

The influence of environmental and genetic factors on the portability of polygenic risk score models across diverse ancestry groups

Nostaeva A.^{1*}, Kuznetsov I.², Zhaivoron A.³, Aulchenko Y.^{1,4}, Karssen L.⁴

¹ *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

² *Skolkovo Institute of Science and Technology, Moscow, Russia*

³ *Novel Software Systems, Novosibirsk, Russia*

⁴ *PolyKnomics, 's-Hertogenbosch, Netherlands*

* *avnostaeva@gmail.com*

Key words: genome-wide association study, summary statistics, polygenic risk score, linkage disequilibrium, single-nucleotide polymorphisms

Motivation and Aim: The use of polygenic risk scores (PRS) has become widespread in biomedical and social science disciplines. Polygenic scoring studies have demonstrated reliable, though modest, prediction using straightforward scoring methods for many complex genetic phenotypes.

However, a historical tendency to use European ancestry samples hinders the use of PRS in non-European ancestry populations. Although there is good reason to anticipate reduced predictive power in non-European ancestry samples, the magnitude of performance decrease, as well as the factors affecting this decline, are not well known.

In this work we try to answer questions about the predictive performance of PRS across diverse ancestry groups and how it depends on environmental and genetic factors.

Introduction: Many studies have shown that PRS is a powerful tool for prediction of the risk of future disease and values of quantitative traits [1, 2]. Through the initiatives of large consortiums and biobanks, sample sizes for genome-wide association studies (GWAS) have reached impressive levels, continuing to improve the predictive power of polygenic risk models [3, 4].

However, there is an over-representation of people of European ancestry in GWAS studies, which results in an under-exploration of non-European populations and a limited ability to use polygenic risk models for prediction in such populations. This problem has been widely recognized [5, 6].

Recent research has focused on the generalizability of polygenic scores to non-European ancestry populations [7]. Due to differences in variant frequencies and patterns of linkage disequilibrium, as well as in environment, predictive power in samples of non-European origin is reduced. However, the dependence of predictive quality on the listed factors is not well known. Thus, there is a need for systematic evaluation of PRS performance across multiple populations and phenotypes.

Here, we developed and validated genome-wide polygenic scores for three anthropometric traits: body mass index (BMI), height and weight, and analysed the influence of environmental and genetic factors on the portability of polygenic risk score models trained and calibrated in people of European ancestry to individuals of non-European descent.

Methods and Algorithms: We developed draft PRS models for each trait using the SBayesR [8] method that utilises summary statistics from GWAS. For GWAS we used a mixed linear model (MLM)-based approach, implemented in the GCTA software package [9], and a training dataset. We selected the best PRS model for each trait, using a validation dataset. Then, we used the selected model to calculate PRS for each person in the calibration dataset and estimated the effects of PRS, sex and age on the phenotype using a linear regression model. This calibrated predictive model was then transferred to other groups. We used not overlapping random samples of individuals of white British born in England from the UK Biobank data (UKB) [10] as the training, the validation, and the calibration datasets.

For quantitative traits, we characterised model performance by the proportion of variance of phenotype, R^2 , explained by observed predictors. Estimation of the predictive performance of the model was made in a series of datasets. We used different cohorts from UKB, such as white British born in England, Wales and Scotland, Ireland, and outside of the UK (“other Europeans”), as well as samples of individuals of Pakistani, Indian, African, Caribbean, and Chinese ancestry. We also added in this analysis cohorts from other consortiums, which include individuals of British, American, Belgian, Croatian and Italian populations. To avoid “testimation” bias, we validated and estimated the models in independent data sets not overlapping with the samples used for GWAS.

As a genetic factor we used the fixation index (F_{st}), which is a measure of population differentiation due to genetic structure. Regarding an environmental factor, we used the average value of a trait for each of the cohorts.

Results: Our results demonstrate a low portability of PRS from European ancestry to non-European for all considered traits. To illustrate, for BMI, it performed reasonably well for different groups of European ancestry (R^2 varies from 9 to 15 %), but had a low predictive accuracy (R^2 less than 6 %) for people of non-European descent.

These differences and a gradient in R^2 have been studied in conjunction with genetic and environmental factors, where Pearson's correlation was observed. Results show, for example, that the correlation between the F_{st} parameter and the quality of BMI prediction was about -0.85 , while for the mean value of this trait, the correlation was around -0.84 .

Conclusion: We concluded that PRS models developed for European populations cannot be directly transferred to other populations. The model performance drops with increased genetic and environmental distances between the population where the model was developed (training and calibration) and the target population.

Acknowledgements: This research has been conducted using the UK Biobank Resource project 41601. This work was funded by PolyKnomics BV.

References

1. Wray N.R., Goddard M.E., Visscher P.M. Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev.* 2008;18:257-263.
2. Torkamani A., Wineinger N.E., Topol E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet.* 2018;19(9):581-590.
3. Lello L. et al. Accurate genomic prediction of human height. *Genetics.* 2018;210:477-497.
4. Lee J.J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet.* 2018;50:1112-1121.
5. Bustamante C.D., Burchard E.G., De la Vega F.M. Genomics for the world. *Nature.* 2011;475:163-165.
6. Petrovski S., Goldstein D.B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* 2016;17:157.

7. Martin A.R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51:584.
8. Lloyd-Jones L.R., Zeng J. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun.* 2019;10(1):5086.
9. Jiang L., Zheng Z., Qi T. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet.* 2019;51:1749-1755.
10. Bycroft C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562:203.