Methods and Tutorials for Building Polygenic Risk Scores 2

Course # 140.721

Ziqiao Wang

Department of Biostatistics
Johns Hopkins University

Review of last lecture

- Advanced methods for building PRS based on statistical high-dimensional modeling (machine learning) techniques is an active area of research
- Different categories of methods, model-free, Bayesian, penalized regression methods
  - Can be applied to summary-statistics data available from GWAS, but tuning and validation may need individual level data
- Current PRS are biased toward European origin populations, active research are being done to reduce this difference
- Interpretations and Applications of PRS

Tutorials for Building Polygenic Risk Scores

Evaluating PGS methods: cross-validation and cross-ethnicity performance

AUC, R2, log odds ratio, log hazard ratio

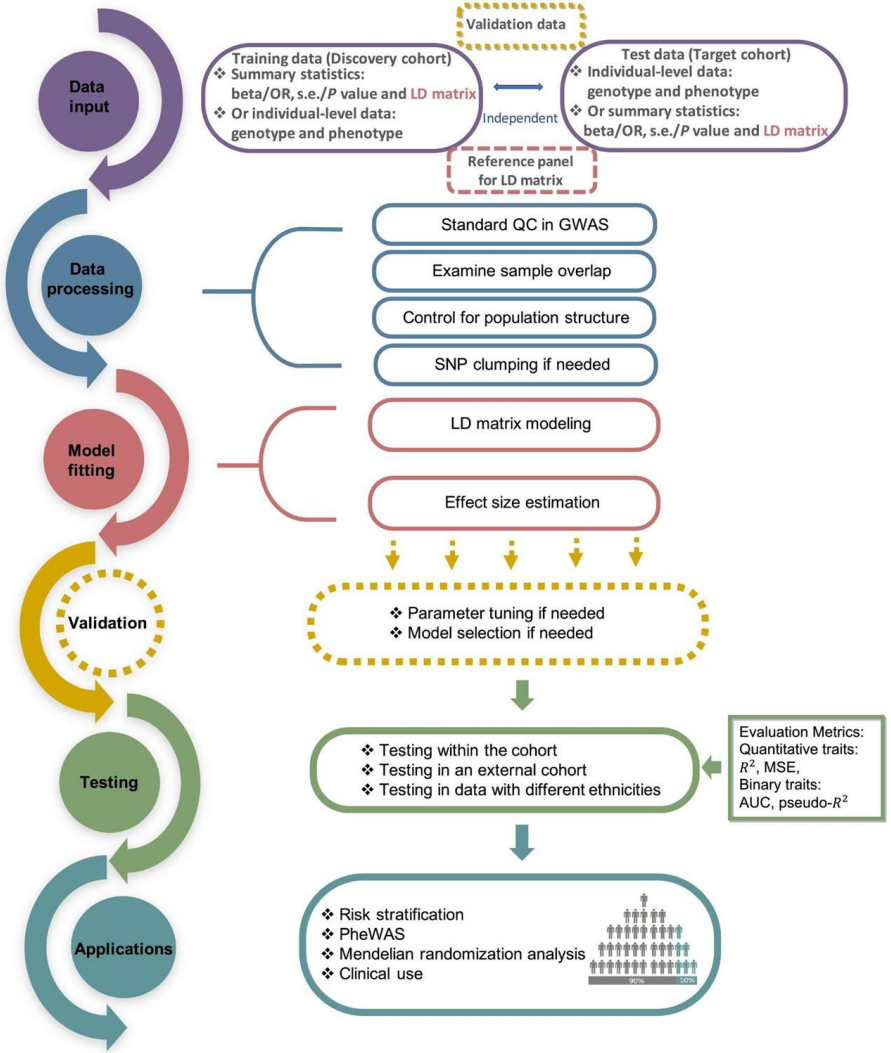
PGSCatalog

R scripts

PLINK Linux scripts

PRS-CS tutorial (download LD panel, etc)

Tutorial: a guide to performing polygenic risk score analyses



Trends in Genetics

Evaluation metrics for binary outcomes (disease risk prediction)

- True positive rate (TPR) and false positive rate (FPR):
  - TPR(m) = P(M>m|D=1) = P(M1 > m) sensitivity
  - FPR(m) = P(M>m|D=0) = P(M0 > m) 1-specificity

Receiver Operating Characteristic (ROC) function:
$ROC(p) = TPR\,[FPR^{-1}(p)],\ p \in (0,1),\ \ AUC = \int_0^1 ROC(p)dp$: area under ROC curve

- AUC: Pr(M1 > M0), Suppose the two biomarker variables M0 and M1 are independent. Then AUC= probability that marker value for a randomly selected case exceeds that for a randomly selected control.
- log odds ratio, 95% CI

Evaluation metrics for continuous outcomes (traits)

- R-Squared (R^2 or the coefficient of determination)
  - the proportion of variance in the dependent variable (true outcome) that can be explained by the independent variable (predicted outcome) in a linear regression
- Incremental R2
  - Fit a linear regression model regress Y on a set of covariates, such as age, BMI, sex, genetic PCs 1-10
  - Fit a second linear regression model regress Y on the same set of covariates + PRS
  - Calculate the difference of R2 from two regressions
- Alternatively, regress PRS on a set of covariates and take the residuals, then regress Y on the residuals using a simple linear regression and report the R2

Evaluation metrics for time-to-event outcomes

- log hazard ratio
- Time-dependent ROC and AUC
- Cumulative sensitivity and dynamic specificity

$$\textbf{Sensitivity}: \; TPR_t^C(m) = P(M > m | T \leq t)$$

$$1-\textbf{Specificity}: \; FPR_t^D = P(M > m | T > t)$$

$$ROC_t^{C/D}(p) = TPR_t^C\{[FPR_t^D]^{-1}(p)\}$$

- Incident sensitivity: TPR(m) = P(M > m |T = t)
- Harrel's C-index (concordant probability)
  - P(Mi < Mj | Ti > Tj), Xi = min(Ti,Ci) is the observed survival time

Tutorials for building PRS

- https://github.com/ziqiaow/PRS-tutorial