Estimating PRS using PRSice-2 software

Muhammad Shoaib

24/11/2021

(1) Usage of PRSice-2 and input files required by the software

In order to estimate PRS using PRSice-2, we need following input files in the correct format

(A) Base file (GWAS file with effect sizes of each SNP)

In our analysis, we used external GWAS data available on the Diagram website https://diagram-consortium.org/downloads.html. We used DIAGRAM 1000G GWAS meta-analysis Stage 1 Summary statistics as published in Scott et al paper (2017). There were 12056346 varaints in the external GWAS, but UKBB had 6588901 variants. The number of common SNPs were, 6468019. After considering Multialleleic SNPs with same SNP positions and using only unique positions, the SNPs remained for analysis were, **5965221** The format of external GWAS file is,

Table 1: External GWAS (DIAGRAM)

Chr.Position	Allele1	Allele2	Effect	StdErr	P.value	TotalSampleSize
5:85928892	T	C	-0.013	0.026	0.61	158186
11:107819621	A	C	-0.071	0.170	0.67	124696

After, splitting the first column,

Table 2: Modified External GWAS (DIAGRAM)

Chr	Pos	Allele1	Allele2	Effect	StdErr	P.value	TotalSampleSize
5	85928892	T	C	-0.013	0.026	0.61	158186
11	107819621	A	C	-0.071	0.170	0.67	124696

In order to find the common SNPs between external GWAS and UKBB, we first combined all the 22 bim files of T2train-control60-splitted (Chk github, Part B of file, PRS_work_flow.pdf) using 'cat' function in linux. we got a single file with rsIDs available in UKBB,

Table 3: UKBB Variants

Chr	rsID	V3	Pos	ref_allele	effect_allele
1	rs3115860	0	753405	A	C
1	rs3131970	0	753425	C	T

We combined external GWAS and UKBB variant file using inner_join function in R based on 'Pos' column (Table 2 & 3). The base file of PRSice-2 is ready for the use. It is a subsetted Diagram GWAS with all rsIDs that are avaialble in UKBB variants. Here A1 is effect allele and A2 ref-allele

Table 4: Base file for PRSice-2

CHR	rsID	BP	A1	A2	BETA	StdErr	Pvalue
1	rs3115860	753405	A	C	0.042	0.023	0.066
1	rs2073813	753541	A	G	-0.036	0.024	0.130

(B) Target file

In the target file required by PRSice-2 we have 60% training data of cases (T2D) and controls. These are binary plink files and we got them earlier (Chk github, Part B of file, PRS_work_flow.pdf) with prefix T2train-control60-splitted (*bed,bim,fam*)

(C) Covariate file

Previously, we extracted the covariate data for 60% T2D cases and controls (Chk github, Part c of file, PRS_work_flow.pdf).

Table 5: Covariate file for PRSice-2

IID	FID	Sex	Age	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
1000019	1000019	1	69	-10.9392	5.49112	-0.502779	-3.21233	-3.009120	-0.062401	0.571605	-1.56643	-0.630085	-0.70346
1000832	1000832	1	69	-15.1206	2.14055	-1.010050	4.35941	-0.114388	1.304450	-1.040090	-1.49936	0.281777	1.56311

(D) Phenotype file

Another file required by the PRSice-2 is the phenotype file. FIDs and IIDs of target file (60% training data of cases and controls) are in the first and second columns respectively, whereas, third column has binary phenotypes (cases =2 & controls=1)

Table 6: Phenotype file for PRSice-2

FID	IID	Pheno
1000832	1000832	2
1001013	1001013	2
4017076	4017076	1
4017099	4017099	1

(2) PRSice-2 code in linux

PRSice-2 can be downloaded from https://www.prsice.info/. The important details can be found at https://choishingwan.github.io/PRS-Tutorial/plink/. Two main files required are PRSice.R and PRSice_linux. Other required files to run the analysis are in the downloaded PRSice directory. This is how, our analysis for the training data was executed,

```
chmod +x PRSice_linux
module load r/4.0.5
Rscript PRSice.R\
    --prsice PRSice_linux\
    --dir /home/shoaib/projects/def-gsarah/shoaib/base-files/T2test/PRSice\
    --base T2baseNoBMI.uniq.txt\
    --target contn.analysis.PRS#\
    --pheno T2.train.control60.txt\
    --cov covariates.cont60.txt\
    --binary-target T\
    --snp rsID\
    --chr CHR\
    --bp BP\
    --A1 A1\
    --A2 A2\
    --stat BETA\
    --pvalue Pvalue\
    --extract PRSice.valid\
    --print-snp\
    --score sum\
    --out PRSice-res\
```

(3) Output files and graphs

Software generated multiple output files and graphs once analysis is completed

a. Bar plot and a high resolution plot
b. A log file with all the processing information
c. A PRSice.best file with best PRS values of all target individuals (cases/controls)
d. A PRSice.summary file with details of best model
e. A PRSice.prsice file which has multiple P-thresholds (between 0 and 1) from the base file and corresponding R2 (phenotypic variance between cases and controls) and number of SNPs lying within every p-threshold value

(4) T2D PRS results

(i) Finding best model using training data

Training sample size	187786
T2D Cases	9171
Controls	178615
Males	86770
Females	101016
SNPs before clump	4654745
SNPs after clump	280722

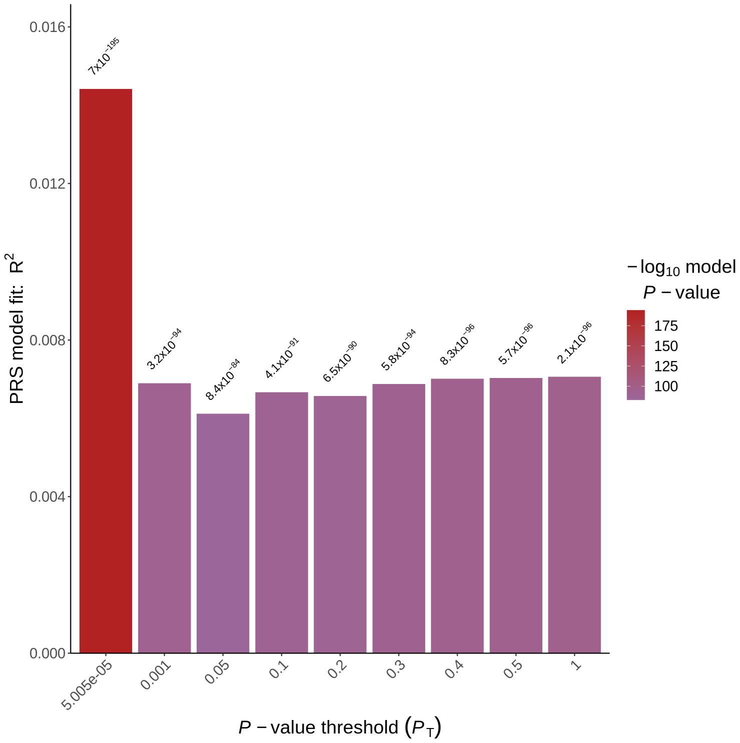

Figure 1: Phenotypic variance R2 at various P-value thresholds

Figure 1 above indicates higher phenotypic variance (R2) between cases and controls (training data) for GWAS p-value threshold = 5e-5

P-value thresholds explaining the highest phenotypic variance

Table 8: Top P-value thresholds with highest R2

Set	Threshold	R2	Coefficient	Standard.Error	Num_SNP
Base	0.00000005	0.0140411	0.550039	0.01864040	60
Base	0.00005005	0.0144168	0.321887	0.01080870	333
Base	0.00010005	0.0142172	0.281208	0.00950508	468
Base	0.00015005	0.0142952	0.266953	0.00899881	551

Graphical illustration of PRS in training data (cases vs controls)

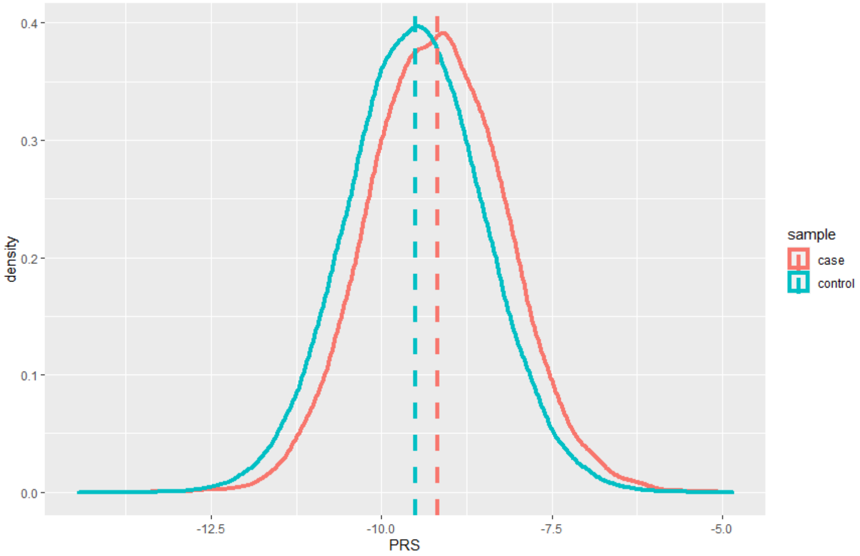

Figure 2: Density plot of PRS

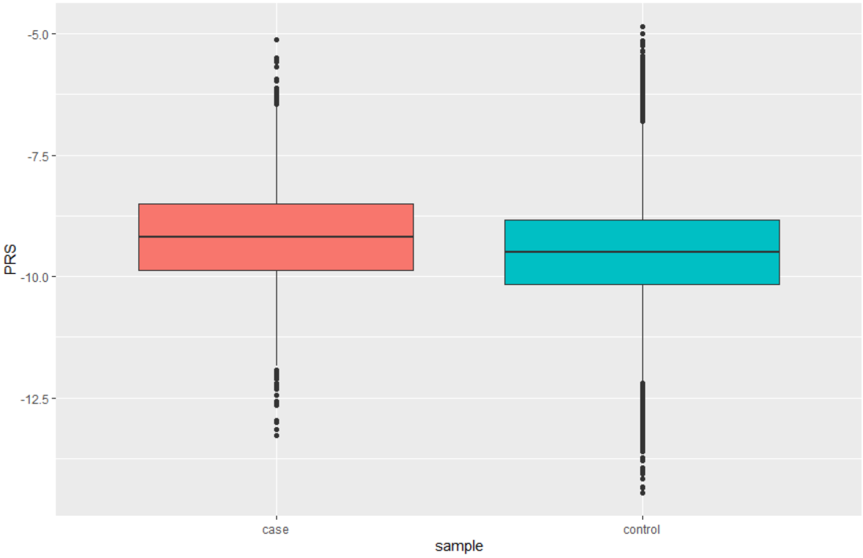

Figure 3: Boxplots of PRS

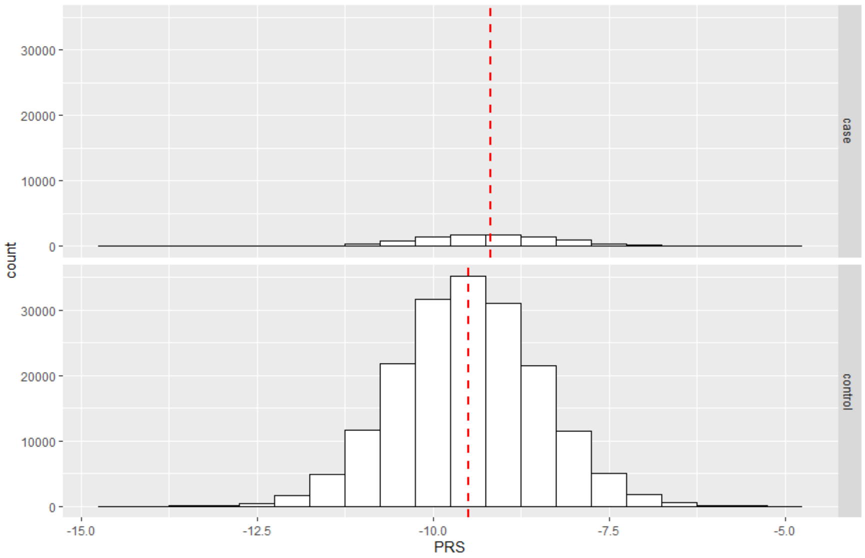

Figure 4: Histograms of PRS

(5) Testing of model in UKBB independent test data

Testing sample size	127572
T2D 40% test Cases	6114
T1D All cases	2383
Controls 40%	119075

The PRSice-2 tool can be used for testing the results in the test data. The p-value threshold with the highest R2 value was actually applied to calculate PRS in test data. The base file is obtained by subsetting the main DIAGRAM GWAS file and only those SNPs were retained which remained after clumping during training data analysis. The clumped rsIDs are avaialble in PRSice.SNP file. The target data is test data (40% T2D, All T1D and 40% controls)

```
module load r/4.0.5
Rscript PRSice.R\
    --prsice PRSice_linux\
    --dir /home/shoaib/projects/def-gsarah/shoaib/base-files/T2test/PRSice\
    --base Diag.test.txt\
    --target con40T1all.T240-#\
    --snp rsID\
    --chr CHR\
    --bp BP\
    --A1 A1\
    --A2 A2\
    --stat BETA\
    --pvalue Pvalue\
    --no-clump\
    --bar-levels 5.005e-05\
    --no-full\
    --fastscore\
    --no-regress\
    --out PRSice-test-res\
```

(6) Box plot of T2D PRS

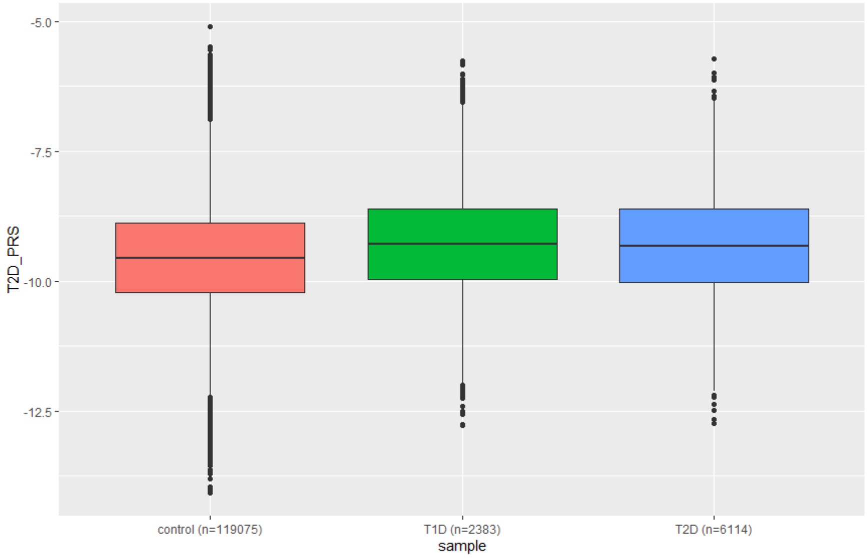

Figure 5: T2D PRS

(7) T1D PRS results

For T1D, the external GWAS was downloaded from the website, https://datadryad.org/stash/dataset/doi:10.5061/dryad.ns8q3. After extracting useful columns (Chr, Pos, A0, A1 which is effect allele, rsID, BetaMeta, PMeta), we got the base file for PRSice 2. Total SNPs in the external GWAS file were, *8537525*, and PRSice used *3558695* SNPs for analysis. *262642* SNPs remained after clumping.

(i) Finding best model using training data

Training sample size	180,044
T1D 60% Cases	1430
Controls 60%	178614
Males	81941
Females	98103

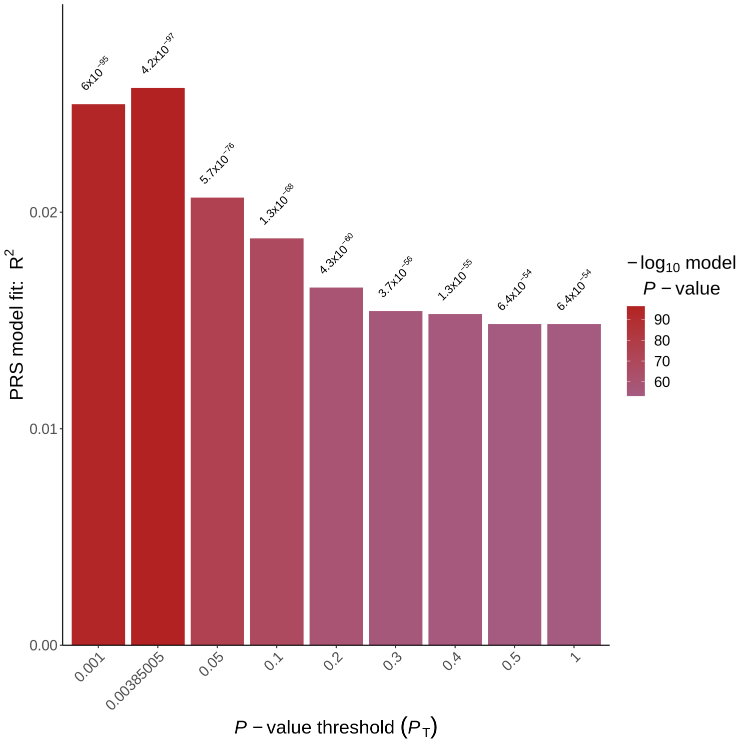

Figure 6: Phenotypic variance R2 at various P-value thresholds

Figure above indicates higher phenotypic variance (R2) between cases and controls (training data) for GWAS p-value threshold = 0.003

(ii) Testing of model in UKBB independent test data

Testing sample size	135312
T1D 40% test Cases	953
T2D All cases	15284
Controls 40%	119075

(iii) Box plot of T1D PRS

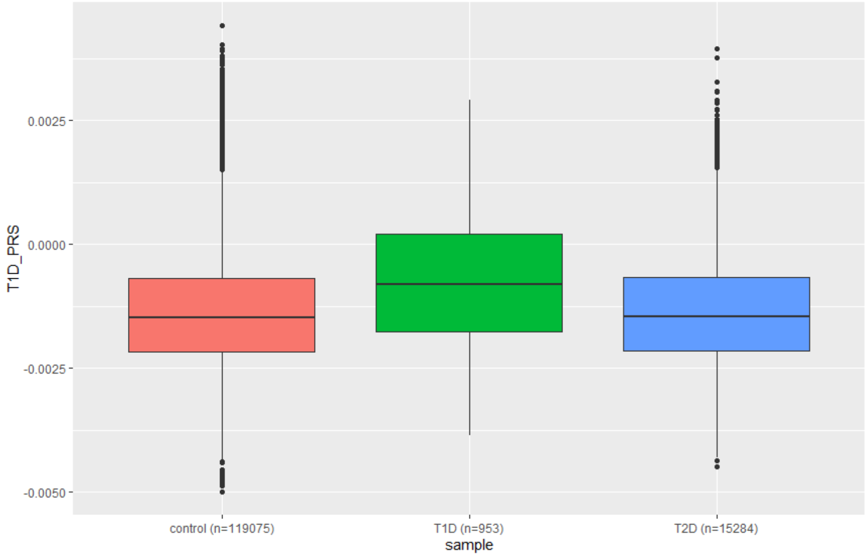

Figure 7: T1D PRS