IMPROVING THE ACCURACY AND INTERPRETATION OF POLYGENIC RISK SCORE THROUGH MODELING THE PATHWAYS OF DISEASE AND MULTIPLE RISK FACTORS

Yihe Yang, Noah Lorincz-Comi, Xiaofeng Zhu, Ph.D.

Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University

ABSTRACT

Motivation

- PRS has become a standard tool for quantifying the genetic risk of complex diseases.
- Traditional PRS typically only explains a limited amount of disease heritability.
- Despite several improvements to the PRS, the gap between PRS-explained variance and disease heritability remains significant.
- Minimax risk theory also implies that the existing PRS methods cannot be improved without additional information on the disease's biological mechanism.

Method

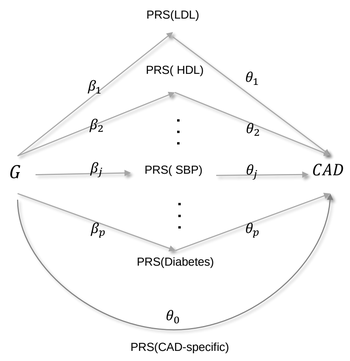
- Genetic variants can influence a disease via the "genetic variants – risk factors – disease" pathway.
- Substantial portion of the genetic associations with complex diseases may be due to intermediary associations with risk factors.
- Using the pathway's information can theoretically increase the accuracy and interpretability of a PRS
- We build a composite PRS with direct and indirect genetic effects, which can increase the accuracy and interpretability of a PRS.

Conclusion

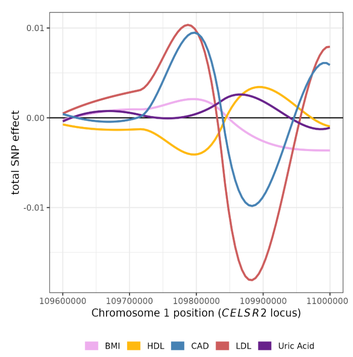
- Composite PRS is epidemiologically interpretable and can improve disease prediction accuracy.
- The significant improvement in prediction accuracy over a traditional PRS can be attributed to the leveraging of information on multiple risk factors for the disease of interest.
- Composite PRS also has a clear interpretation of risk factors affecting an outcome.
- We will apply the new composite PRS method to the prediction of cardiovascular disease outcomes in UK Biobank data and compare our method to existing alternatives.

Step I: Construction of CAD Pathway

Pathway of Composite PRS

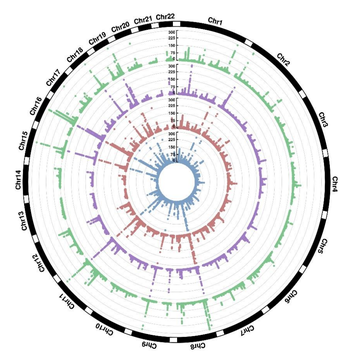


GWAS Effects in CELSR2 Locus

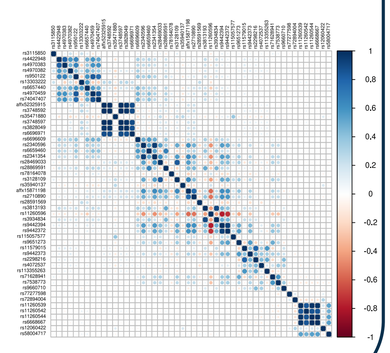


Step II: Estimation of Exposure PRS

GWAS Summary Statistics

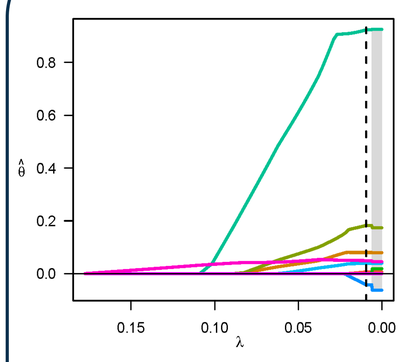


LD Matrix of Variants

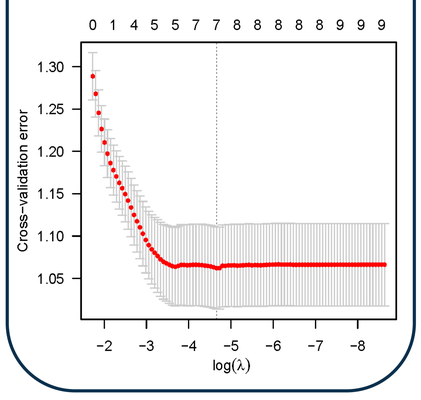


Step III: Selection of Significant PRS

Selection Path of Lasso

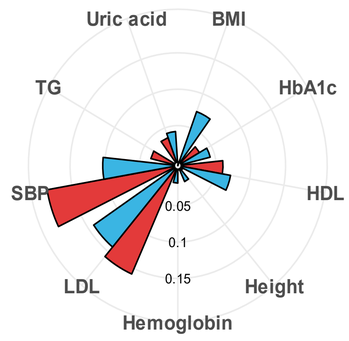


Tuning Parameter Selection of Lasso

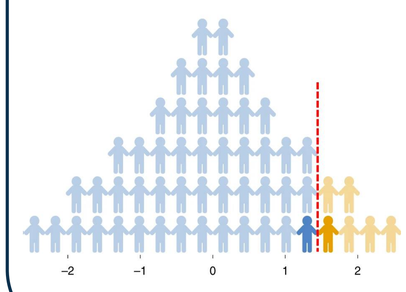


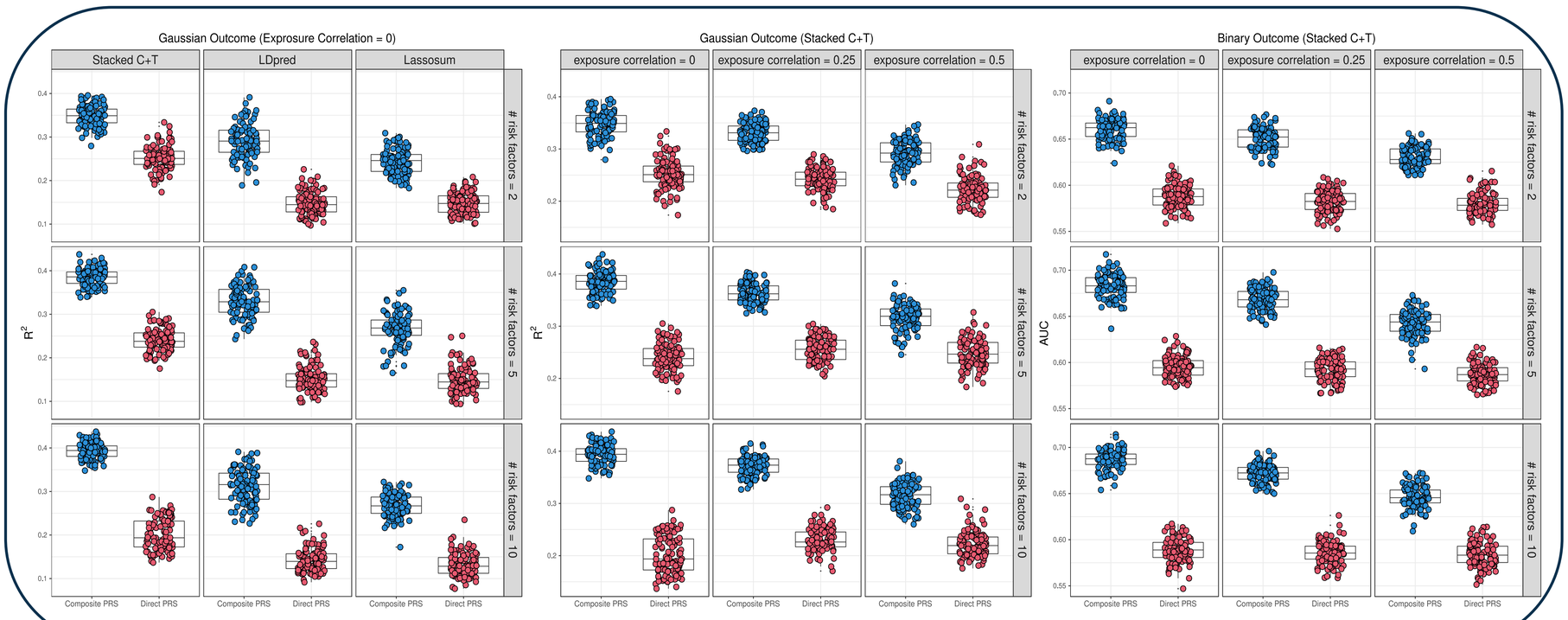
Step IV: Inference of Composite PRS

Contribution Proportion of Exposures



Risk Prediction of CAD patient





Model

- $x_{i1}, \dots, x_{ip}$ are p exposures and y_i is an outcome.
- The model for x_{ij} is

$$x_{ij} = \eta_{ij} + u_{ij},$$

where $\eta_{ij} = \sum_{s=1}^{m} g_{is}\,\beta_{js}$ is the PRS for x_{ij}.

- The model of y_i is

$$y_i = \eta_{i0}\theta_0 + \sum_{j=1}^{p} x_{ij}\,\theta_j + \mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma} + u_{i0}$$

$$= \sum_{j=0}^{p} \eta_{ij}\,\theta_j + \mathbf{W}_i^{\mathsf{T}}\boldsymbol{\gamma} + u_j^*$$

where $\mathbf{W}_i$ is a vector of covariate.

PRS Estimation

- Let $\mathbf{G}$ is the sample matrix of G_i, $\mathbf{x}_\mathrm{j}$ is the sample vector of x_{ij}. Then

$$\widehat{b_\mathrm{j}} = \left(\hat{b}_{1\mathrm{j}}, \dots, \hat{b}_{m\mathrm{j}}\right)^{\mathsf{T}} = \widehat{\mathbf{D}}^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{x}_\mathrm{j}/n$$

where $\widehat{\mathbf{D}}$ is the diagonal variance matrix of $\mathbf{G}$.

- PRS methods estimate the effect size β by

$$\widehat{b_\mathrm{j}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}\beta_\mathrm{j} + \boldsymbol{\epsilon}_\mathrm{j},$$

where

$$\boldsymbol{\epsilon}_\mathrm{j} \sim \mathcal{N}\left(0, \sigma^2_{u_\mathrm{j}u_\mathrm{j}}\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{D}^{-1\frac{1}{2}}\right).$$

- The PRS estimate of x_i is

$$\widehat{\boldsymbol{\eta}}_j = \mathbf{G}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_j.$$

Coefficient Selection

- Suppose we have obtained the estimated PRS $\widehat{\boldsymbol{\eta}}_0, \widehat{\boldsymbol{\eta}}_1, \dots, \widehat{\boldsymbol{\eta}}_p$.
- For normalization, we reweight each PRS by

$$\widehat{\boldsymbol{\eta}}_j = \frac{\widehat{\boldsymbol{\eta}}_j}{se\left(\widehat{\boldsymbol{\eta}}_j\right)}.$$

- We apply a penalized likelihood to estimate $\boldsymbol{\vartheta} = \left(\theta_0, \theta_1, \dots, \theta_p, \boldsymbol{\gamma}^{\mathsf{T}}\right)^{\mathsf{T}}$:

$$\widehat{\boldsymbol{\vartheta}} = \underset{\boldsymbol{\vartheta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{n} -\log\left(l_i(\boldsymbol{\vartheta})\right) + \sum_{j=1}^{p}\lambda_j\left|\vartheta_j\right|\right\},$$

where $l_i(\boldsymbol{\vartheta})$ is the likelihood function.

Inference

- The variant weights of the novel composite PRS is

$$\widehat{\boldsymbol{\beta}}_{com} = \sum_{j=0}^{p}\widehat{\boldsymbol{\beta}}_j\,\hat{\theta}_j$$

- The risk prediction of y_i is yielded by the conditional probability

$$p(y_i|\mathbf{G}_i, \mathbf{W}_i) = \frac{\exp\left(\mathbf{G}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{com} + \mathbf{W}_i^{\mathsf{T}}\hat{\boldsymbol{\gamma}}\right)}{\exp\left(\mathbf{G}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{com} + \mathbf{W}_i^{\mathsf{T}}\hat{\boldsymbol{\gamma}}\right) + 1}.$$

- The contribution of each exposures can be model by the Pratt index:

$$PI\left(\mathbf{x}_j\right) = \hat{\theta}_j \times \hat{r}_j$$

where $\hat{r}_j$ is the marginal regression coefficient between $\widehat{\boldsymbol{\eta}_j}$ and $\mathbf{y}$.

Contact Us

yxy1234@case.edu

xxz110@case.edu

Acknowledgement

This work was supported by grant HG011052 (to X.Z.) from the NHGRI.