blocks	–
SBayesR	$\beta_j \pi, \sigma_j^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2 \sigma_j^2), & \text{with probability of } \pi_2 \\ \vdots \\ N(0, \gamma_c \sigma_j^2), & \text{with probability of } 1 - \sum_{c=1}^{c-1} \pi_c \end{cases}$ $\sigma_j^2 \sim \text{Inv-}\chi^2(df, \nu)$ ($df = 4$) $\pi_j \sim \text{Dir}(\mathbf{1})$, estimated from discovery GWAS in SBayesR software γ_j are scaling parameters	No	LD radius in cM or kb $C = 4$ γ software default values	–
MegaPRS	Lasso: $\beta_j \sim DE(\lambda/\sigma_j)$ Ridge regression: $\beta_j \sim N(0, \nu\sigma_j^2)$ BOLT-LMM: $\beta_j \sim \begin{cases} N\left(0, \frac{(1-f_2)\sigma_j^2}{\pi}\right), & \text{with probability of } \pi \\ N\left(0, \frac{f_2\sigma_j^2}{1-\pi}\right), & \text{with probability of } 1 - \pi \end{cases}$	Yes	LD radius in cM or kb Parameters used in BLD-LDAK Grid search parameter values for each method	The tuning cohort is used to estimate the parameters that maximize prediction in BLD-LDAK from these the model that maximizes prediction is selected

f_2 is the proportion of the total mixture variance in the second normal distribution
BayesR: similar to SBayesR with $C = 4$, and π_j and γ_j estimated in the tuning sample
 σ_j^2 is the expected per SNP-heritability under BLD-LDAK model using SumHer

- 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

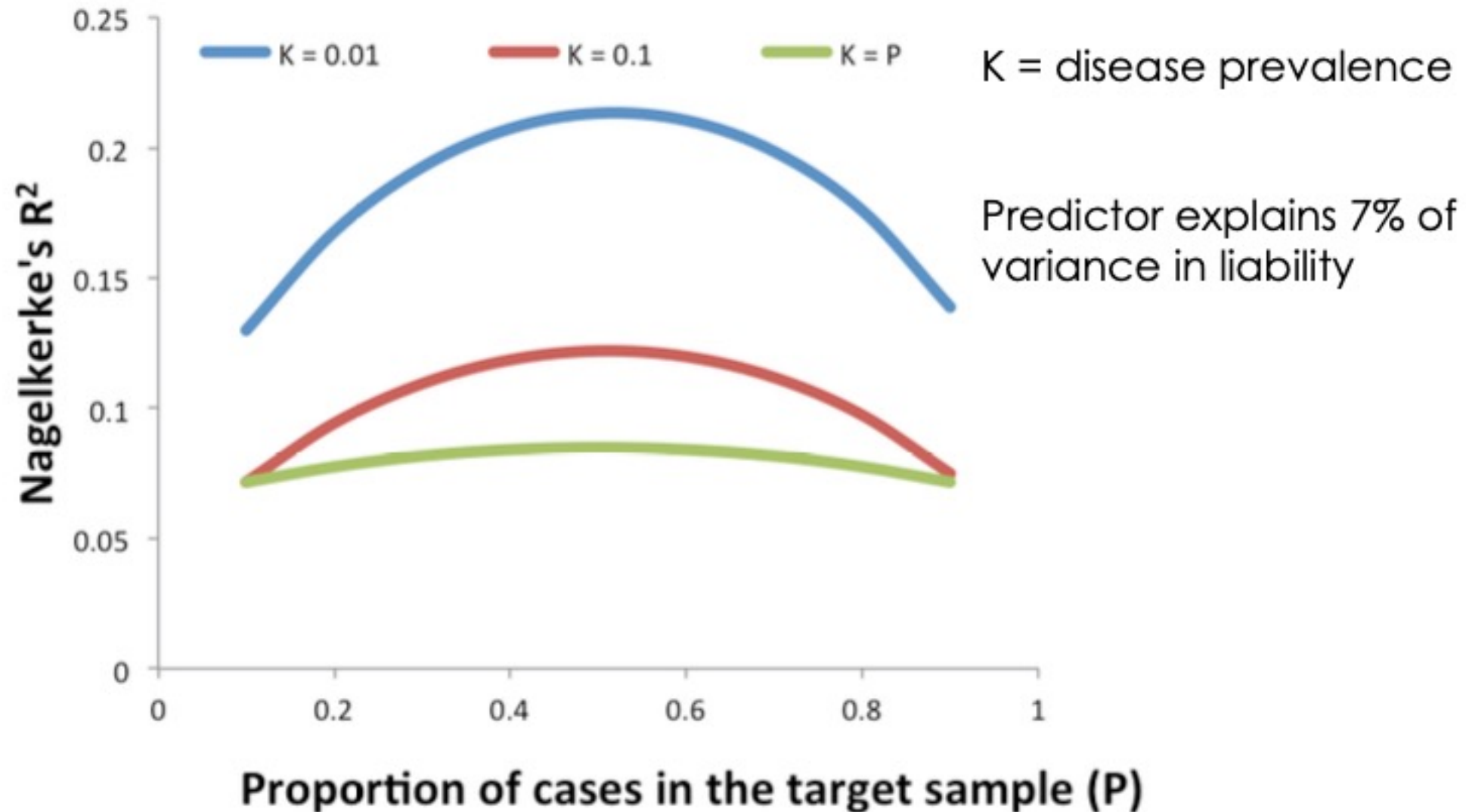
- Nagelkerke's R^2

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

<https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds>

Pseudo R-Squared	Formula	Description
Efron's	$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$ \hat{y} = model predicted probabilities	Efron's mirrors approaches 1 and 3 from the list above—the model residuals are squared, summed, and divided by the total variability in the dependent variable, and this R-squared is also equal to the squared correlation between the predicted values and actual values. When considering Efron's, remember that model residuals from a logistic regression are not comparable to those in OLS. The dependent variable in a logistic regression is not continuous and the predicted value (a probability) is. In OLS, the predicted values and the actual values are both continuous and on the same scale, so their differences are easily interpreted.
McFadden's	$R^2 = 1 - \frac{\ln(L(M_{full}))}{\ln(L(M_{intercept}))}$ M_{full} = Model with predictors $M_{intercept}$ = Model without predictors L = Estimated likelihood	McFadden's mirrors approaches 1 and 2 from the list above. The log likelihood of the intercept model is treated as a total sum of squares, and the log likelihood of the full model is treated as the sum of squared errors (like in approach 1). The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model (like in approach 2). A likelihood falls between 0 and 1, so the log of a likelihood is less than or equal to zero. If a model has a very low likelihood, then the log of the likelihood will have a larger magnitude than the log of a more likely model. Thus, a small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model. If comparing two models on the same data, McFadden's would be higher for the model with the greater likelihood.
McFadden's (adjusted)	$R_{adj}^2 = 1 - \frac{\ln(L(M_{full})) - K}{\ln(L(M_{intercept}))}$ L = Estimated likelihood	McFadden's adjusted mirrors the adjusted R-squared in OLS by penalizing a model for including too many predictors. If the predictors in the model are effective, then the penalty will be small relative to the added information of the predictors. However, if a model contains predictors that do not add sufficiently to the model, then the penalty becomes noticeable and the adjusted R-squared can decrease with the addition of a predictor, even if the R-squared increases slightly. Note that negative McFadden's adjusted R-squared are possible.
Cox & Snell	$R^2 = 1 - \left(\frac{L(M_{intercept})}{L(M_{full})} \right)^{\frac{2}{N}}$	Cox & Snell's mirrors approach 2 from the list above. The ratio of the likelihoods reflects the improvement of the full model over the intercept model (the smaller the ratio, the greater the improvement). Consider the definition of $L(M)$: $L(M)$ is the conditional probability of the dependent variable given the independent variables. If there are N observations in the dataset, then $L(M)$ is the product of N such probabilities. Thus, taking the n^{th} root of the product $L(M)$ provides an estimate of the likelihood of each Y value. Cox & Snell's presents the R-squared as a transformation of the $-2 \ln(L(M_{intercept})/L(M_{full}))$ statistic that is used to determine the convergence of a logistic regression. Note that Cox & Snell's pseudo R-squared has a maximum value that is not 1: if the full model predicts the outcome perfectly and has a likelihood of 1, Cox & Snell's is then $1 - L(M_{intercept})^{2/N}$, which is less than one.
Nagelkerke / Cragg & Uhler's	$R^2 = \frac{1 - \left(\frac{L(M_{intercept})}{L(M_{full})} \right)^{\frac{2}{N}}}{1 - L(M_{intercept})^{2/N}}$	Nagelkerke/Cragg & Uhler's mirrors approach 2 from the list above. It adjusts Cox & Snell's so that the range of possible values extends to 1. To achieve this, the Cox & Snell R-squared is divided by its maximum possible value, $1 - L(M_{intercept})^{2/N}$. Then, if the full model perfectly predicts the outcome and has a likelihood of 1, Nagelkerke/Cragg & Uhler's R-squared = 1. When $L(M_{full}) = L(M_{intercept})$, then $R^2 = 0$.
McKelvey & Zavoina	$R^2 = \frac{\text{Var}(\hat{y}^*)}{\text{Var}(\hat{y}^*) + \text{Var}(\epsilon)}$	McKelvey & Zavoina's mirrors approach 1 from the list above, but its calculations are based on predicting a continuous latent variable underlying the observed 0-1 outcomes in the data. The model predictions of the latent variable can be calculated using the model coefficients (NOT the log-odds) and the predictor variables. McKelvey & Zavoina's also mirrors approach 3. Because of the parallel structure between McKelvey & Zavoina's and OLS R-squareds, we can examine the square root of McKelvey & Zavoina's to arrive at the correlation between the latent continuous variable and the predicted probabilities. Note that, because y^* is not observed, we cannot calculate the variance of the error (the second term in the denominator). It is assumed to be $\pi^2/3$ in logistic models.
Count	$R^2 = \frac{\# \text{Correct}}{\text{Total Count}}$	Count R-Squared does not approach goodness of fit in a way comparable to any OLS approach. It transforms the continuous predicted probabilities into a binary variable on the same scale as the outcome variable (0-1) and then assesses the predictions as correct or incorrect. Count R-Squared treats any record with a predicted probability of .5 or greater as having a predicted outcome of 1 and any record with a predicted probability less than .5 as having a predicted outcome of 0. Then, the predicted 1s that match actual 1s and predicted 0s that match actual 0s are tallied. This is the number of records correctly predicted, given this cutoff point of .5. The R-squared is this correct count divided by the total count.
Adjusted Count	$R^2 = \frac{\text{Correct} - n}{\text{Total} - n}$ n = Count of most frequent outcome	The Adjusted Count R-Square mirrors approach 2 from the list above. This adjustment is unrelated to the number of predictors and is not comparable to the adjustment to OLS or McFadden's R-Squareds. Consider this scenario: if you are asked to predict who in a list of 100 random people is left-handed or right-handed, you could guess that everyone in the list is right-handed and you would be correct for the majority of the list. Your guess could be thought of as a null model. The Adjusted Count R-Squared controls for such a null model. Without knowing anything about the predictors, one could always predict the more common outcome and be right the majority of the time. An effective model should improve on this null model, and so this null model is the baseline for which the Count R-Square is adjusted. The Adjusted Count R-squared then measures the proportion of correct predictions beyond this baseline.

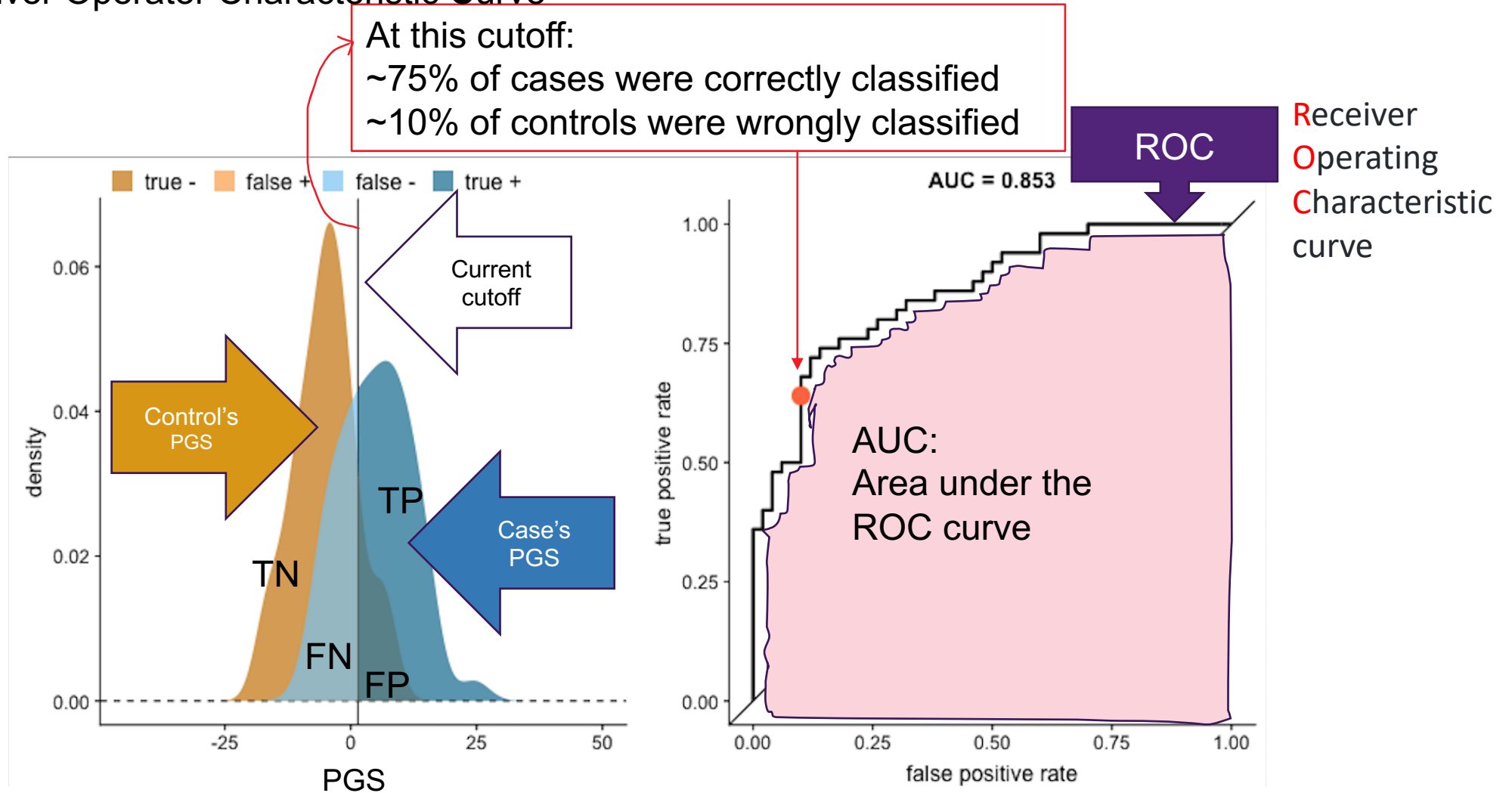
Nagelkerke's R^2 depends on case proportion in the sample



2) AUC

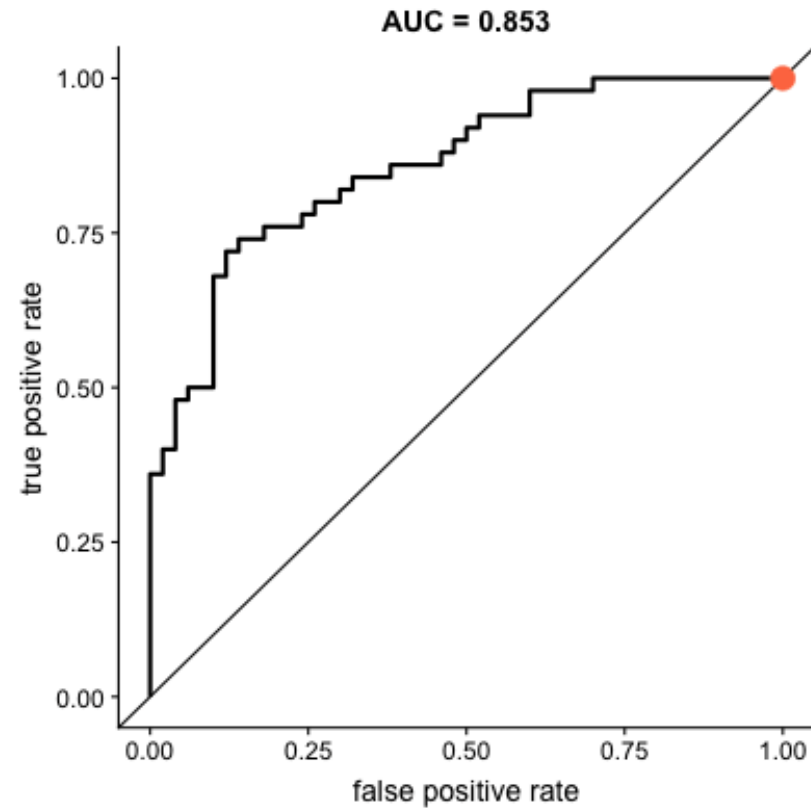
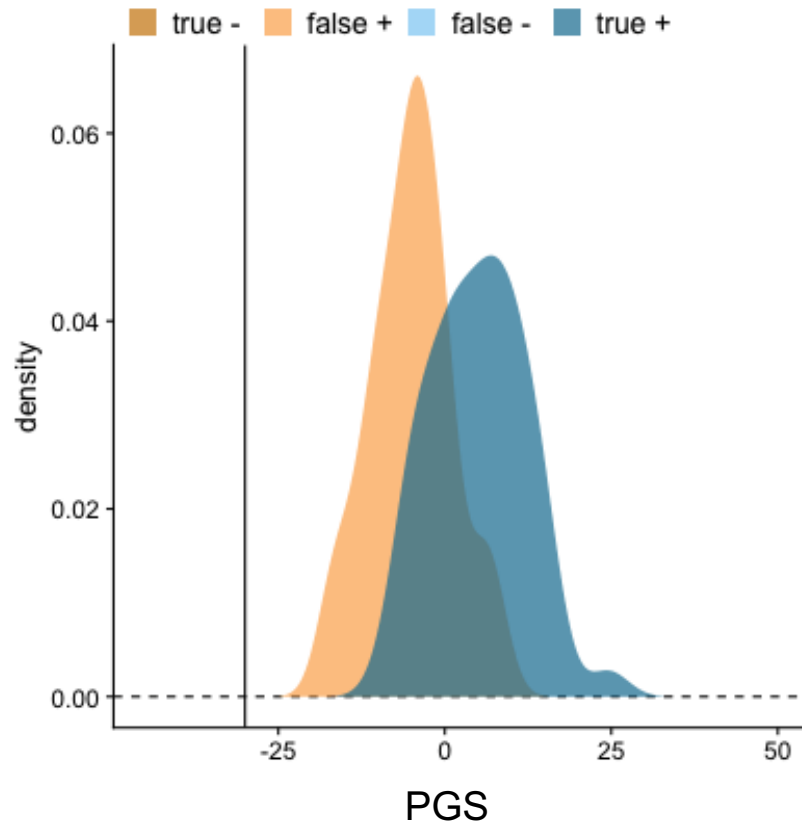
Area Under Receiver Operator Characteristic Curve

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

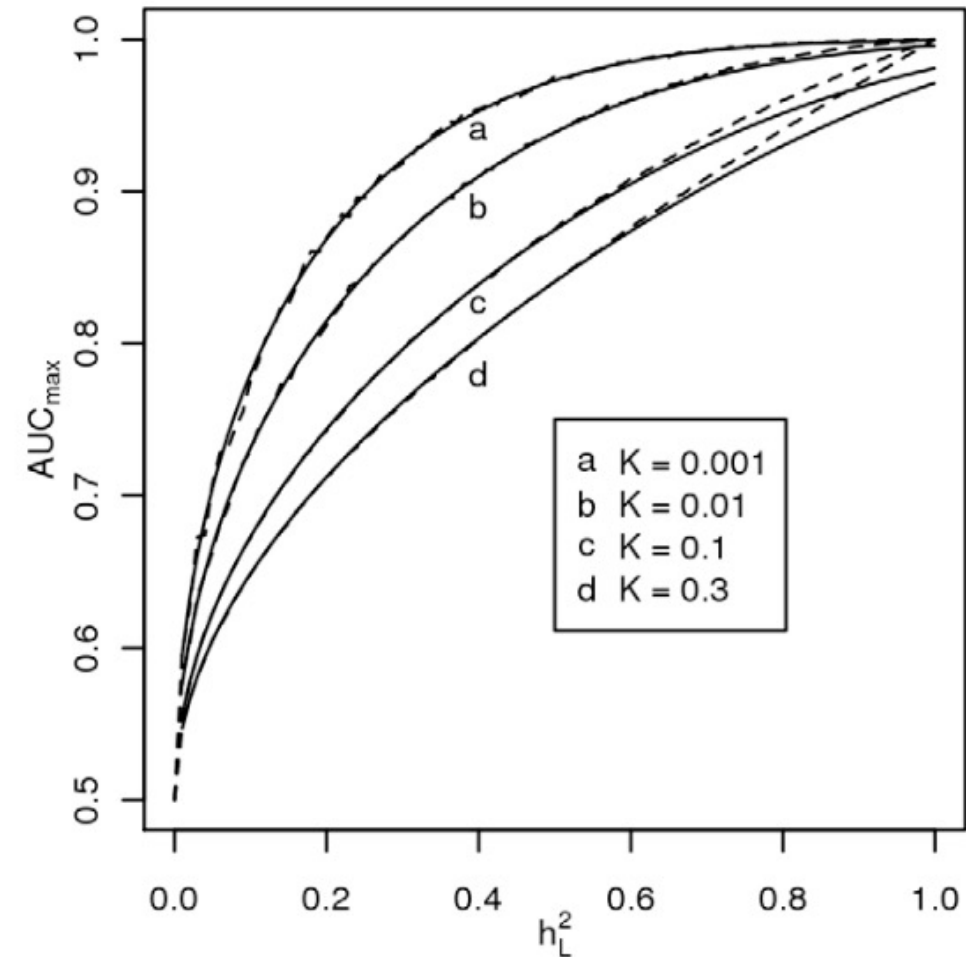
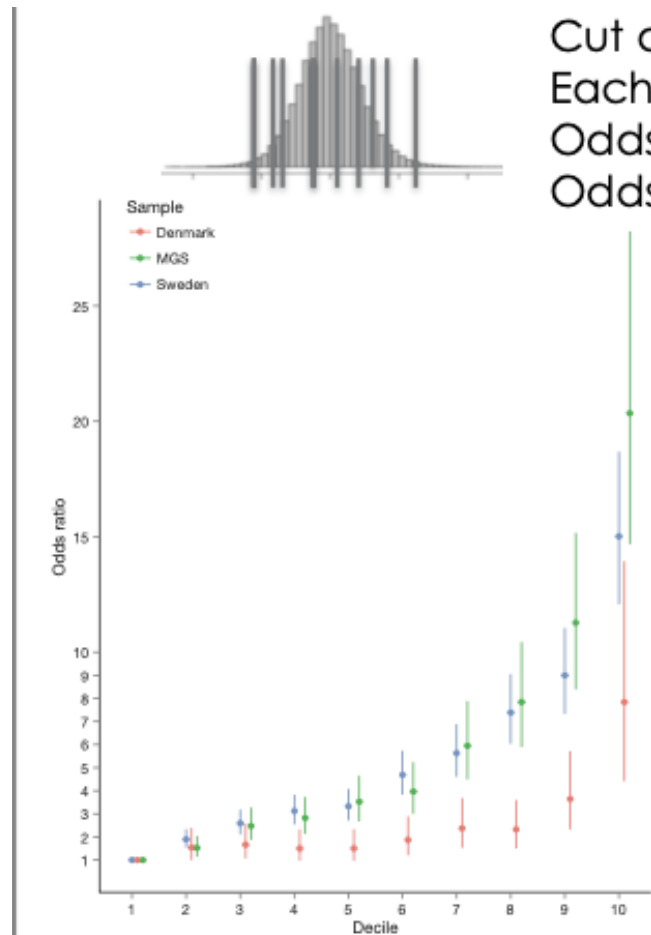


Figure 2. Relationship between maximum AUC (AUC_{max}) from a genomic profile and heritability on the liability scale h_L^2 . For

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 1st decile /Middle decile

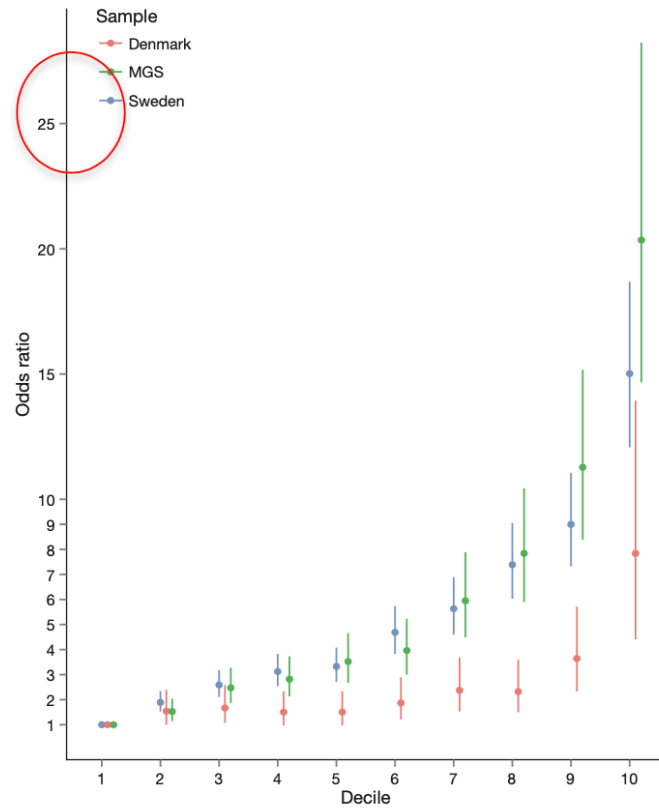
- Good visualisation
- Shows that there could be utility in using high vs low profile risk scores
- But remember case-control samples are 50% cases
- Would look less impressive if a population sample

PGC-SCZ 2014 108 loci Nature

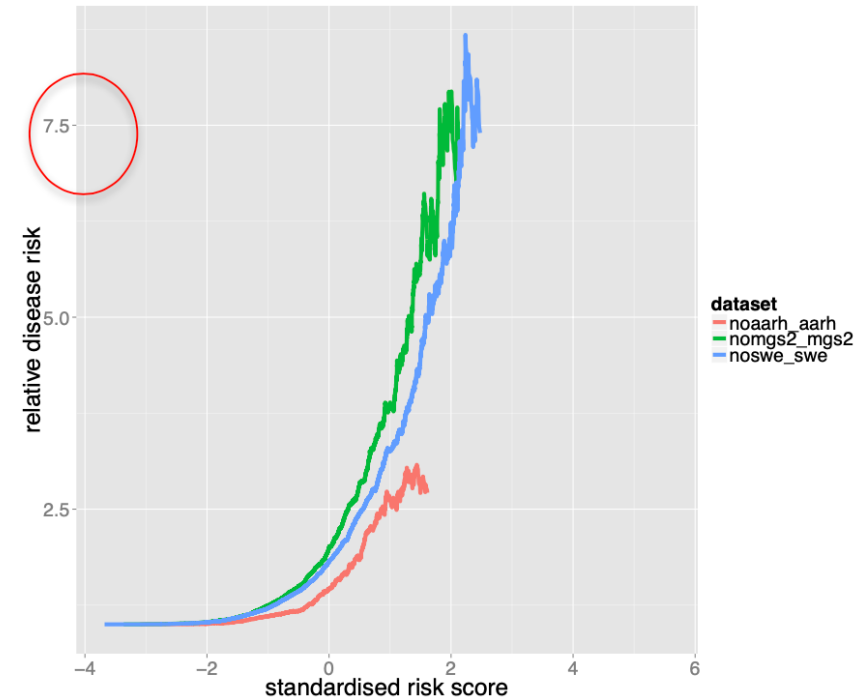
13

3) Odds ratio

In case control samples



Same data scaled to population risk



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

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

P = probability of being case

Toy example:

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

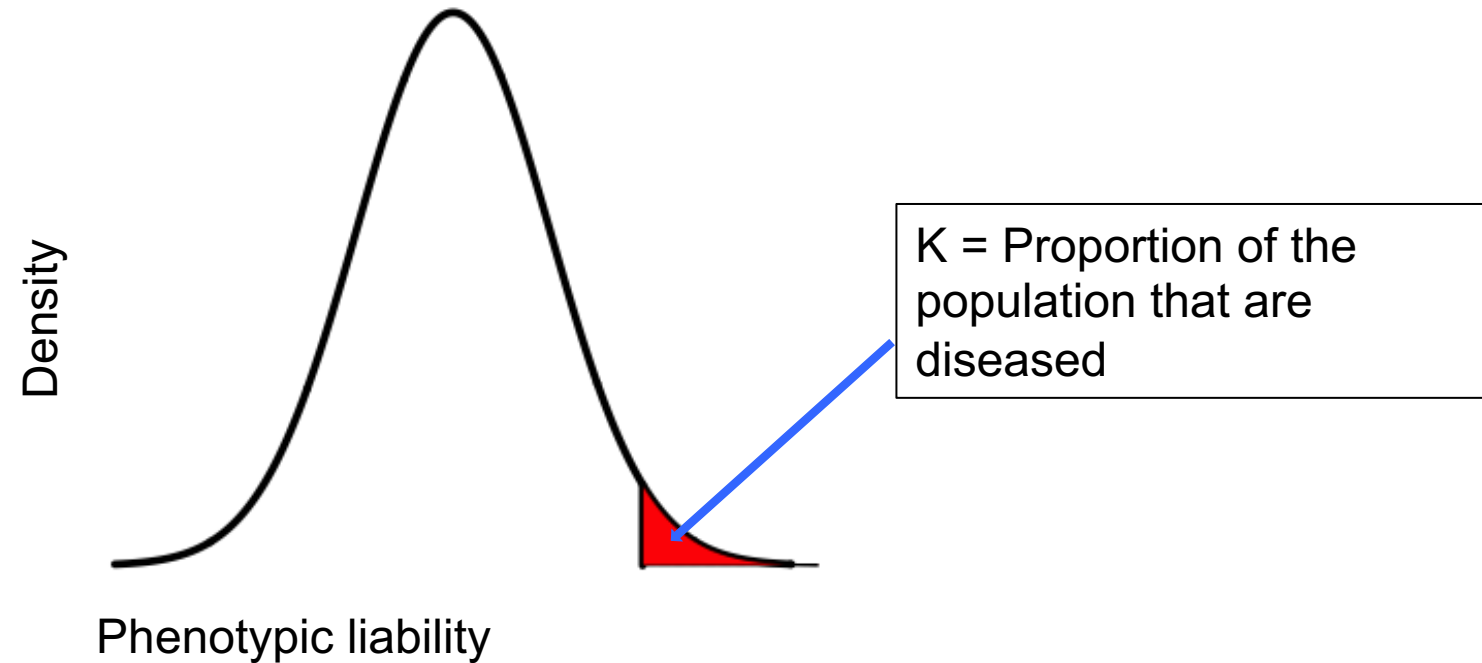
Odds being a case in 1st decile
= 23/103

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

Odds ratio between 10th and 1st decile
= (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

4) R² on liability scale

R² on the liability scale when using ascertained case-control studies

Linear regression; Y are 0s and 1s

Null: Y = cov + e

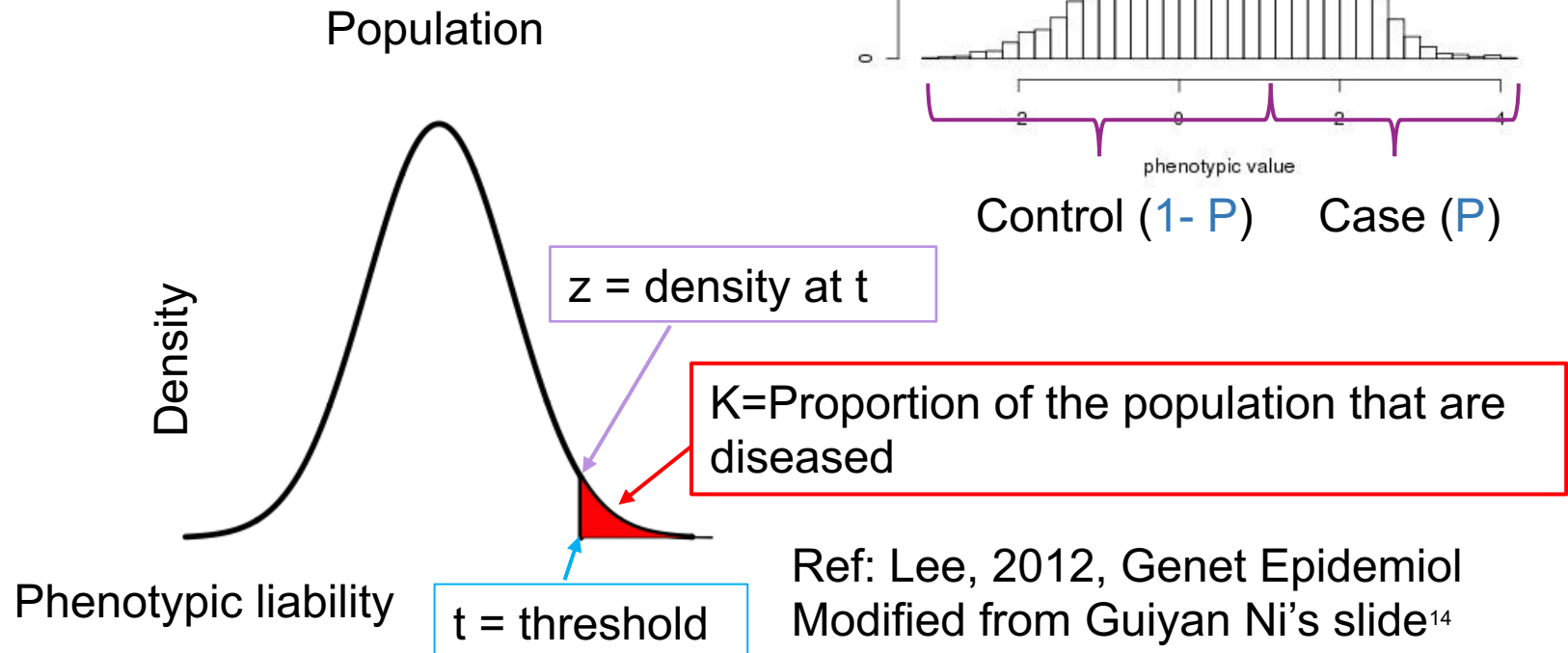
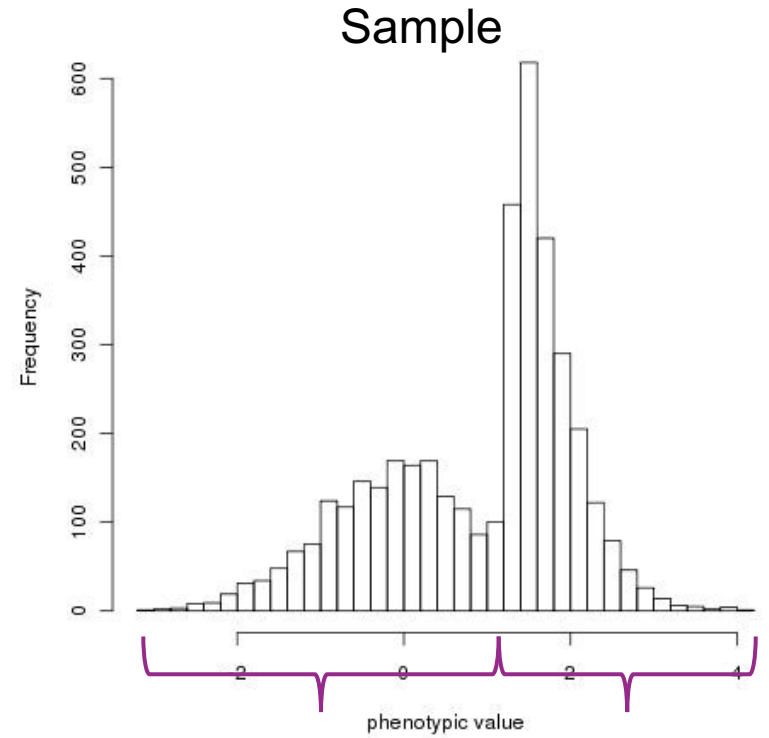
Full: Y = cov + PGS + e

$$R_{l_cc}^2 = \frac{R_{o_cc}^2 * C}{1 + R_{o_cc}^2 * \theta * C}$$

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

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

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



Ref: Lee, 2012, Genet Epidemiol
Modified from Guiyan Ni's slide¹⁴

5) Net reclassification index (NRI)

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 is⁴

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

“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 for CVD

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

“New” model: “Old”+PRS

Time-to-event:

- From assessment to disease (indicate cases) - PRS + traditional risk model
- From birth to disease (age of onset) – PRS alone

Method: Cox proportional hazard analysis

Statistics:

- Hazard ratio per SD
- Harrell's C-index

R package:

“coxph” function in “survival” package

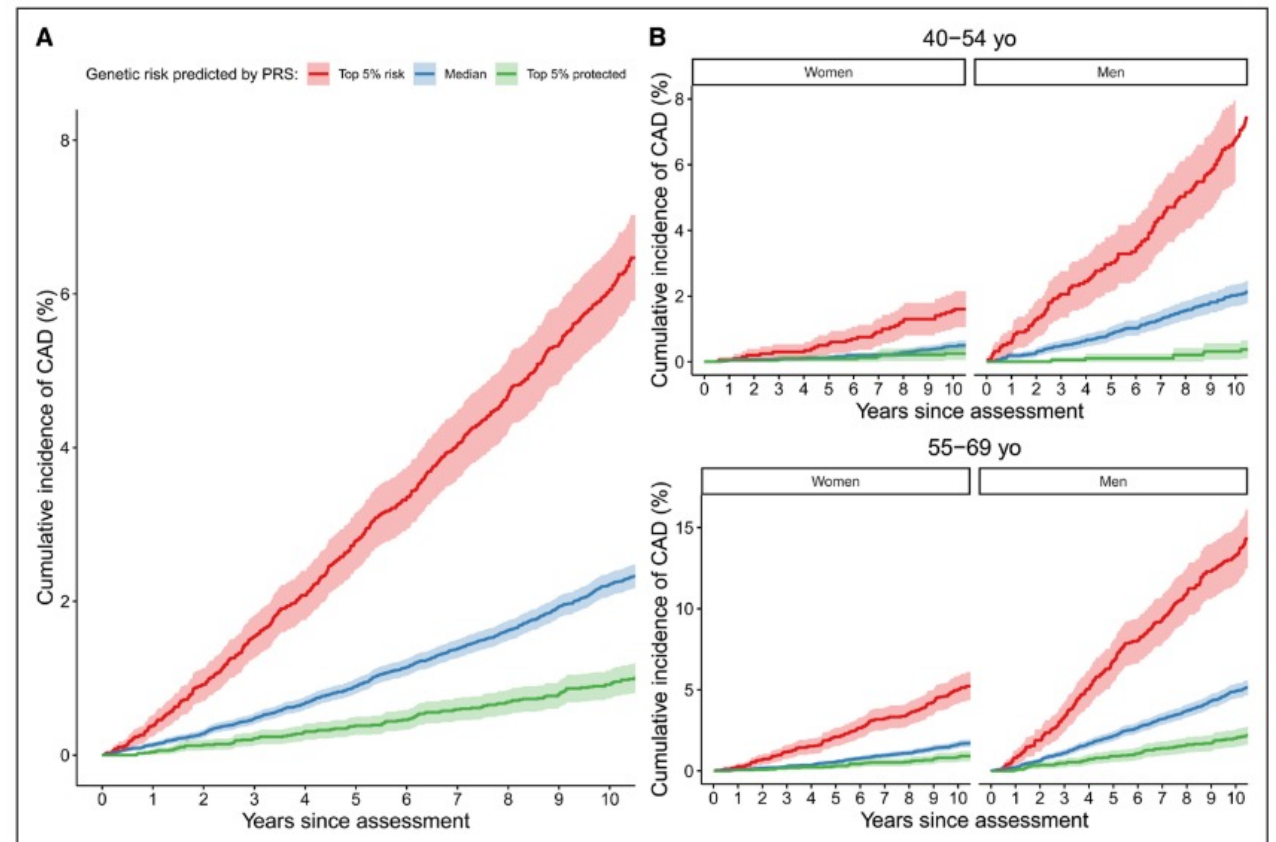


Figure 1. Cumulative incidence of coronary artery disease (CAD) in UK Biobank incident cases in group III.

A, All of group III. **B**, Group III stratified into 4 subgroups according to age (45–54- and 55–69-y-old age ranges) and sex. Individuals are further stratified by polygenic risk score (PRS)-defined risk into the top 5% of PRS risk (red), the median 40% to 60% distribution of risk (blue), and the bottom 5% of risk distribution (green).

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

assume LD-independent

$$y = \sum_i^M b_i x_i + e; \hat{y} = \sum_i^M \hat{b}_i x_i$$

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

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

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

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

- Discovery/Training/Derivation

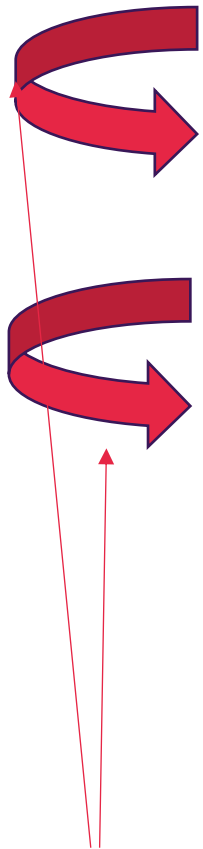
- Estimate the effect sizes (\hat{b}) of SNPs on a trait (y) – GWAS

- 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^{1*}, Stephen Richards^{2*}, Eric A. Stone^{1*}, Antonio Barbadilla^{3*}, Julien F. Ayroles^{1†}, Dianhui Zhu², Sònia Casillas^{3†}, Yi Han², Michael M. Magwire¹, Julie M. Cridland⁴, Mark F. Richardson⁵, Robert R. H. Anholt⁶, Maite Barrón³, Crystal Bess², Kerstin Petra Blankenburg², Mary Anna Carbone¹, David Castellano³, Lesley Chaboub², Laura Duncan¹, Zeke Harris¹, Mehwish Javaid², Joy Christina Jayaseelan², Shalini N. Jhangiani², Katherine W. Jordan¹, Fremiet Lara², Faye Lawrence¹, Sandra L. Lee², Pablo Librado⁷, Raquel S. Linheiro⁵, Richard F. Lyman¹, Aaron J. Mackey⁸, Mala Munidasa², Donna Marie Muzny², Lynne Nazareth², Irene Newsham², Lora Perales², Ling-Ling Pu², Carson Qu², Miquel Ràmia³, Jeffrey G. Reid², Stephanie M. Rollmann^{1†}, Julio Rozas⁷, Nehad Saada², Lavanya Turlapati¹, Kim C. Worley², Yuan-Qing Wu², Akihiko Yamamoto¹, Yiming Zhu², Casey M. Bergman⁵, Kevin R. Thornton⁴, David Mittelman⁹ & Richard A. Gibbs²

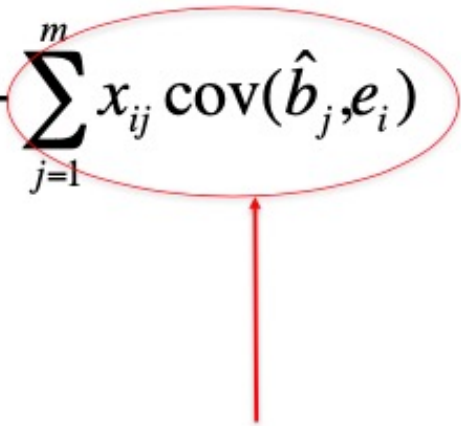
Predicting phenotypes from genotypes

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.)

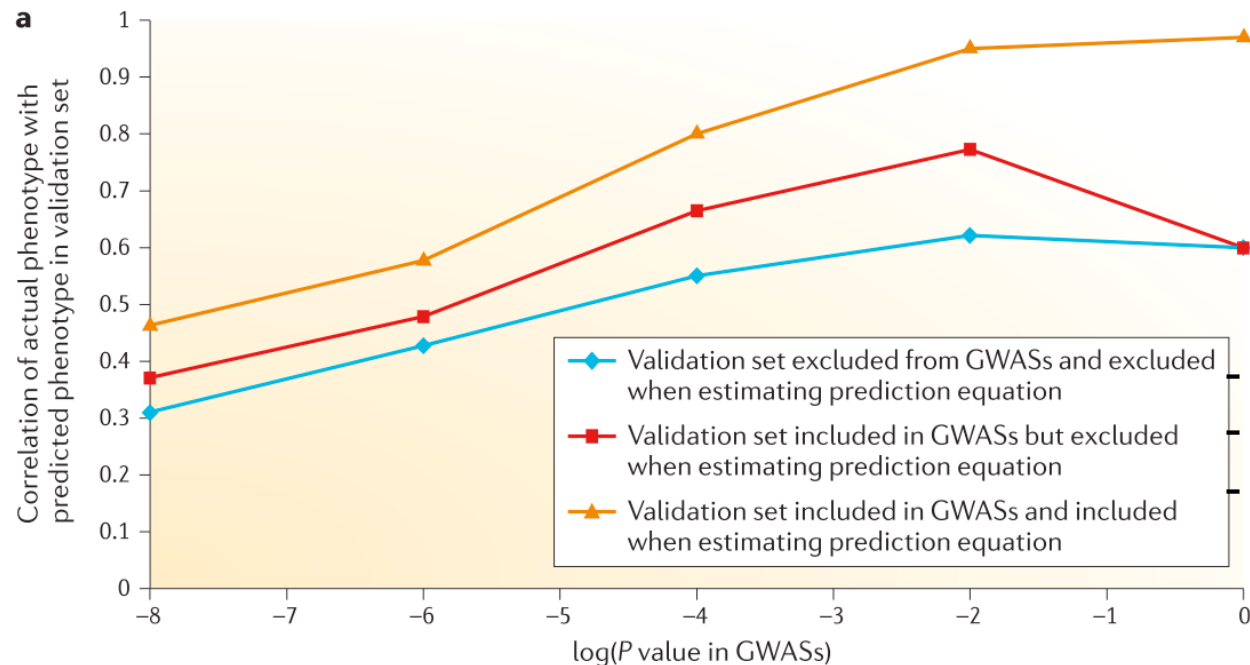
- Overlapping target and discovery sample
- Greater similarity between target and discovery sample (such as relatedness)
 - Cross-validation: not a pitfall, but to be aware

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


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 R2 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)
 - Time-to-event analysis
- 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