Fine-tuning Polygenic Risk Scores with GWAS Summary Statistics

Zijie Zhao[1,*], Yanyao Yi[2,*], Jie Song[2], Yuchang Wu[1], Xiaoyuan Zhong[3], Yupei Lin[3], Timothy J. Hohman[4,5], Jason Fletcher[6,7,8], Qiongshi Lu[1,2,8,#]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI
[2] Department of Statistics, University of Wisconsin-Madison, WI
[3] University of Wisconsin-Madison, Madison, WI
[4] Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Vanderbilt University School of Medicine, Nashville, TN
[5] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN
[6] La Follette School of Public Affairs, University of Wisconsin-Madison, Madison, WI
[7] Department of Sociology, University of Wisconsin-Madison, Madison, WI
[8] Center for Demography of Health and Aging, University of Wisconsin-Madison, Madison, WI

Background

- Most PRS models have tuning parameters. These parameters need to be properly selected in applications. In practice, they are typically selected in one of the two ways:
  - ➢ Cross-validation on the training GWAS samples
  - ➢ Tune the model on a validation set independent from the training GWAS
- What if all you have is a GWAS summary statistics file? We introduce PUMAS (Parameter-tuning Using Marginal Association Statistics), a general statistical framework to fine-tune PRS models with GWAS summary statistics.

Methods

- There are two key steps in the PUMAS model-tuning framework (Figure 1)
  - ➢ Simulate sumstats for a subset of samples using the complete sumstats file
  - ➢ Evaluate model performance using a validation set of GWAS sumstats
- Simulating sumstats for a subset of samples

$$x^{(tr)\top} y^{(tr)} \mid x^\top y \sim \mathbf{N}\left(\frac{(N-n)}{N} x^\top y, \frac{(N-n)n}{N}\Sigma\right)$$

$$x^{(v)\top} y^{(v)} = x^\top y - x^{(tr)\top} y^{(tr)}$$

$$\hat{\beta}_j^{(tr)} = \left[(N-n)\hat{\sigma}_j^2\right]^{-1} x_j^{(tr)\top} y^{(tr)}$$

- Calculating predictive R^2 on the validation sumstats

$$\hat{R}^2 = \frac{(\frac{1}{n}\sum_{j=1}^{p} \omega_j x_j^{(v)^T} y^{(v)})^2}{N \max_j [SE(\hat{\beta}_j)^2 \hat{\sigma}_j^2][\omega^T D \omega]}$$

Data

- Education attainment (EA): [1] EA3 GWAS from SSGAC with HRS, Addhealth, WLS, 23&me removed (N=742,903). [2] HRS samples with EUR ancestry (N=10,214). [3] AddHealth samples with EUR ancestry (N=4,775).
- Alzheimer's disease (AD): [1] IGAP 2013 GWAS stage-I analysis (N=54,162). [2] ADGC samples not used in IGAP 2013 (N=7,050). [3] UKBB GWAS with an AD-proxy phenotype (N=355,583).

Results

- PUMAS demonstrates highly consistent results compared with external validations on real GWAS summary statistics under various genetic architecture.
- PUMAS delivers immediate benefits to downstream analysis using PRS as inputs.

Have tuning parameters in your polygenic risk score model? We can perform cross-validation on GWAS summary statistics!

Figure 1. PUMAS workflow. (A) Traditional model tuning approach based on individual-level data **(B)** Model tuning based on summary statistics.

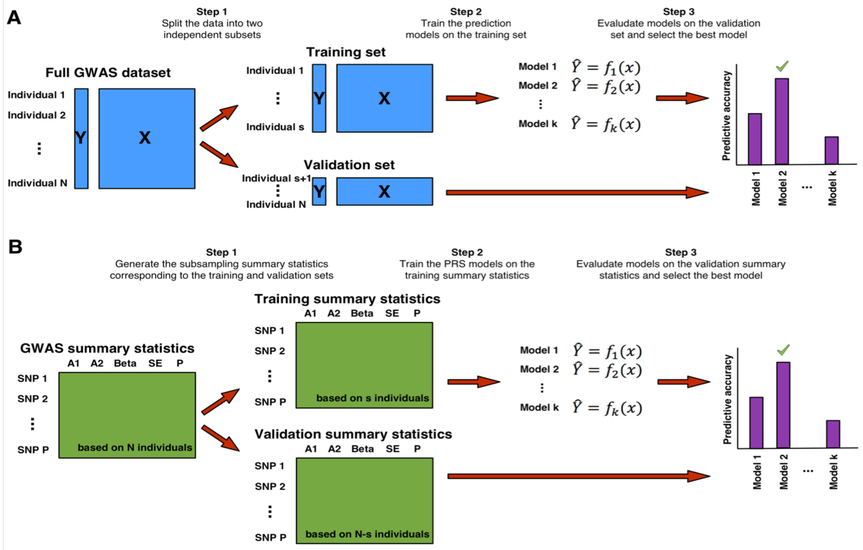


Figure 2. Simulation results. (**A** and **C**) PUMAS (**B** and **D**) Repeated learning based on individual-level data. Heritability (h^2)=0.2(**A-B**)/0.8(**C-D**), number of causal variants (m)=50/1k/4k, total number of variants (M)=5k, sample size (N)=100k(**A-B**)/20k(**C-D**).

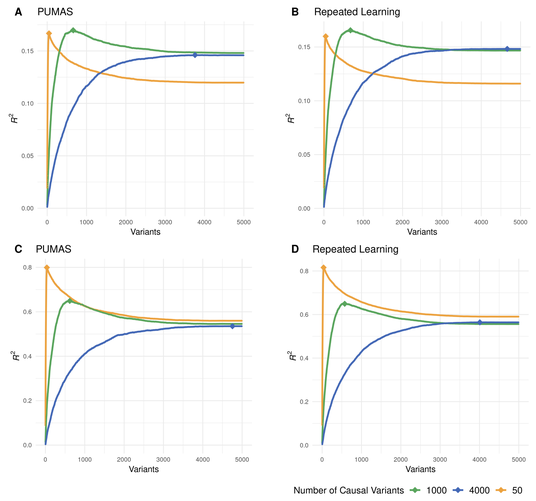


Figure 3. External validation (A-B) educational attainment **(C-D)** Alzheimer's disease. (**A** and **C**) PUMAS (**B** and **D**) PRS evaluated on external validation datasets.

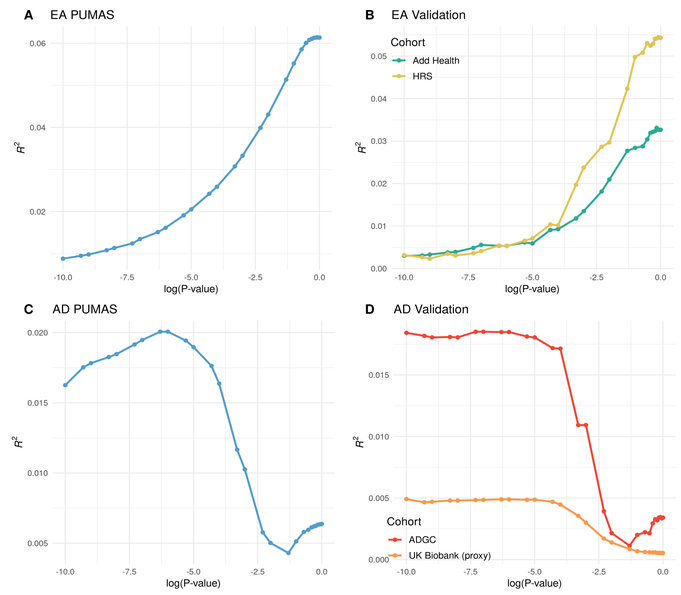


Figure 4. Simulation on WTCCC genotype. (A-B) simple PRS **(C-D)** LDpred PRS

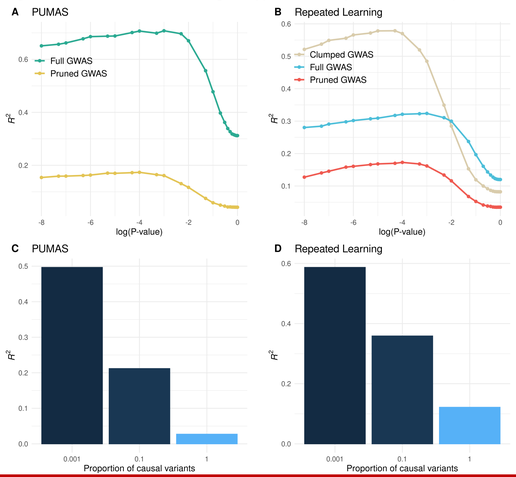


Figure 5. An atlas of optimized PRSs for complex diseases and traits. 45 diseases/traits with optimized R2 > 0.005 are included. X-axis: Optimal P-value cutoff of PRSs; Y-axis: Predictive R^2.

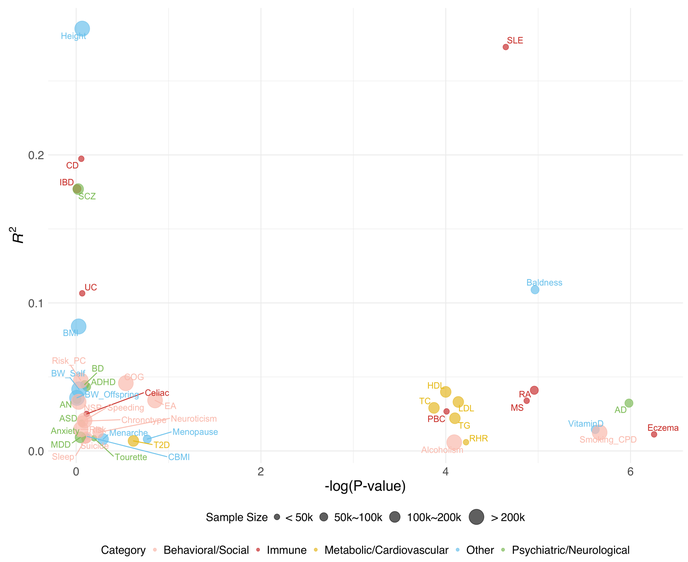
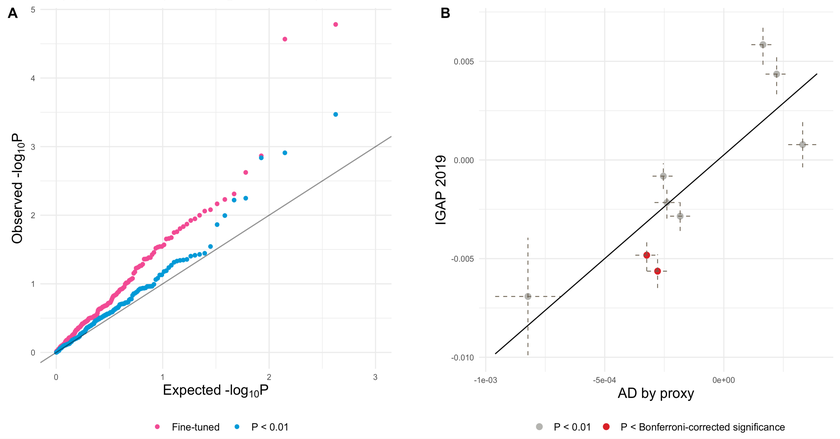


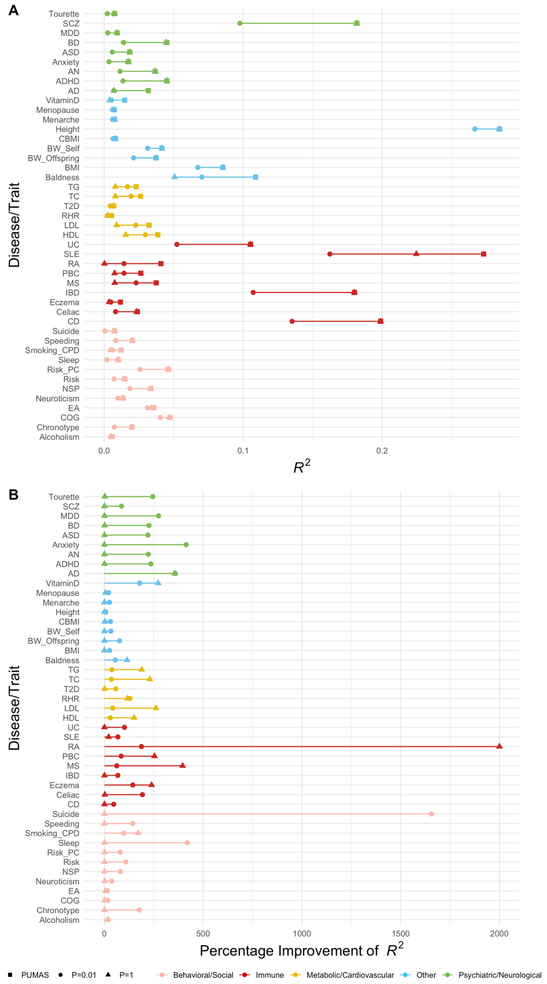
Figure 6. Identifying neuroimaging risk factors for AD using fine-tuned and regular PRSs. (A) QQ plot of two approaches, respectively **(B)** consistency of findings in two independent datasets.



Conclusion

- We provide an innovative solution to a long-standing problem – tuning PRS models with GWAS summary statistics.
- We apply PUMAS to 65 complex diseases and traits. The average gains in predictive R^2 by optimized PRS are 0.0106 (205.6% improvement) and 0.0034 (62.5% improvement) compared to PRS with p-value cutoffs of 0.01 and 1, respectively.
- So far, we have used p-value threshold tuning on pruned sumstats to demonstrate the performance, but the framework can be generalized to more complex settings, as shown in **Figure 4**.

Supplementary Tables and Figures



Supplementary Figure 1. Improvement of predictive R2 by optimized PRS compared to PRS with P=0.01 and 1. (A-B) numerical and percentage improvement

	N=20,000				N=100,000			
	h^2=0.2		h^2=0.8		h^2=0.2		h^2=0.8	
m	PUMAS	RL	PUMAS	RL	PUMAS	RL	PUMAS	RL
50	35	32	37	37	46	45	45	45
1000	335	266	618	574	662	676	851	763
4000	3589	4799	4754	4009	3758	4660	4262	4214

Supplementary Table 1. Additional simulation results. The number in each cell denotes the optimal number of variants to include in the PRS model. m: number of causal variants

Reference

Zhao et al. (2020) Fine-tuning Polygenic Risk Scores with GWAS Summary Statistics. *bioRxiv*

LU LABORATORY
statistical genetics & genome informatics

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON