PLOS GENETICS

RESEARCH ARTICLE

Transfer learning with false negative control improves polygenic risk prediction

Xinge Jessie Jeng[1], Yifei Hu[1], Vaishnavi Venkat[2], Tzu-Pin Lu[3,4], Jung-Ying Tzeng[1,2,3¤]*

1 Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Institute of Health Data Analytics and Statistics, National Taiwan University, Taipei, Taiwan, **4** Department of Public Health, National Taiwan University, Taipei, Taiwan

¤ Current address: Department of Statistics and Bioinformatics Research Center, Campus Box 7566, Raleigh NC, United States of America
* jytzeng@ncsu.edu

Citation: Jeng XJ, Hu Y, Venkat V, Lu T-P, Tzeng J-Y (2023) Transfer learning with false negative control improves polygenic risk prediction. PLoS Genet 19(11): e1010597. https://doi.org/10.1371/journal.pgen.1010597

Editor: Michael P. Epstein, Emory University, UNITED STATES

Received: January 1, 2023

Accepted: November 9, 2023

Published: November 27, 2023

Data Availability Statement: • For simulation data, the relevant R code used in the simulation are available at https://github.com/JessieJeng/FNC-Lasso. • For the CoLaus/PsyCoLaus data, it can be applied from the database of Genotypes and Phenotypes (dbGaP) (dbGaP Study Accession: phs000145.v4.p2) at NCBI website: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2. • The numerical results underlying the tables and figures are included as supporting materials (S2 Table).

Abstract

Polygenic risk score (PRS) is a quantity that aggregates the effects of variants across the genome and estimates an individual's genetic predisposition for a given trait. PRS analysis typically contains two input data sets: base data for effect size estimation and target data for individual-level prediction. Given the availability of large-scale base data, it becomes more common that the ancestral background of base and target data do not perfectly match. In this paper, we treat the GWAS summary information obtained in the base data as knowledge learned from a pre-trained model, and adopt a transfer learning framework to effectively leverage the knowledge learned from the base data that may or may not have similar ancestral background as the target samples to build prediction models for target individuals. Our proposed transfer learning framework consists of two main steps: (1) conducting false negative control (FNC) marginal screening to extract useful knowledge from the base data; and (2) performing joint model training to integrate the knowledge extracted from base data with the target training data for accurate trans-data prediction. This new approach can significantly enhance the computational and statistical efficiency of joint-model training, alleviate over-fitting, and facilitate more accurate trans-data prediction when heterogeneity level between target and base data sets is small or high.

Author summary

Polygenic risk score (PRS) can quantify the genetic predisposition for a trait. PRS construction typically contains two input datasets: base data for variant-effect estimation and target data for individual-level prediction. Given the availability of large-scale base data, it becomes common that the ancestral background of base and target data do not perfectly match. In this paper, we introduce a PRS method under a transfer learning framework to effectively leverage the knowledge learned from the base data that may or may not have similar background as the target samples to build prediction models for target individuals. Our method first utilizes a unique false-negative control strategy to extract useful

Funding: This work is partially supported by National Institutes of Health (https://www.nih.gov/) grants RF1AG074328 and U24AG041689 to JYT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

information from base data while ensuring to retain a high proportion of true signals; it then applies the extracted information to re-train PRS models in a statistically and computationally efficient fashion. We use numerical studies based on simulated and real data to show that the proposed method can increase the accuracy and robustness of polygenic prediction across different ranges of heterogeneities between base and target data and sample sizes, reduce computational cost in model re-training, and result in more parsimonious models that can facilitate PRS interpretation and/or exploration of complex, non-additive PRS models.

Introduction

Polygenic risk score (PRS), first proposed in [1], is a quantity that aggregates the effects of variants across the genome and estimates an individual's genetic predisposition for a given trait. PRS can be used to explore the polygenetic architecture of a trait, to predict disease risk, to identify groups of individuals with substantially increased risks, and to study shared polygenic signals between different traits.

PRS is typically calculated as a weighted sum of trait-associated alleles, where the weights are the estimated effect sizes of the alleles on the trait. PRS analysis usually contains two input data sets: (i) base data, containing summary statistics of association tests from GWAS such as the estimated regression coefficients with the traits and their corresponding p-values, and (ii) target data, consisting of individual-level genome-wide genotypes data and traits used for prediction. Accumulating evidence suggests that for a variety of complex traits, variants selected based on a large ancestry-diversified base data are expected to harbor high fractions of causal variants carried by the target data, yet their effect sizes may not be the same across base and target data sets [2–4].

Because not all genetic variants influence the trait and because SNPs are correlated due to linkage disequilibrium (LD), two types of selection approaches have been adopted to construct PRS using GWAS summary statistics from the base data. The first type constructs PRS purely based on GWAS summary statistics from marginal model, i.e., removing correlated redundant SNPs via linkage disequilibrium (LD) pruning/clumping and computing the weighted allele sums only on those SNPs whose GWAS p-values are lower than a certain threshold (e.g., pruning/clumping + thresholding (C+T) [5], PRSice [6], and PRSice2 [7]). In these approaches, *p*-value threshold is usually set as a tuning parameter and optimized by data. The second type considers a joint additive model of all SNPs to estimate the effect sizes and to account for SNP LD, either by imposing Lasso regularization (e.g., lassosum [8]) or prior distributions (e.g., LDpred [9], PRS-CS [10] and LDpred2 [11]). PRS based on joint modeling tend to have higher prediction accuracy [12] due to the increased accuracy in estimated SNP effects.

Given the utility of large-scale base data, it becomes more and more common to encounter the scenarios where the ancestral background of the base sample and target sample may or may not be perfectly matched. For example, in the CoLaus/PsyCoLaus study [13, 14], the target sample is composed of Switzerland Caucasians while the base sample is composed of general European descents from US, central Europe, southern Europe and northern Europe. In multi-ethnic PRS prediction, the base sample tend to comprise European descents in large sample size as well as descents from other populations including the target population in moderate sample size. Consequently, although causal variants in the target data are expected to be included in a larger set of causal variants carried by the base data, the same variant may have different effect sizes in the base and target data sets.

To accommodate the potential ancestral heterogeneities between base sample and target sample, in this work, we treat the GWAS summary information obtained in the base data as knowledge learned from a pre-trained model, and adopt a transfer learning framework to leverage the knowledge learned from the relatively ancestry-diversified base data to build the PRS of target individuals. Our proposed trans-learning framework consists of two main steps: (1) conducting false-negative control (FNC) screening to extract useful knowledge from the base data; (2) conducting joint-model training to integrate the extracted knowledge with the target training data for accurate trans-data prediction. In (1), as SNP effects may differ across base and target datasets, marginal screening for promising SNPs should not just focus on strong base signals, but should have good capacity for weak base signals as well. Therefore, we use FNC to ensure the retention of a high proportion of strong and weak base signals and, at the same time, effectively exclude noise variants that are distinguishable from base signals. This is crucial for the second step of joint-model training. To ensure computational efficiency and numerical stability, it is essential to only include a reduced set of SNPs when training the joint PRS model; however, if the majority of signals were not included in this reduced set, the prediction accuracy of the PRS model would be compromised. In (2), we use joint regression to train the target PRS model with individual-level target data on the reduced promising-SNP set, so as to ensure accurate SNP effect estimates that better reflect the underlying causal SNP architecture of the target sample.

We use simulation to illustrate the performance of the proposed transfer learning procedure. In settings with various overlapping proportions of signals between base and target data sets, the new procedure leads to a better PRS model in terms of achieving a higher predictive R^2 with a more parsimonious model. Application to the CoLaus/PsyCoLaus GWAS of lipid plasma indicates that besides being substantially faster than C+T, lassosum, and LDped methods, the new procedure can achieve the highest predictive R^2 while retaining the least number of SNPs in the PRS model. A parsimonious PRS model that includes fewer SNPs whilst maintaining the same or higher predictive R^2 can facilitate PRS interpretation and enable downstream analyses with more complex modeling techniques, such as incorporating SNP-SNP interactions in polygenic prediction models as discussed in [15, 16]. We also explore interactive models with the lipid prediction in the CoLaus/PsyCoLaus study.

Materials and methods

Formulation of trans-data signals

Let $\mathcal{S}$ be the set of causal variants carried by the target data and $\mathcal{S}^+$ be the set of signal variants carried by the base data. Signal variants in $\mathcal{S}^+$ are defined as variants whose true effect sizes in the GWAS marginal model are not zero. In this paper, we use "causal variants" to refer to variants truly affecting the phenotype, and use "signal variants" to refer to those variants that have non-zero true effect sizes in a certain statistical model. While causal variants are considered signal variants, the reverse is not always true. For example, a non-causal SNP that is in high LD with some causal variants would be a signal variant in the GWAS marginal model, but would not be a signal variant in a joint model that accounts for the causal variants. We consider heterogeneities between base data and target data from two sources: (a) the nested ethnicity relationship between base and target samples, and (b) the possible distortion of signal variants information from the pre-trained GWAS marginal model. A general assumption on the trans-data information sharing can be

$$\mathcal{S} \subseteq \mathcal{S}^+.$$

That is, the causal variants in the target data are included in the set of signal variants in the

more ancestry-diverse base data. $\mathcal{S}^+$ could also include non-causal variants, which have non-zero marginal effects because of their dependence with the causal variants in LD.

The nested assumption of $\mathcal{S} \subseteq \mathcal{S}^+$ could be satisfied by using multi-ancestry base data, which can be obtained by conducting meta-analyses on multiple GWAS of different ancestral groups, such as the multi-ancestry GWAS base data considered in [17]. The multi-ancestry base samples may be predominantly from a specific ancestral group; hence it is possible that the effect of a causal variant in $\mathcal{S}$ gets diluted in the ancestrally diverse base data and becomes fairly weak in $\mathcal{S}^+$. In addition, the effect sizes of a variant can differ between base and target data due to the use of marginal association models in base data. An ideal PRS method should address these issues, incorporate both strong and weak base signals and re-train PRS model with target data.

Trans-data polygenic risk prediction

We propose a transfer learning polygenic prediction procedure which performs FNC marginal screening to effectively retain both strong and weak signals in $\mathcal{S}^+$ and applies the learned information from base data to identify causal variants in $\mathcal{S}$ and improve PRS for target individuals. Different from the classical power analysis in hypothesis testing, FNC screening does not hinge on a pre-fixed control level of type I error (or some form of cumulative type I errors in multiple testing). Instead, FNC screening adapts to a user-specified level of false negative proportion (FNP) to facilitate more powerful discovery of weak signals while keeping out noise variants (i.e., non-signal variants) that are distinguishable from weak signals.

Fig 1 shows a flow chart of the transfer-learning procedure with FNC marginal screening and joint model training. In Step 1, we apply FNC marginal screening to the base data to extract useful knowledge as follows. First, we pre-fix a sequence of control levels, $\epsilon_1, \ldots, \epsilon_K$, on FNP, which is the number of false negatives divided by the total number of signals $|\mathcal{S}^+|$. In this paper, we set ϵ_k's in grid values of {0.02, 0.04, . . ., 0.4}. Second, for a given ϵ_k, we apply the

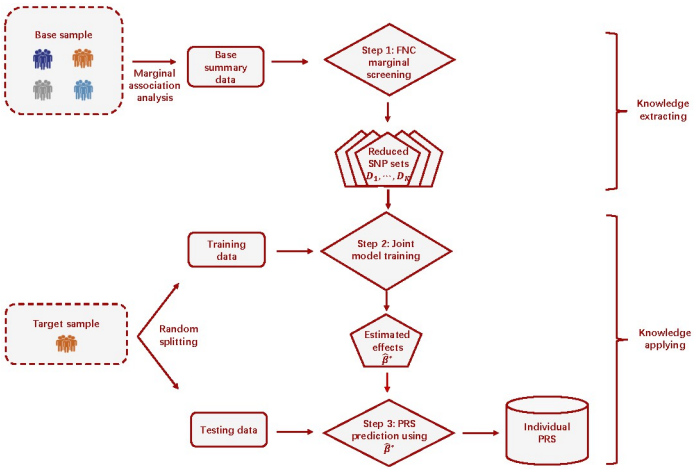

Fig 1. Overview of the PRS transfer learning procedure. Step 1: Given a pre-specified grid of control level for false negative proportion, $\{\epsilon_1, \cdots, \epsilon_K\}$ (e.g., {0.02, 0.04, ⋯, 0.4}), apply false negative control (FNC) marginal screening to the base summary statistics and obtain K candidates of reduced SNP sets $\mathcal{D}_1, \cdots, \mathcal{D}_K$, each of which retains a specific high proportion (i.e., $1 - \epsilon_1, \cdots, 1 - \epsilon_K$, respectively) of base signals. Step 2: For each reduced SNP set $\mathcal{D}_k$, train the joint prediction model using the target training sample. The joint model that yields the largest R^2 is selected as the final model and the corresponding SNP effect estimates are denoted as $\hat{\beta}^*$. Step 3: Apply the estimated effects $\hat{\beta}^*$ from the final model of Step 2 to the target testing sample to calculate individual PRS.

https://doi.org/10.1371/journal.pgen.1010597.g001

FNC screening procedure of [18] on p-values of the base summary statistics and obtain a reduced SNP set $\mathcal{D}_k$, which is expected to contain a $(1 - \epsilon_k)$ proportion of base signals in $\mathcal{S}^+$ as well as some noise SNPs. However, because FNC screening excludes noise SNPs that are distinguishable from the $(1 - \epsilon_k)$ proportion of signals in $\mathcal{S}^+$, each $\mathcal{D}_k$ can be much smaller than the full set of SNPs. More detailed descriptions of FNC screening is provided in the next subsection.

In Step 2, we equally and randomly divide the individual-level target sample into training and testing samples, and use the training sample and Lasso regression to build the PRS model. Specifically, given a reduced SNP set $\mathcal{D}_k$, $k = 1, \ldots, K$, we fit a joint prediction model using regularized regressions, select the regularization parameter using cross-validation or information criterion, and obtain the regularized estimates of SNP effects and compute the model R^2 value. Among the K PRS models, the model yields the largest R^2 value is selected as the final model, and the corresponding effect size estimates (denoted by $\hat{\beta}^*$) are used to construct PRS. In Step 2, working with the FNC reduced SNP set $\mathcal{D}_k$ instead of the entire SNP set enhances prediction accuracy and computational efficiency of the joint-model training. In Step 3, we apply the estimated effects from the final model $\hat{\beta}^*$ to the target testing sample and calculate PRS for each testing individual.

The proposed transfer learning procedure is general and can accommodate a spectrum of prediction models such as linear mixed effects models (LMM) [19] and penalized linear regression [20, 21] in Step 2. In this paper, we adopt the popular Lasso regression to obtain sparse effect estimates and utilize cross-validation to determine the hyper-parameters. We denote the whole procedure as FNC+Lasso and show the complete algorithm of FNC+Lasso in Algorithm 1 of S1A Appendix. The computer code for FNC+Lasso, together with complete documentation is publicly available at Github: https://github.com/JessieJeng/FNC-Lasso.

We note that the default implementation of FNC+Lasso at Github does not use LD information. When LD is not accounted for, the total number of base signal variants, denoted as $|S^+|$, could be overestimated, but the computational time required for estimating $|S^+|$ can be substantially reduced. An overestimated $|S^+|$ would result in more conservative screening results, i.e., more variants would be selected to enter the joint-modeling step. Given the computational gain and the screening results, we opt not to incorporate LD information among SNPs in the screening step.

FNC marginal screening

The original FNC screening procedure was designed to preserve a high proportion of signals via effectively controlling false negatives [18]. In comparison to other existing false negative control methods [22–24], FNC screening exhibits the capacity to accommodate arbitrary covariance dependencies [18]. FNC screening can be applied as follows. Given a sequence of p-values for all the SNPs in the base data: $p_1, \ldots, p_m$, rank the SNPs by their p-values such at $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. For each $j \in \{1, \ldots, m\}$, calculate

$$\widehat{FNP}_j = \max\{1 - j/\hat{s} + (m - \hat{s})p_{(j)})/\hat{s}, 0\}, \quad (1)$$

where $\hat{s}$ is an estimate for $|\mathcal{S}^+|$, the number signal variants in the base data. Now, for a control level ϵ_k, derive

$$j_k^* = \min\{j : \widehat{FNP}_j < \epsilon_k\}. \quad (2)$$

Then, construct the SNP set $\mathcal{D}_k$ as the collection of SNPs whose p-values are less than or equal to $p_{(j_k^*)}$. Repeat the above for each of $\epsilon_1, \ldots, \epsilon_K$, and obtain the reduced SNP sets $D_1, \ldots, D_K$.

The FNC screening procedure is based on direct estimation for the true FNP if only the top j ranked SNPs are selected. Rationale of the estimator in (1) is as follows. When the top j variants are selected, the total number of signal variants equals to the sum of true positives (TP_j) and false negatives (FN_j), i.e. $|\mathcal{S}^+| = s = TP_j + FN_j$. Then we have $FNP_j = FN_j/s = 1 - TP_j/s$, and we would need to estimate s and TP_j for each j.

FNC screening uses $\hat{s}$ from [25] as recommended in [18] to estimate s, whose consistency has been proved under arbitrary covariance dependence. For TP_j, note that selecting the top j variants incurs true positives and false positives with $j = TP_j + FP_j$. Then $TP_j = j - FP_j$. It is reasonable to use $E(FP_j) = (m - \hat{s})p_{(j)}$ to approximate FP_j based on the null distribution of the p-values. Then TP_j can be approximated by $j - (m - \hat{s})p_{(j)}$ and, consequently, FNP_j can be approximated by $1 - j/\hat{s} + (m - \hat{s})p_{(j)}/\hat{s}$ as in Eq (1).

Moreover, because the true FNP_j is non-increasing with respect to j, stopping at the first time $\widehat{FNP}_j$ is less than ϵ_k can result in the smallest subset of variants with FNP controlled at ϵ_k with high probability. This step, as stated in Eq (2), can effectively exclude noise variants that are distinguishable from the $(1 - \epsilon_k)$ signals. More detailed justifications in theory and simulation of FNC screening can be found in [18].

We note that one can consider all possible p-value thresholds to search for the optimal reduced SNP subset $\mathcal{D}_{k^*}$ for joint-model training. However, considering all possible thresholds will result in an extremely large number of candidate models, K, to be trained in the joint modeling step using the target data. Thus the practical and conceptual benefits of FNC screening are to provide a moderate K (e.g., $K = 20$ in our default setting) based on the principle of FN control and to obtain the K candidate reduced SNP sets, each of which retains a high proportion of signal variants while minimizes the false positives by selecting the smallest subset of promising variants with respect to a specific FN proportion ϵ_k.

Results

Simulation design

For the target data in simulation, we obtain the genotype data of Chromosome 21 from the CoLaus/PsyCoLaus study [13, 14]. After data pre-process described in S1C Appendix, the genotype data for Chromosome 21 has $m = 5053$ SNPs. We randomly select n individuals to form the target sample, resulting in a genotype matrix X consisting of m columns and n rows. Let β be a m-dimensional vector of the allelic effects. We assume that β is a sparse vector with $\beta_j \neq 0$ for $j \in \mathcal{S}$, where $\mathcal{S}$ is the set of causal variants for the target sample. The trait vector Y are simulated from $Y = X\beta + W$, where $W \sim N_n(0, I)$, I is an identity matrix. The simulated target sample is randomly split into training set and testing set with a 1 : 1 ratio. The training set is used to build PRS models using different methods. The testing set is used to evaluate the predictive performances of these methods.

In our main simulations, we assume that $\mathcal{S} \subseteq \mathcal{S}^+$. The base summary data are generated by $Z \sim N_m(\mu^+, \Sigma/n_0)$, where μ^+ is the mean vector of Z, n_0 represents the base sample size, Σ is the correlation matrix of all the $m = 5053$ SNPs, which is estimated from the CoLaus/PsyCoLaus samples. Such marginal effect model has been used in GWAS analysis in e.g. [26]. The mean vector μ^+ has $\mu_j^+ \neq 0$ for $j \in \mathcal{S}^+$ and 0 otherwise, where $\mathcal{S}^+$ is the set of base signal variants. In the simulation studies, we consider 3 sets of (n_0, n), which are (4000, 2000), (4000, 1000) and (2000, 1000). We control the overlap between base signal variants ($\mathcal{S}^+$) and target causal variants ($\mathcal{S}$) using a proportion parameter $\delta = |\mathcal{S}|/|\mathcal{S}^+|$, where $|\cdot|$ represents the cardinality of a set. It is expected that the more diverse the base samples are, the smaller the δ would be. In the simulation examples, we randomly select 50 SNPs as signal variants in the base

Table 1. Summary of data generation schemes in the simulations. β^+ and β are the true effect size of base and target data, respectively, from the joint models; μ^+ is the mean of the summary statistics, typically obtained from marginal models; Σ is the correlation matrix of the $m = 5,053$ SNPs; n_0 and n are the sample size of base and target data, respectively.

	Main Simulations $\mathcal{S} \subseteq \mathcal{S}^+$	**Additional Simulations** $\mathcal{S} \nsubseteq \mathcal{S}^+$
Causal Variant Effect $(\beta_j^+;\beta_j)$ or Signal Variant Effect (μ_j^+)	$\mu_j^+ \sim Unif(0.05, 0.15)$ $\beta_j \sim Unif(0.05, 0.15)$	$\begin{bmatrix}\beta_j^+\\ \beta_j\end{bmatrix} \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau^2\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$ $\rho \neq 0$ if j shared between base and target; $\rho = 0$ otherwise
Base Summary Data Z	$Z \sim N_m(\mu^+, \Sigma/n_0)$	$Z \sim N_m(\mu^+ = \Sigma\beta^+, \Sigma/n_0)$
Target Data Y	$Y \sim N_n(X\beta, 1)$	$Y \sim N_n(X\beta, 1)$
Results	Fig 2 and Table 2	S1 Fig and S1 Table

https://doi.org/10.1371/journal.pgen.1010597.t001

sample (i.e., $|\mathcal{S}^+| = 50$) and set the number of target causal variants as $|S| = 50 \times \delta$. We consider $\delta = 0.3$, 0.5 and 0.7, representing low, median, and high overlaps between $\mathcal{S}$ and $\mathcal{S}^+$, respectively. The non-zero effect sizes of μ_j^+ and β_j are generated independently from Uniform (0.05, 0.15) and separately for base and target data. The range of the uniform distribution is calibrated so that R^2 roughly range from 5% to 30% for different methods across different scenarios. The heritability h^2, quantified by $V(X\beta)/V(Y)$, is about 0.14, 0.21 and 0.28 for $\delta = 0.3$, 0.5 and 0.7, respectively.

We also conduct additional simulations where the nested assumption is not satisfied (i.e., $\mathcal{S} \nsubseteq \mathcal{S}^+$) and the effect sizes of the causal/signal variants are generated from more conventional settings, e.g., allowing the true effect sizes of base and target causal variants (denoted by β^+ and β, respectively) in the joint-SNP models to be correlated and to be positive or negative, as well as incorporating SNP LD in both mean and variance of the base summary statistics Z when generating Z. Specifically, we control the overlap proportion between the target causal variants ($\mathcal{S}$) and the base causal variants (denoted by $\mathcal{S}_{\beta^+}^+$) using parameter $\delta^* = \left|\mathcal{S} \cap \mathcal{S}_{\beta^+}^+\right| / |\mathcal{S}|$. We consider $|\mathcal{S}| = |\mathcal{S}_{\beta^+}^+| = 50$, $\delta^* = 0.3$, 0.5, and 0.7, $(n_0, n) = (4000, 1000)$, and $h^2 \approx 0.33$. We provide the detailed designs of the additional simulations in S1B Appendix and summarize the data generation schemes in Table 1 for the simulation study.

Performance assessment and baseline methods

We simulate 100 replicates under each simulation scenario, and evaluate the predictive performance, model fitting and model parsimony for the final PRS model. Specifically, we examine two metrics: (1) R^2, which quantifies the outcome variation explained by the PRS on the target testing set; and (2)Akaike Information Criterion (AIC), which evaluates how well the PRS model fits the outcome data and has been used to evaluate polygenic prediction in [27] and [28]. AIC is calculated by $AIC = 2q + n_{test} \cdot \log(RSS/n_{test})$), where q is the number of SNPs retained in the PRS model, n_{test} is the sample size of the target testing set, and RSS is the residual sum of square (RSS) calculated on the target testing set. As an ancillary metric, we also report the number of SNPs included in the final PRS model.

We benchmark the proposed FNC+Lasso transfer learning method against five representative methods for PRS analysis, which are Clumping+Thresholding (CT; [5]), lassosum [8], LDpred [9, 11], Lasso [29], and SIS+Lasso. The first three methods use the LD information obtained from Chromosome 21 of the CoLaus/PsyCoLaus data, estimate the effect sizes using base summary data, and are implemented using R package `bingsnpr` [30] in our numerical studies. In contrast, Lasso only uses the individual-level target data, see for example [21]. The

last method, SIS+Lasso, can be viewed as a naive transfer learning procedure that also includes dimension reduction and joint-model training. Instead of using FNC screening and searching adaptively for the optimal reduced SNP set, SIS+Lasso applies sure independence screening (SIS) [31] to the base summary statistics to select the top $n_{train} - 1$ SNPs, with n_{train} being the target training sample size, and then performs Lasso regression on the selected SNPs with the target training data. It is well-known that SIS can be applied to marginal summary statistics to quickly reduce data dimension. We include SIS+Lasso in the baseline methods to evaluate potential advantages of using FNC screening in the transfer learning framework.

Simulation results

Fig 2 shows the boxplots of R^2 from 100 replications under the scenario of $\mathcal{S} \subseteq \mathcal{S}^+$, for (n_0, n) = (4000, 2000) (top row), (4000, 1000) (middle row) and (2000, 1000) (bottom row). In each replication, the target sample is randomly and equally divided into training and testing sets. In all scenarios, we observe that FNC+Lasso tend to give the highest R^2 among all methods. We also observe that methods using a trans-learning framework (i.e., FNC+Lasso and SIS+Lasso) generally outperform other baseline methods. The only exception is when the target training sample size is relatively small (e.g., n = 1000 and hence $n_{training}$ = 500) and the overlap between base signals and target causal variants is high (e.g., δ = 0.7), where CT, lassosum and LDpred can perform comparable to or better than SIS+Lasso. Nevertheless, FNC+Lasso, which aims to reserve an optimal high proportion of signals during screening, consistently yield better R^2 than SIS+Lasso and other baseline methods.

The relative performances of the non-trans-learning baseline methods are sensitive to the target sample size n and the overlap proportion δ. Specifically, when target sample size is relatively large (e.g., n = 2000; top row of Fig 2), Lasso tends to have higher R^2 than those methods

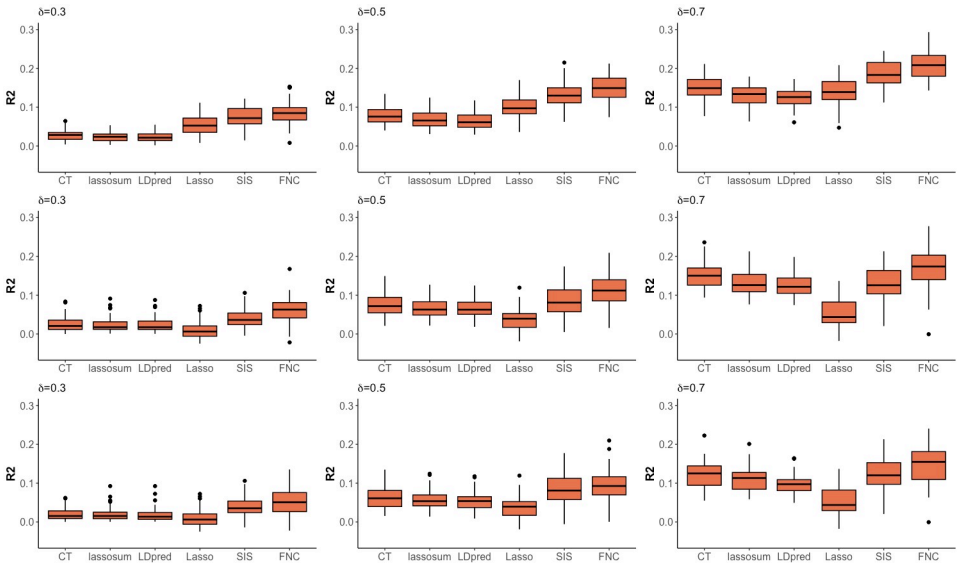

Fig 2. Results of R^2 for prediction accuracy of different PRS methods in the main simulations assuming $\mathcal{S} \subseteq \mathcal{S}^+$. The results are based on 100 simulation replicates under different sample-size combinations of base data (n_0) and target data (n), and under different overlap proportion between base signal variants ($\mathcal{S}^+$) and target causal variants ($\mathcal{S}$), i.e., $\delta \equiv |\mathcal{S}|/|\mathcal{S}^+|$ with $|\mathcal{S}^+| = 50$. From the top to bottom rows are for (n_0, n) = (4000, 2000), (4000, 1000) and (2000, 1000), respectively. In each replication, the target sample is randomly and equally divided into training and testing sets. From the left to right columns are for δ = 0.3, 0.5, and 0.7, respectively. Methods considered include Clumping +Thresholding (CT), lassosum, LDpred, Lasso, SIS+Lasso (SIS), and FNC+Lasso (FNC).

https://doi.org/10.1371/journal.pgen.1010597.g002

relying on the effect sizes estimated from base data, i.e., CT, lassosum, and LDpred. The gap is more obvious when the overlap proportion δ is small. The results suggest that re-estimating SNP effects using target training data (i.e., Lasso) can improve predictive performance when the target sample size is sufficiently large, and that the gain is more substantial when the base signals and target causal variants are less homogeneous (i.e., smaller δ).

In contrast, when target sample size is relatively small (e.g., $n = 1000$; middle and bottom rows of Fig 2), CT, lassosum, and LDpred generally outperform Lasso. The gap is more obvious when the overlap proportion δ becomes larger. The results suggest that when target sample size n (and hence the target training sample size n_{train}) is not sufficiently large, it becomes inefficient to re-estimate SNP effects using target training data, even when the base signals and target causal variants become heterogeneous (e.g., overlap proportion $\delta = 0.3$ or 0.5). When the overlap between the base signal set and the target causal set increases, relying on the base effect estimates, which are obtained on a larger sample size n_0, can enhance predictive R^2 (e.g., in the middle and bottom rows, the gap between Lasso and CT/lassosum/LDpred becomes larger as δ increases from 0.3 to 0.7). We also observe that the results of $(n_0, n) = (4000, 1000)$ and (2000, 1000) are very similar, implying that the base sample size n_0 has less impact on relative performance of the baseline methods.

Table 2 shows AIC and the number of SNPs included in the final PRS model under the scenario of $\mathcal{S} \subseteq \mathcal{S}^+$. A smaller AIC value indicates a better model fitting in simplicity and accuracy. When the overlap proportion is not high (e.g., $\delta = 0.3$ and 0.5), FNC+Lasso tends to have the lowest AIC and select the least number of SNPs, closely followed by SIS+Lasso. When δ is

Table 2. Results of Akaike Information Criterion (AIC) and number of SNPs in the final PRS model of different methods for assessing model fit and parsimony under the main simulations assuming $\mathcal{S} \subseteq \mathcal{S}^+$.

Overlap (δ)	Method	$(n_0, n) = (4000, 2000)$		$(n_0, n) = (4000, 1000)$		$(n_0, n) = (2000, 1000)$	
		Selected SNPs	AIC	Selected SNPs	AIC	Selected SNPs	AIC
0.3	CT	62 (50)	239 (102)	77 (130)	214 (267)	53 (41)	168 (96)
	lassosum	222 (182)	564 (361)	248 (287)	558 (567)	232 (258)	529 (513)
	LDpred	2732 (207)	5583 (419)	2634 (233)	5329 (463)	2454 (283)	4972 (570)
	Lasso	62 (32)	212 (87)	37 (35)	143 (83)	37 (35)	143 (83)
	SIS+Lasso	55 (23)	178 (77)	33 (17)	119 (50)	33 (18)	**119 (51)**
	FNC+Lasso	40 (20)	**131 (65)**	37 (28)	**117 (77)**	35 (29)	**119 (79)**
0.5	CT	54 (13)	274 (63)	59 (54)	198 (114)	51 (21)	188 (58)
	lassosum	187 (97)	551 (205)	173 (114)	430 (234)	151 (91)	391 (188)
	LDpred	2701 (179)	5582 (363)	2753 (219)	5592 (441)	2548 (276)	5187 (554)
	Lasso	107 (38)	356 (101)	57 (41)	215 (85)	57 (41)	215 (85)
	SIS+Lasso	83 (24)	273 (62)	55 (27)	186 (60)	54 (26)	185 (56)
	FNC+Lasso	62 (39)	**214 (109)**	47 (23)	**155 (64)**	43 (25)	**155 (63)**
0.7	CT	51 (9)	271 (57)	50 (7)	**187 (39)**	44 (12)	**193 (42)**
	lassosum	174 (73)	539 (155)	146 (71)	392 (151)	140 (63)	392 (136)
	LDpred	2719 (163)	5637 (339)	2734 (200)	5572 (399)	2576 (279)	5270 (552)
	Lasso	148 (47)	479 (108)	77 (50)	294 (100)	77 (50)	294 (100)
	SIS+Lasso	115 (27)	357 (64)	69 (24)	240 (55)	67 (23)	238 (55)
	FNC+Lasso	79 (54)	**257 (139)**	61 (35)	199 (95)	59 (38)	208 (93)

The reported values are the mean (and standard deviation) based on 100 simulation replicates under different sample-size combinations of base data (n_0) and target data (n), and under different overlap proportion between base signal variants ($\mathcal{S}^+$) and target causal variants ($\mathcal{S}$), i.e., $\delta \equiv |\mathcal{S}|/|\mathcal{S}^+|$ with $|\mathcal{S}^+| = 50$. Methods considered include Clumping+Thresholding (CT), lassosum, LDpred, Lasso, SIS+Lasso (SIS+Lasso), and FNC+Lasso (FNC+Lasso). Models with the smallest AIC as shown in bold.

https://doi.org/10.1371/journal.pgen.1010597.t002

high (e.g., 0.7), CT tends to have the lowest AIC and select the least number of SNPs, closely followed by FNC+Lasso.

The results of the additional simulations considering the non-nested scenario (i.e., $\mathcal{S} \subseteq \mathcal{S}^+$) are summarized in S1 Fig and S1 Table. For R^2 (S1 Fig), we see that across different δ^*'s (i.e., overlap proportions between the base and target causal variants) and different ρ's (i.e., effect correlations between the overlapping base and target causal variants), the methods utilizing a trans-learning framework (i.e., FNC+Lasso and SIS+Lasso) generally perform better than or comparable to CT, lassosum, and LDpred; however, the trans-learning methods tend to perform worse than Lasso, especially when there is limited overlap between the base signals and target causal variants. These outcomes are not surprising because leveraging information from the base data cannot enhance the prediction accuracy for the target individuals when a fair proportion of the causal variants are unique to the target population. Between the two transfer-learning methods, FNC+Lasso and SIS+Lasso have similar R^2 across all scenarios, which is perhaps because SIS always selects a fixed large number of SNPs into the joint-model training regardless of the informativeness of base data and consequently, retains some robustness against the base uninformativeness. As the overlap proportion δ^* increases from 0.3 to 0.7, the performance of Lasso remains stable, whereas all the other methods exhibit improvement with higher R^2 values; when $\delta^* = 0.7$ (i.e., the third column of S1 Fig), FNC+Lasso and SIS+Lasso yield R^2 comparable to Lasso. Finally, the effect correlation ρ does not seem to have much impact on the relative performance between FNC+Lasso and the baseline methods, except that when the effect sizes of the base and target causal variants are highly correlated (i.e., $\rho = 0.9$) and the overlap proportion is high ($\delta^* = 0.7$), all methods have similar R^2. For model fit and parsimony (S1 Table), SIS+Lasso yields the most parsimonious results with the smallest AIC measures in all settings, closely followed by FNC+Lasso and then Lasso..

Real data applications

We applied the proposed FNC+Lasso method and the baseline methods to construct the PRSs for plasma lipids, including total cholesterol (CHOL), triglycerides (TRIG), high-density lipoprotein (HDL) and low-density lipoprotein (LDL). Plasma lipids are risk factors for cardiovascular diseases, a leading cause of death worldwide [32], and CHOL, TRIG, HDL and LDL are effective biomarkers and risk factors for heart attack and heart diseases. At present, over 1,900 common SNPs have been shown to be associated with CHOL with different significant levels and the corresponding number for TRIG is 4,002 [33]. However, majority of these variants confer small risk individually and have limited predictive accuracy for lipid traits [34], although the variance proportions explained by PRS for Whites and Hispanics tend to be higher than other populations [35].

Our target data is obtained from the CoLaus/PsyCoLaus GWAS study of the Cohort Lausanne, Switzerland [13, 14], consisting of 5,247 Switzerland Caucasians individuals. We acquired the base data from [36] and [37], which includes samples of general European descents from US, central Europe (i.e., Australia, Denmark, France Germany, Netherlands, Switzerland, UK), southern Europe (Italy) and northern Europe (i.e., Iceland, Finland, and Sweden). The base sample has broader ancestral background and larger sample sizes (i.e., 94,595, 100,184, 99,900 and 95,454 for CHOL, TRIG, LDL and HDL, respectively). We preprocess the genotype data by conducting quality control (QC) and imputation as detailed in S1C Appendix and matched SNPs between the base and target data. The number of matched SNPs for CHOL, TRIG, LDL and HDL are 1,833,233, 1,756,492, 1,833,283 and 1,833,233, respectively.

To construct PRSs, we randomly split the target sample into training and testing sets, each containing half of the target sample. We apply the six methods, i.e., the 5 baseline methods (CT, lassosum, LDpred, Lasso and SIS+Lasso) and the proposed FNC+Lasso. For CT, lassosum and LDpred, the LD information is obtained from the CoLaus/PsyCoLaus samples, and the effect sizes in the PRS model are estimated using the base summary data. In all real data analyses, we include the top 10 principal components for population substructure, sex and age as covariates. We optimize the tuning parameters and obtain the effect size estimates for the final PRS model in training set, and calculate the prediction R^2 of the final PRS model in testing set. To account for the variations involved in the random data splits, we repeat the process using 50 different splitting of training and testing sets. In this part, we use the R package *bigsnpr* for Lasso to ensure computational efficiency with whole genome SNPs.

Table 3 shows the R^2, AIC and number of selected SNPs of different methods. Across all four traits, the proposed FNC+Lasso yields more robust performance in R^2 and AIC by yielding the best or comparable values to the best-performing methods, while the best-performing methods vary depending on traits, R^2 and AIC. For example, for CHOL and TRIG, FNC +Lasso has the highest R^2 and lowest AIC. For HDL and LDL, FNC+Lasso and SIS+Lasso have the highest R^2 while lassosum has the smallest AIC. We also observe that although giving reasonable AICs, the classical Lasso has the lowest R^2 for all four traits, suggesting that re-training

Table 3. Prediction R^2, number of selected SNPs, AIC, and computing time of different PRS methods for the CoLaus/PsyCoLaus analysis of four lipid traits.

Trait	Method	Predictive R^2	Selected SNPs	AIC	Computing Time (sec)
CHOL	CT	4.6% (0.6%)	50729 (122314)	101572 (244630)	1143 (101)
	lassosum	4.7% (0.5%)	957 (6514)	2032 (13023)	7320 (1)
	LDPred	4.6% (0.5%)	928258 (17319)	1856637 (34651)	7200 (1)
	Lasso	3.2% (0.2%)	315 (276)	782 (54)	326 (69)
	SIS + Lasso	7% (0.6%)	344 (47)	735 (88)	**15 (1)**
	FNC + Lasso	**7.2% (0.7%)**	229 (28)	**499 (79)**	616 (36)
TRIG	CT	5.9% (0.6%)	118645 (194875)	238203 (389599)	763 (272)
	lassosum	5.4% (0.5%)	832732 (54187)	1666391 (108373)	6900 (1)
	LDPred	5.4% (0.5%)	1183748 (59460)	2368424 (118933)	6781 (1)
	Lasso	3.4% (0.4%)	260 (257)	1503 (694)	143 (21)
	SIS + Lasso	7% (0.7%)	239 (44)	1359 (413)	**13 (1)**
	FNC + Lasso	**7.4% (0.8%)**	170 (21)	**1211 (402)**	533 (39)
HDL	CT	19% (1%)	15 (16)	-4865 (70)	1826 (310)
	lassosum	20.6% (1%)	9 (0)	**-4929 (62)**	7200 (1)
	LDPred	20.3% (1%)	865651 (3866)	1726365 (7729)	7800 (1)
	Lasso	18.9% (0.6%)	699 (361)	-3494 (732)	110 (27)
	SIS + Lasso	**23.8% (0.9%)**	365 (51)	-4326 (106)	**18 (5)**
	FNC + Lasso	**23.7% (0.9%)**	205 (38)	-4642 (96)	596 (76)
LDL	CT	4.2% (0.6%)	125 (302)	-258 (624)	961 (39)
	lassosum	4.2% (0.6%)	109 (0)	**-289 (64)**	8280 (1)
	LDPred	4.2% (0.6%)	893486 (27412)	1786465 (54829)	8881 (1)
	Lasso	1.8% (0.2%)	356 (290)	270 (587)	508 (151)
	SIS + Lasso	5.6% (0.5%)	331 (49)	117 (122)	**17 (3)**
	FNC + Lasso	**5.9% (0.6%)**	171 (26)	-211 (88)	579 (59)

The reported values are the mean (and standard deviation) from 50 random splits for each of the four lipid traits, including total cholesterol (CHOL), triglycerides (TRIG), high-density lipoprotein (HDL) and low-density lipoprotein (LDL). Methods considered include Clumping+Thresholding (CT), lassosum, LDpred, Lasso, SIS +Lasso (SIS+Lasso), and FNC+Lasso (FNC+Lasso). Best performing results across different methods are shown in bold.

https://doi.org/10.1371/journal.pgen.1010597.t003

a PRS model using Lasso on all SNPs with target training sample does not yield ideal predictive accuracy. Incorporating a screening procedure using base data (i.e., trans-learning-based SIS +Lasso and FNC+Lasso) can substantially enhance the predictive R^2 of Lasso. Similar to what is observed in the main simulations, Fig 3 shows that the trans-learning based methods consistently yield higher R^2 than other baseline methods for all four traits, and FNC+Lasso tends to have higher R^2 than SIS+Lasso. We also observe that FNC+Lasso tends to use the less number of SNPs while achieving a higher predictive R^2.

Table 3 also shows that the proposed FNC+Lasso is computationally efficient. It is slower than SIS+Lasso and Lasso because FNC+Lasso involves a grid search to determine two hyper-parameters (FN proportion ϵ_k and the shrinkage parameter of Lasso). However, the required computing time is still less than CT, lassosum and LDpred, e.g., FNC+Lasso can be $1.4 \sim 15$ times faster than these baseline methods.

The parsimonious modeling of FNC+Lasso could foster interpretability and facilitate more complex modeling in downstream analyses, such as incorporating SNP-SNP interactions in polygenic prediction models. Here we explore pairwise interactions of the SNPs selected in FNC+Lasso on the lipid traits. To avoid potential overestimation of the interaction effects, we fit an linear regression model that satisfies strong hierarchy (i.e., whenever a SNP-SNP interaction is estimated to be non-zero by Lasso, the main effects of both SNPs are also included in the model) so that the interaction effects are not identified due to their collinearity with the omitted main effects. This procedure utilizes hierarchical group-Lasso regulation introduced in [38] and can be applied using the R package *glinternet*. We investigate the predictive R^2

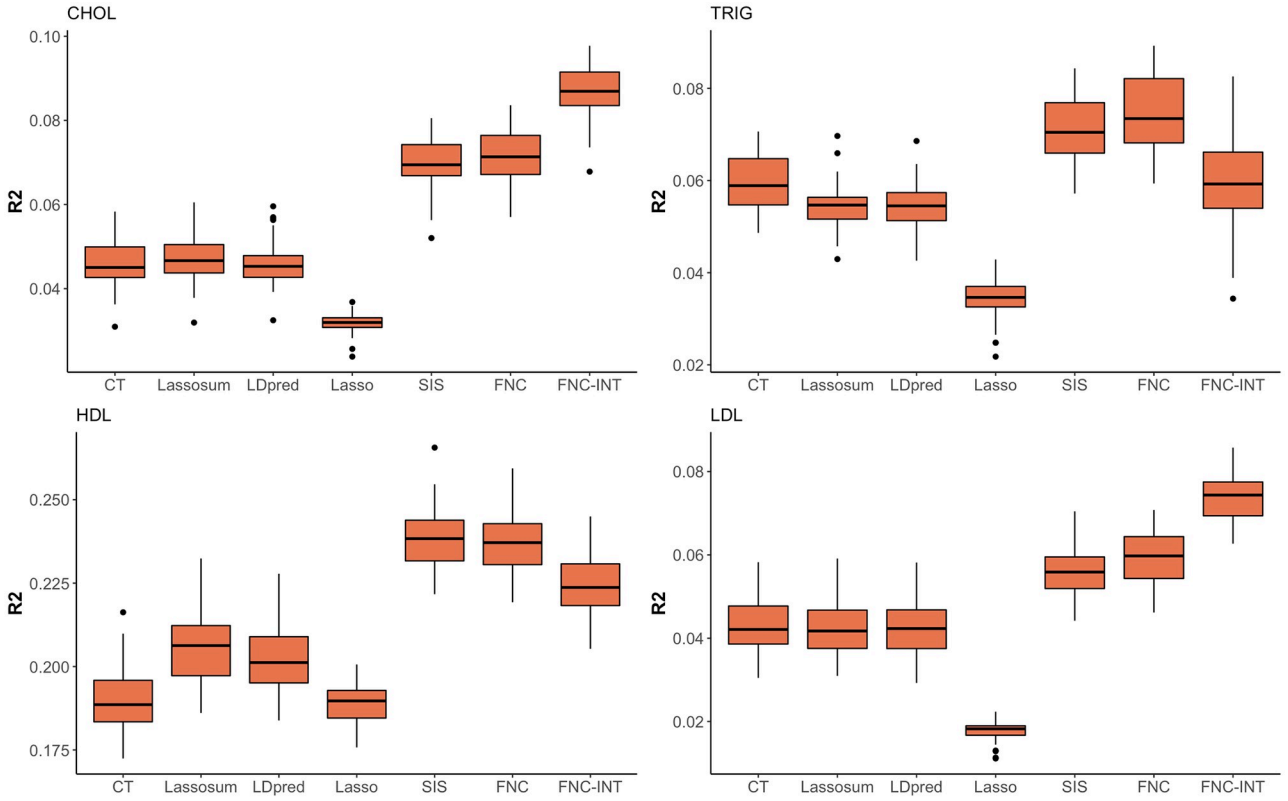

Fig 3. R^2 values for prediction accuracy of different PRS methods. Methods considered include Clumping+Thresholding (CT), lassosum, LDpred, Lasso, SIS+Lasso (SIS), FNC+Lasso (FNC), and interaction model based on FNC+Lasso identified SNPs (FNC-INT) for four lipid traits.

https://doi.org/10.1371/journal.pgen.1010597.g003

performance of the interaction PRS model based on FNC-Lasso selected SNPs (denoted as FNC-INT) for the four traits. The results in Fig 3 show that incorporating SNP-SNP interaction terms can further increase the predictive performance for CHOL and LDL, although interactions do not help in TRIG and HDL. Recent studies have gain valuable insights into the functional consequences of some gene-gene interactions on LDL through systematic experiments that simultaneously knockdown candidate gene pairs in cultured HeLa cells [39]. Further follow-up studies could relate our identified SNP-SNP pairs to candidate genes and gene pairs through eQTL analysis and evaluate their roles in biological pathways of lipid traits.

Conclusion and discussion

Inspired by the idea of transfer learning, the proposed FNC+Lasso procedure aims to leverage the base data information for target data prediction modeling, where the target data may or may not have similar ancestral background as the base data. FNC+Lasso utilizes a unique false negative control strategy to extract useful information from the base data, and applies the extracted information to re-train the PRS models using target data in a statistically and computationally efficient fashion. The benefits are multi-fold: (1) FNC screening leverages the base information to effectively retain a high proportion of base signal variants and yields a moderate number of candidate joint models, which reduces the cost in model re-training. (2) Only the identities of SNPs, not their effect sizes, are carried over to joint modeling, which encourages accurate PRS estimation and prediction for the target samples. (3) Through the efficient exclusion of noise variants by FNC screening, dimensions of the candidate joint models are largely reduced, which helps to alleviate the challenges associated with the "large p small n" joint modeling and leads to enhanced prediction accuracy. (4) The resulting PRS model of FNC+Lasso are parsimonious with better fit, which facilitates model interpretation and complex modeling in follow-up analyses.

The proposed PRS transfer learning framework can accommodate continuous and categorical traits by applying linear or logistic models in the joint modeling step. It is also applicable to target data with only summary statistics available. In that case, the Lasso step on individual-level target data can be replaced by using the lassosum algorithm with GWAS summary statistics from target data. Just like lassosum, the specific summary statistics needed would include: (1) the LD matrix from reference panel with similar ancestral background as the target samples, and (2) the SNP-phenotype correlation from target summary statistics.

FNC+Lasso adopts a transfer learning framework and relies on the nested assumption of $\mathcal{S} \subseteq \mathcal{S}^+$, i.e., the target causal variants $\mathcal{S}$ are nested within the base signal variants $\mathcal{S}^+$. In real practice, such nested assumption could be satisfied by adopting multi-ancestry GWAS base data, e.g., using the large-scale meta-analysis from Global Biobank Meta-analysis Initiative as discussed in Wang et al. [17]. Using simulations, we explore the impact of overlapping proportion $\delta \equiv |\mathcal{S}|/|\mathcal{S}^+|$ under the scenario of $\mathcal{S} \subseteq \mathcal{S}^+$. We observed that when δ is low (i.e., heterogeneous effect patterns between base and target data), re-training PRS models with sufficient target samples can achieve better predictive performance. When δ is high (i.e., similar effect patterns between base and target data), utilizing effect estimates from base data can ensure higher predictive accuracy, especially when target samples are of limited size. Our numerical studies also show that FNC+Lasso adapts to overlap level—FNC+Lasso ensures effective PRS joint-model training even with limited target samples, and results in a more robust and better predictive performance than the baseline methods across a range of δ values and different base/target sample sizes.

We also conduct additional simulations where the nested assumption is not satisfied. The results show that FNC+Lasso could still outperform existing PRS methods (e.g., C+T, lassosum

and LDpred), has similar R^2 as SIS+Lasso, but has smaller R^2 than Lasso. However, Lasso exhibits unstable performance across different nested and non-nested scenarios in our numerical studies. Given that the actual informativeness level of base data on the target causal variants is unknown and can vary from trait to trait in real applications, we would recommend to use methods utilizing a transfer learning framework (i.e., FNC+Lasso and SIS+Lasso) for PRS analysis. If multi-ancestry GWAS base data could be used, FNC+Lasso may provide further improvement upon SIS+Lasso.

The PRS models of FNC+Lasso tend to be more parsimonious compared to baseline models while maintaining a comparable or higher R^2. A parsimonious PRS model that consists of key SNPs and excludes noise SNPs could aid the interpretablity of PRS and help to inform underlying molecular and functional basis, such as by examining the SNPs involved in the PRS model as well as those SNPs in LD with them to identify biologically important variants or to identify hub genes that capture larger polygenicity [40]. By focusing on the SNPs in a parsimonious PRS model, it also allows further exploration of non-additive PRS models, such as the SNP-SNP interaction PRS model considered in the CoLaus/PsyCoLaus data analysis. PRS models incorporating epistasis have been challenging because existing PRS models tend to be comprised of a large number of SNPs, and the corresponding interactions quickly leads to a prohibitive number of predictors in the model. While machine-learning PRS methods such as random forest have been proposed to account for unknown epistasis [41], it remains an open question how the base data information can be incorporated into these learning approaches. Complex models based on FNC+Lasso SNPs may offer an alternative for constructing non-additive PRS.

Supporting information

S1 Appendix. FNC+Lasso algorithm, details of additional simulations, and preprocess of the CoLaus/PsyCoLaus GWAS data. Section A shows the FNC+Lasso algorithm. Section B presents the designs and results of the additional simulations. Section C describes the QC and imputation of the CoLaus/PsyCoLaus GWAS data.
(PDF)

S1 Fig. Results of R^2 for prediction accuracy of different PRS methods in the additional simulations assuming $\mathcal{S} \nsubseteq \mathcal{S}^+$, with $(n_0, n) = (4000, 1000)$, $|\mathcal{S}| = |\mathcal{S}^+_{\beta^+}| = 50$, $\delta^* = 0.3, 0.5$ and 0.7, and $\rho = 0.5, 0.7$ and 0.9.
(PDF)

S1 Table. Results of Akaike Information Criterion (AIC) and number of SNPs in the final PRS model of different methods for assessing model fit and parsimony in the additional simulations assuming $\mathcal{S} \nsubseteq \mathcal{S}^+$, with $(n_0, n) = (4000, 1000)$, $|\mathcal{S}| = |\mathcal{S}^+_{\beta^+}| = 50$, $\delta^* = 0.3, 0.5$ and 0.7, and $\rho = 0.5, 0.7$ and 0.9.
(PDF)

S2 Table. Numerical data underlying graphs and tables.
(XLSX)

Acknowledgments

The authors thank Dr. Peter Vollenweider, Dr. Gerard Waeber, Dr. Julien Vaucher and Dr. Martin Preisig, PIs of the CoLaus/PsyCoLaus study, and collaborators at GSK for providing the CoLaus/PsyCoLaus phenotype and genotype data.

Author Contributions

Conceptualization: Xinge Jessie Jeng, Yifei Hu, Jung-Ying Tzeng.

Data curation: Xinge Jessie Jeng, Yifei Hu, Jung-Ying Tzeng.

Formal analysis: Xinge Jessie Jeng, Yifei Hu, Vaishnavi Venkat, Tzu-Pin Lu.

Funding acquisition: Xinge Jessie Jeng, Jung-Ying Tzeng.

Investigation: Xinge Jessie Jeng, Yifei Hu, Vaishnavi Venkat, Tzu-Pin Lu, Jung-Ying Tzeng.

Methodology: Xinge Jessie Jeng, Yifei Hu, Jung-Ying Tzeng.

Project administration: Xinge Jessie Jeng, Jung-Ying Tzeng.

Resources: Xinge Jessie Jeng.

Software: Xinge Jessie Jeng.

Supervision: Xinge Jessie Jeng, Jung-Ying Tzeng.

Validation: Xinge Jessie Jeng.

Writing – original draft: Xinge Jessie Jeng, Yifei Hu, Jung-Ying Tzeng.

Writing – review & editing: Xinge Jessie Jeng, Yifei Hu, Vaishnavi Venkat, Tzu-Pin Lu, Jung-Ying Tzeng.

References

1. Consortium IS. Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. Nature. 2009; 460(7256):748. https://doi.org/10.1038/nature08185
2. Franceschini N, Fox E, Zhang Z, Edwards TL, Nalls MA, Sung YJ, et al. Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. The American Journal of Human Genetics. 2013; 93(3):545–554. https://doi.org/10.1016/j.ajhg.2013.07.010 PMID: 23972371
3. Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS biology. 2013; 11(9):e1001661. https://doi.org/10.1371/journal.pbio.1001661 PMID: 24068893
4. Coram MA, Fang H, Candille SI, Assimes TL, Tang H. Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. The American Journal of Human Genetics. 2017; 101 (2):218–226. https://doi.org/10.1016/j.ajhg.2017.06.015 PMID: 28757202
5. Privé F, Vilhjálmsson BJ, Aschard H, Blum MG. Making the most of Clumping and Thresholding for polygenic scores. The American Journal of Human Genetics. 2019; 105(6):1213–1221. https://doi.org/10.1016/j.ajhg.2019.11.001 PMID: 31761295
6. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. Bioinformatics. 2015; 31 (9):1466–1468. https://doi.org/10.1093/bioinformatics/btu848 PMID: 25550326
7. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019; 8(7):giz082. https://doi.org/10.1093/gigascience/giz082 PMID: 31307061
8. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genetic epidemiology. 2017; 41(6):469–480. https://doi.org/10.1002/gepi.22050 PMID: 28480976
9. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The american journal of human genetics. 2015; 97 (4):576–592. https://doi.org/10.1016/j.ajhg.2015.09.001 PMID: 26430803
10. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications. 2019; 10(1):1–10. https://doi.org/10.1038/s41467-019-09718-5 PMID: 30992449
11. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. BioRxiv. 2020;.

12. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics. 2019; 211 (4):1131–1141. https://doi.org/10.1534/genetics.119.301859 PMID: 30967442

13. Firmann M, Mayor V, Vidal PM, Bochud M, Pécoud A, Hayoz D, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC cardiovascular disorders. 2008; 8(1):6. https://doi.org/10.1186/1471-2261-8-6 PMID: 18366642

14. Preisig M, Waeber G, Vollenweider P, Bovet P, Rothen S, Vandeleur C, et al. The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. BMC Psychiatry. 2009; 9(1):9. https://doi.org/10.1186/1471-244X-9-9 PMID: 19292899

15. Franco NR, Massi MC, Ieva F, Manzoni A, Paganoni AM, Zunino P, et al. Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. Radiotherapy and Oncology. 2021; 159:241–248. https://doi.org/10.1016/j.radonc.2021.03.024 PMID: 33838170

16. Cope JL, Baukmann HA, Klinger JE, Ravarani CN, Böttinger EP, Konigorski S, et al. Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. Frontiers in Genetics. 2021; 12. https://doi.org/10.3389/fgene.2021.744557 PMID: 34745218

17. Wang Y, Namba S, Lopera E, Kerminen S, Tsuo K, Lall K, et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. Cell Genomics. 2023; 3(1):100241. https://doi.org/10.1016/j.xgen.2022.100241 PMID: 36777179

18. Jeng XJ, Hu Y. Weak Signal Inclusion Under Sparsity and Dependence. arXiv preprint arXiv:200615667. 2022;.

19. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS genetics. 2013; 9(7):e1003608. https://doi.org/10.1371/journal.pgen.1003608 PMID: 23874214

20. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. The American Journal of Human Genetics. 2013; 92(6):1008–1012. https://doi.org/10.1016/j.ajhg.2013.05.002 PMID: 23731541

21. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS genetics. 2014; 10(11):e1004754. https://doi.org/10.1371/journal.pgen.1004754 PMID: 25393026

22. Jeng XJ, Daye ZJ, Lu W, Tzeng JY. Rare variants association analysis in large-scale sequencing studies at the single locus level. PLoS computational biology. 2016; 12(6). https://doi.org/10.1371/journal.pcbi.1004993 PMID: 27355347

23. Cai TT, Sun W. Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies. Journal of the Royal Statistical Society: Series B. 2017; 79(1):197–223. https://doi.org/10.1111/rssb.12171

24. Jeng XJ, Chen X. Variable selection via adaptive false negative control in linear regression. Electron J Statist. 2019; 13(2):5306–5333. https://doi.org/10.1214/19-EJS1649

25. Jeng XJ. Estimating the proportion of signal variables under arbitrary covariance dependence. Electronic Journal of Statistics. 2023; 17(1):950–979. https://doi.org/10.1214/23-EJS2119

26. Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence. Journal of the American Statistical Association. 2012; 107(499):1019–1035. https://doi.org/10.1080/01621459.2012.720478 PMID: 24729644

27. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using summary statistics. Nature Genetics. 2020; 52(4):458–462. https://doi.org/10.1038/s41588-020-0600-y PMID: 32203469

28. Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nature Communications. 2021; 12:4192. https://doi.org/10.1038/s41467-021-24485-y PMID: 34234142

29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996; 58(1):267–288.

30. Privé F, Aschard H, Ziyatdinov A, Blum MG. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018; 34(16):2781–2787. https://doi.org/10.1093/bioinformatics/bty185 PMID: 29617937

31. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(5):849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

32. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics-2021 update: A report from the American Heart Association. Circulation. 2021; 143(8): e254–e743. https://doi.org/10.1161/CIR.0000000000000950 PMID: 33501848

33. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47(D1):D1005–D1012. https://doi.org/10.1093/nar/gky1120 PMID: 30445434

34. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nature genetics. 2018; 50 (11):1514–1523. https://doi.org/10.1038/s41588-018-0222-9 PMID: 30275531

35. Cupido AJ, Tromp TR, Hovingh GK. The clinical applicability of polygenic risk scores for LDL-cholesterol: Considerations, current evidence and future perspectives. Current Opinion in Lipidology. 2021; 32 (2):112. https://doi.org/10.1097/MOL.0000000000000741 PMID: 33560669

36. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466(7307):707–713. https://doi.org/10.1038/nature09270 PMID: 20686565

37. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nature genetics. 2013; 45(11):1274. https://doi.org/10.1038/ng.2797 PMID: 24097068

38. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. Journal of Computational and Graphical Statistics. 2015; 24(3):627–654. https://doi.org/10.1080/10618600.2014.938812 PMID: 26759522

39. Zimoń M, Huang Y, Trasta A, Halavatyi A, Liu JZ, Chen CY, et al. Pairwise effects between lipid GWAS genes modulate lipid plasma levels and cellular uptake. Nature communications. 2021; 12(1):1–16. https://doi.org/10.1038/s41467-021-26761-3 PMID: 34741066

40. Baker E, Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. Frontiers in Digital Health. 2020; 2. https://doi.org/10.3389/fdgth.2020.00014 PMID: 34713027

41. Lau M, Wigmann C, Kress S, Schikowski T, Schwender H. Evaluation of tree-based statistical learning methods for constructing genetic risk scores. BMC Bioinformatics. 2022; 23(1):97. https://doi.org/10.1186/s12859-022-04634-w PMID: 35313824