FairPRS: a fairness framework for Polygenic Risk Scores

Diego Machado Reyes†, Aritra Bose†, Ehud Karavani, and Laxmi Parida‡

†Biomedical Engineering Department, Rensselaer Polytechnic Institute, Troy, NY, USA
‡IBM Research, Haifa, Israel

† Equal Contribution
‡ Email: parida@us.ibm.com

Polygenic risk scores (PRS) are increasingly used to estimate the personal risk of a trait based on genetics. However, most genomic cohorts are of European populations, with strong under-representative of multi-ethnic minority groups. Given that PRS poorly transport across racial groups, this has the potential exacerbate health disparities if used in clinical care. Hence there is a need to generate PRS that perform comparably across ethnic groups. Borrowing recent advancements in the domain adaption field of machine learning, we propose *FairPRS* - an Invariant Risk Minimization (IRM) approach for estimating fair PRS or debiasing pre-computed ones. We test our method on both a diverse set of synthetic data and real data from the UK Biobank. We show our method can create ancestry-invariant PRS distributions that are both racially unbiased and largely improve phenotype prediction. We hope that fair PRS will contribute to fairer characterization of patients by genetics rather than by race.

**Keywords**: Polygenic Risk Scores; Fairness; Racial Disparity; Invariant Risk Minimization; Machine Learning; Precision medicine

1. Introduction

Genome wide association studies (GWAS) were developed for finding statistical association between single nucleic polymorphisms (SNPs) and phenotype traits. Later, these associations were then aggregated into a score – a polygenic (risk, for diseases) score (PRS) – for predicting traits. PRS became extremely popular due to their promise at harnessing one’s genome to act as a biomarker for personalizing medical risk estimation. This capacity for personalization can also translate to heterogeneity on the population level with PRS helping to identify subpopulations who are at higher risk of disease.

Unfortunately, PRS are plagued by many issues. Primarily GWAS cohorts strongly suffer from lack of sample diversity. For example, 79% of all participants in the NHGRI-EBI GWAS catalog are of European descent despite being only 16% of the global population. The under-representation of minority groups in cohorts leads to inferior PRS because PRS derived from European ancestry tend to perform poorly in genetically diverse populations and even
within other admixed European populations. As a simple example, polygenic scores for height predicts all Africans to be shorter than Europeans, contrary to empirical evidence. Thus, using PRS for precision medicine at its current form may exacerbate health disparities until the lack of representation is not solved.

Reducing racial bias in genomic prediction may contribute to a more equitable healthcare for all. But to establish health equity in precision medicine we require better genetic cohorts whose multi-ethnic representation matches real life. This solution, however, is resource heavy and is a long-term goal. In the meantime, we can apply advances in machine intelligence to mitigate bias in trait prediction from PRS.

There exists prior work of using computational framework in making PRS generalize better across subgroups. These include deconvoluting ancestry and partial PRS computation and computing ancestry-specific PRS to showcase their utility as predictors across different populations. Advances in machine learning such as using transfer learning-based methods and deep learning based methods have been applied to make PRS more portable across ancestries. However, either some of these methods assume part of the background genome is still of European origin or considers pre-computed associated markers as input to reduce search space which can contain significant bias or spurious associations.

In this work, we apply a domain-adaptation-based paradigm called Invariant Risk Minimization (IRM) in the context of PRS. We consider the problem of generalizability of PRS as an out of distribution generalization problem, a common machine learning problem where models are developed in one domain but are deployed in another. IRM’s goal is to generate invariant predictors given multiple training domains. In our context, these different domains are adapted to be the different ancestry groups, therefore allowing for race-invariant phenotype prediction from PRS. Our goal is then to learn a generalizable PRS that contain as little ancestry information as possible, while still accurately predicting the phenotype of interest.

We present FairPRS, a straightforward and efficient framework for finding and mitigating bias in PRS which improves generalizability across populations and make it portable while increasing the prediction accuracy of phenotype of interest. FairPRS is robust across both rigorous simulation studies involving arbitrary population structure and pre-computed PRS obtained from UK Biobank (UKB). As use of PRS is being advocated in clinical care, FairPRS can be a very important tool to achieve equity in healthcare as well as further our understanding of true genetic causes of disease risk.

2. Methods

FairPRS offers an entire pipeline from genetic data to trait prediction. It has three possible access points for input: genotypes, genotypes with summary statistics, or a pre-computed PRS. We will explain the FairPRS framework herein, followed by the autoencoder architecture, training and evaluation phases of the pipeline. Thereafter, we will discuss the simulation and real data used in the study for evaluating FairPRS including computational details.
2.1. FairPRS framework

The FairPRS pipeline is designed for easy and customizable use with multiple access points based on the needs of the user. Moreover, the pipeline can be run for a user-determined number of iterations for all or specific portions. The first stage focuses on processing the genotype data towards the PRS computation. It allows to calculate summary statistics, namely GWAS, and principal components (PCs) of the genotype data. The PCs can be used as covariates for the GWAS and as input to the FairPRS model. The summary statistics are computed using Plink 2.0 and the PCs are efficiently calculated for large scale data using TeraPCA. Next, the pipeline allows to start at the PRS computation step if the user has previously calculated the summary statistics. The betas are extracted from the summary statistics and used for PRS computation through PRSice2. Lastly, the third stage is the FairPRS model which uses the pipeline-computed or user-provided PRS and PCs as input, while the phenotype and the PRS will be used for the training supervision. The model is implemented as a dual task autoencoder and MLP as shown in Figure 1. Briefly, first, the data is encoded into a shared latent representation. The latent representation is then fed into two tasks in parallel: decoding back the inputted PRS and predicting the phenotype. The losses are then combined with the ancestry information to obtain the IRM loss. Finally, the fair PRS estimates are obtained from the PRS decoder output. A key point in this step is the automatic multi-thread hyperparameter tuning per iteration with allows the pipeline to train high performing models in an efficient manner. After the model training and evaluation, the average performance over the iterations is reported and a dictionary with all the results per iteration is saved for further analysis and reproducibility purposes.

2.1.1. Implementation and Evaluation

**Detailed architecture** The encoder is a single layer with ReLU activation, and latent space size determined as a hyperparameter. Both the PRS decoder and the phenotype prediction head perform a 10% dropout and then apply a single linear layer. The ERM loss is obtained by adding the two MSE losses with equal weight. The final loss is a weighted sum of the...
ERM and IRM losses, with the weight being a hyperparameter. Adam method was used for optimization.\textsuperscript{17} Framework is implemented using PyTorch 1.11.\textsuperscript{18}

**Training** The proposed model allows using regression losses for the double task network and employ multiple environments which are automatically adjusted based on the number of populations present in the PRS data. Moreover, as machine learning models rely on fine-tuning and hyperparameter selection to achieve state-of-the-art performances, an automatic hyperparameter search is used at the training stage. Hyperparameter random search was done for the learning rate (log-uniform $[10^{-5}, 0.1]$), the dimension of the latent space (uniform from $2^i : i \in [2, 9]$), and the relative weight of the IRM loss (uniform $[0.5, 1.5]$). The search space was defined based on preliminary experiments allowing for a wide search without a prohibitively computationally expensive search space. Tuning was done using Ray Tune.\textsuperscript{19} UK Biobank data was also randomly split to train (70%), validation (20%), and test (10%) sets. The best hyperparameter configuration was selected based on a validation set and was subsequently used for evaluation.

**Evaluation** To test the model against a baseline in a fair way, both the original PRS and those resulting from the model were regressed separately against the outcome using ordinary least squares. The covariate adjusted coefficients of determination ($\text{adjusted} R^2$ scores) for both models are reported. Regression was done in Python using statsmodels.\textsuperscript{20} Results per iteration are computed to finally report the mean performance across all iterations.

### 2.2. Data

**FairPRS** was evaluated on multiple simulated and real datasets. The simulated datasets included a wide array of configurations and were generated using the data simulator in.\textsuperscript{21} Additionally, UK Biobank enhanced PRS (ePRS-UKB) for multiple binary and continuous phenotypes were used to further evaluate the model in real world scenarios across different disease outcomes.

**Simulated data** Three models for simulating genetic datasets with arbitrary population structure: Balding-Nichols (BN), Prichard-Stephens-Donelly (PSD), and 1000 Genomes Project (TGP) with 3 variance proportion configurations for genetic, environment and noise, $\{v_{\text{gen}}, v_{\text{env}}, v_{\text{noise}}\}$, totaling in 9 different simulation scenarios were used to evaluate FairPRS. We used three populations for BN and PSD and ten populations for TGP. For each model we generated 10 iterations resulting in 90 different datasets. The 3 proportions configurations used were $\{v_{\text{gen}} : v_{\text{env}} : v_{\text{noise}}\} = \{5 : 5 : 90, 10 : 20 : 70, 20 : 40 : 40\}$. The number of causal SNPs was set at 5% for all simulated datasets. Moreover, for all configurations the simulated datasets included 100,000 SNPs, 10,000 samples for GWAS, 1000 for PRS training and 400 for PRS testing.

**Real data** PRS and ancestry data were obtained from the UKB for further model validation.\textsuperscript{22} Enhanced PRS (ePRS-UKB) for 6 different conditions across 104,231 multi-ethnic individuals were used in our analysis, these are height, body mass index (BMI), glycated
hemoglobin (HBA1C), high density lipoprotein cholesterol (HDL), and low-density lipoprotein cholesterol (LDL).

3. Results

3.1. Simulated data

Across all simulation scenarios with 10 iterations for reproducibility, FairPRS consistently achieved higher or comparable phenotype prediction accuracy with respect to the original PRS computed by PRSice2, measured in terms of adjusted $R^2$ after correcting for top eight principal components (PCs) computed by TeraPCA (Supplementary Figure 1). FairPRS achieved better results on all models across all simulation scenarios (Figure 2). Kolmogorov-Smirnov (KS) two-sample tests, a goodness of fit test of equality of the original vs. observed PRS distributions were done to test the null hypothesis of whether the two distributions were sampled from the same unknown distribution. This resulted in very low $p$-values ($p < 10^{-160}$) across all simulation scenarios which rejected the null hypothesis that the FairPRS distributions and the original PRS distribution were sampled from the same distribu-

---

**Fig. 2.** Simulation study results for three simulation models, BN, PSD and TGP. A. Distributions of ancestry-specific PRS computed by (i) PRSice2 and (ii) FairPRS. B. Box-and-whisker plot of adjusted $R^2$ between the phenotype and PRS computed by PRSice2 and FairPRS across the variance proportions for $\{v_{gen} : v_{env} : v_{noise}\}$. 
bution (see Supplementary Table 1). The KS tests were done using SciPy package in python.

The Net Reclassification Index (NRI), a percentage score of the difference in adjusted $R^2$ by using FairPRS on top of pre-computed PRS using PRSice2 shows that as the genetic variance ($v_{gen}$) increases in contribution to the phenotype, the NRI also increases (Figure 3). Hence, when a pre-computed PRS is augmented with FairPRS not only do we observe a higher $R^2$ across all the simulation scenarios, we also obtain a relatively unbiased PRS estimate with negligible ancestry influence.

3.2. Real data

To demonstrate how FairPRS estimates real-world traits, we applied it on UKB-ePRS across six traits as mentioned above. FairPRS achieves considerably higher $R^2$ compared to the pre-computed ePRS-UKB for all traits analyzed (Figure 4).

FairPRS was run 10 times for each ePRS-UKB trait analyzed for replicability and hyperparameter tuning. The NRI was computed by the percentage difference in $R^2$ when using FairPRS vs. pre-computed PRS. Maximum NRI was observed in glycated hemoglobin (HbA1c)
which is a biomarker for Type 2 Diabetes and has been shown to have high predictivity accuracy for PRS. This was followed by BMI and Height, respectively which are very well-studied in terms of phenotypic variance explained by PRS.\footnote{KS test for the two PRS distributions, \textit{FairPRS} and pre-computed ePRS-UKB also resulted in rejection of the null hypothesis (see Supplementary Table 2) and demonstrated that \textit{FairPRS} learns a domain invariant distribution different from its input. This shows how \textit{FairPRS} can result in better predictive accuracy in large biobanks such as UKB and can be integrated in precision medicine efforts.} KS test for the two PRS distributions, \textit{FairPRS} and pre-computed ePRS-UKB also resulted in rejection of the null hypothesis (see Supplementary Table 2) and demonstrated that \textit{FairPRS} learns a domain invariant distribution different from its input. This shows how \textit{FairPRS} can result in better predictive accuracy in large biobanks such as UKB and can be integrated in precision medicine efforts.

4. Discussion

In this work, we combined notions from classical genetics–the polygenic risk scores (PRS)–with notions from machine learning and domain adaptation. We developed a model that applies Invariant Risk Minimization (IRM) approach to estimating PRS. Using both synthetic data and pre-computed PRS from the UK Biobank, we obtained PRS that are indistinguishable across races, while improving prediction accuracy in terms of $R^2$.

Despite being very promising, GWAS are often plagued by over-representing European ancestry populations in their cohorts. If left uncorrected, this disproportional representation of populations structure lead to both spurious associations and only being able to explain a small fraction of heritability, among others.\footnote{As PRS are computed from GWAS summary statistics, PRS inherits many of these challenges which contributes to its poor generalizability and transferability across populations due to underlying influence of LD structure and environmental factors.} Our method for finding fair estimates of PRS based on domain adaptation, learns an ancestry-invariant estimates which provide both qualitative and quantitative advantages.

Domain adaptation is a sub-field of machine learning focusing on model performance across multiple domains. The simplest driving example is when the distribution of data used for development, shifts during deployment of the model. For example, using images of Swiss cows in the grassy Alps for training, while deploying the model to identify cows in the sandy beaches of Corsica.\footnote{By having training data from multiple such sources and by training in an environment-aware approach - as with IRM, we can reduce the amount of spurious correlations our model learns, like the grassy Alpine background.} By having training data from multiple such sources and by training in an environment-aware approach - as with IRM, we can reduce the amount of spurious correlations our model learns, like the grassy Alpine background.

In this work, we extend the notion of “domains” to different population ancestries. We apply the IRM framework, a form of supervised domain adaptation, to adjust the pre-computed PRS scores to be ancestry ignorant. Intuitively, we try to learn the most phenotype-predictive PRS, while forcing ourselves to ignore (or “forget”) any residual race information. Using IRM means we encourage the model to learn only information that is shared across ancestries. By constraining the PRS distribution of ancestries to coalesce, we ensure that when using the PRS for phenotype prediction, we get equal performance across ancestries. Thus, leading to a fairer PRS.

Different ancestries do exhibit disparities in health-related measures, and, therefore, different phenotypic distribution. However, these differences are rarely inherently biological. More often they are the result of how different ethnic subgroups interact with the healthcare system differently.\footnote{Different ancestries do exhibit disparities in health-related measures, and, therefore, different phenotypic distribution. However, these differences are rarely inherently biological. More often they are the result of how different ethnic subgroups interact with the healthcare system differently.} (More formally, race disparities are more of an acquisition shift, rather than...
population or prevalence shift \(^2\)). Consequently, forcing to disentangle race information from genetic information will (at least partially) remove race bias and will lead to a fairer usage of genetic data when assessing genetic risk.

**FairPRS** can be used as a tool to find unbiased estimates of pre-computed PRS or from GWAS summary statistics which would better predict the phenotype of interest and hence it is fast and efficient. **FairPRS** estimates can be used as a step forward to achieve equity in precision medicine and evaluating disease risk in large clinical cohorts. It can be extensively used for out-of-sample prediction with pre-computed PRS to obtain ancestry-robust PRS which transport better across ancestries and datasets. In future work, we want to compare performance of PRS computed by state-of-the-art methods and ancestry-robust **FairPRS** and evaluate their portability to other ancestries.\(^29\)

We hope that **FairPRS** will contribute to fairer characterization of patients by genetics rather than by race.

**Code Availability** A Pytorch based implementation of **FairPRS**, along with scripts, descriptions and sample data to run experiments are available at [https://github.com/ComputationalGenomics/FairPRS](https://github.com/ComputationalGenomics/FairPRS)

**Data Availability** Simulated data is made available upon request. UKB-ePRS are available from UK Biobank.

**Supplementary Material** Supplementary material is hosted in the Supplementary directory in [https://github.com/ComputationalGenomics/FairPRS](https://github.com/ComputationalGenomics/FairPRS)

**Acknowledgements** This work was funded by IBM. We would like to thank Kenney Ng for helping with data access.

**References**

7. D. Marnetto, K. Pärna, K. Läll, L. Molinaro, F. Montinaro, T. Haller, M. Metspalu, R. Mägi, K. Fischer and L. Pagani, Ancestry deconvolution and partial polygenic score can improve sus-


