PRS Reporting

Goal

Our goal is to calculate disease-specific patient-level PRS based on GWAS summary statistics for different disease

Todo List

Input:

Select traits & Download GWAS ssumary statistics from GWAS catlog (prepare GWAS summary statistics to be in the same form)

Prepare UKB synthetic genotype/ phenotype data

(Individual-level) Compute PRS:

Compute PRS based on various p-value thresholds and GWAS summary statistics

(Cohort-level) Build predictive models:

Build a predictive model using generated PRS scores on UKB synthetic data

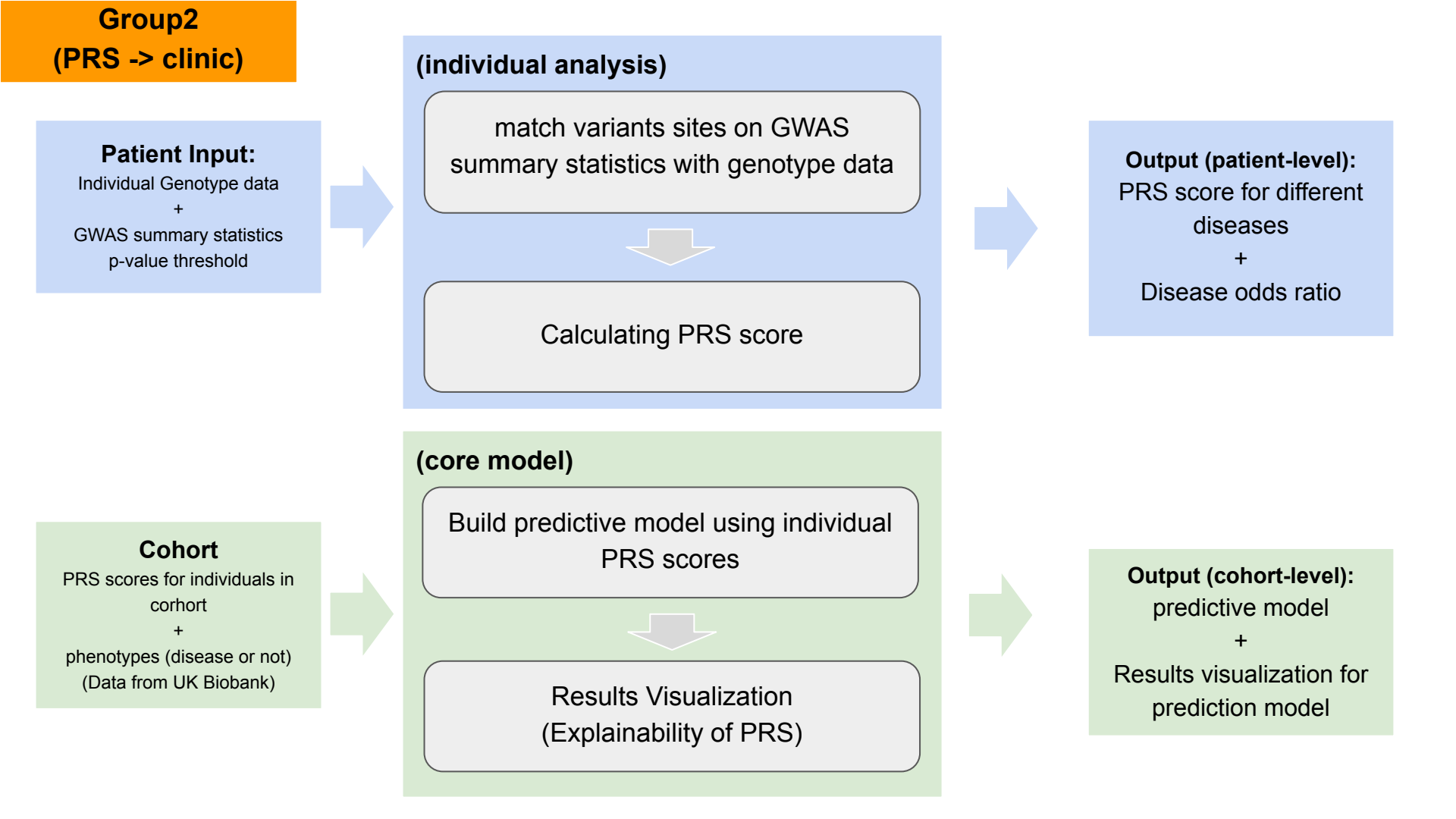
Input/ Output

Input: GWAS summary statistics; individual genotype variant data (vcf/ plink bed/ ...), disease phenotype?
Output: 1. PRS scores; 2. Predictive model 3. Some visualizations

Feedback from UX team

What is the model output? Related likelihood? Some more clear output information(scores, images, pdf)
How are these information given to the other team?(database)
How do we associate the phenotype and PRS? (can customize it based on gene information, probability for other disease, sample selection for clinical trials), Link to our github.
###

GWAS summary statistics preparation


Group2
(PRS -> clinic)
Patient Input:
Individual Genotype data
+
GWAS summary statistics
p-value threshold
(individual analysis)
match variants sites on GWAS summary statistics with genotype data
Calculating PRS score
Output (patient-level):
PRS score for different diseases
+
Disease odds ratio
Cohort
PRS scores for individuals in corhort
+
phenotypes (disease or not)
(Data from UK Biobank)
(core model)
Build predictive model using individual PRS scores
Results Visualization
(Explainability of PRS)
Output (cohort-level):
predictive model
+
Results visualization for prediction model

Just for reference

Methods of construction [edit]

applied to Type 2 Diabetes in humans. Individuals with Type 2 diabetes (white bars) have a higher score than controls (black bars).[18]

A polygenic score (PGS) is constructed from the "weights" derived from a genome-wide association study (GWAS), or from some form of machine learning algorithm. In a GWAS, a set of genetic markers (usually SNPs) is genotyped on a training sample, and effect sizes are estimated for each marker's association with the trait of interest. These weights are then used to assign individualized polygenic scores in an independent replication sample.[1] The estimated score, $\hat{S}$, generally follows the form

$$\hat{S} = \sum_{j=1}^{m} X_j \hat{\beta}_j,$$

where the $\hat{S}$ of an individual is equal to the weighted sum of the individual's marker genotypes, X_j, at m SNPs.[1] Weights are estimated using some form of regression analysis. Because the number of genomic variants is usually larger than the sample size, one cannot use OLS multiple regression ($p > n$ problem[23][24]). Researchers have proposed various methodologies that deal with this problem as well as how to generate the weights of the SNPs, $\hat{\beta}_j$, and how to determine which m SNPs should be included.